Inaugural dissertation

for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of the
Ruprecht - Karls – University
Heidelberg

Presented by

M.Sc. Ioannis Sarropoulos

born in Thessaloniki, Greece

Oral examination: January 17th 2023

Non-coding genome contributions to the development and evolution of mammalian organs

<u>Referees:</u>

Prof. Dr. Henrik Kaessmann

Prof. Dr. Duncan Odom

Summary

Protein-coding sequences only cover 1-2% of a typical mammalian genome. The remaining non-coding space hides thousands of genomic elements, some of which act via their DNA sequence while others are transcribed into non-coding RNAs. Many well-characterized non-coding elements are involved in the regulation of other genes, a process essential for the emergence of different cell types and organs during development. Changes in the expression of conserved genes during development are in turn thought to facilitate evolutionary innovation in form and function. Thus, non-coding genomic elements are hypothesized to play important roles in developmental and evolutionary processes. However, challenges related to the identification and characterization of these elements, in particular in non-model organisms, has limited the study of their overall contributions to mammalian organ development and evolution. During my dissertation work, I addressed this gap by studying two major classes of non-coding elements, long non-coding RNAs (lncRNAs) and *cis*-regulatory elements (CREs).

In the first part of my thesis, I analyzed the expression profiles of lncRNAs during the development of seven major organs in six mammals and a bird. I showed that, unlike protein-coding genes, only a small fraction of lncRNAs is expressed in reproducibly dynamic patterns during organ development. These lncRNAs are enriched for a series of features associated with functional relevance, including increased evolutionary conservation and regulatory complexity, highlighting them as candidates for further molecular characterization. I then associated these lncRNAs with specific genes and functions based on their spatiotemporal expression profiles. My analyses also revealed differences in lncRNA contributions across organs and developmental stages, identifying a developmental transition from broadly expressed and conserved lncRNAs towards an increasing number of lineage- and organ-specific lncRNAs.

Following up on these global analyses, I then focused on a newly-identified lncRNA in the marsupial opossum, *Female Specific on chromosome X (FSX)*. The broad and likely autonomous female-specific expression of *FSX* suggests a role in marsupial X-chromosome inactivation (XCI). I showed that *FSX* shares many expression and sequence features with another lncRNA, *RSX* — a known regulator of XCI in marsupials. Comparisons to other marsupials revealed that both *RSX* and *FSX* emerged in the common marsupial ancestor and have since been preserved in marsupial genomes, while their broad and female-specific expression has been retained for at least 76 million years of evolution. Taken together, my analyses highlighted *FSX* as a novel candidate for regulating marsupial XCI.

In the third part of this work, I shifted my focus to CREs and their cell type-specific activities in the developing mouse cerebellum. After annotating cerebellar cell types and states based on single-cell chromatin accessibility data, I identified putative CREs and characterized their spatiotemporal activity across cell types and developmental stages. Focusing on progenitor cells, I described temporal changes in CRE activity that are shared between early germinal zones, supporting a model of cell fate induction through common developmental cues. By examining chromatin accessibility dynamics during neuronal differentiation, I revealed a gradual divergence in the regulatory programs of major cerebellar neuron types.

In the final part, I explored the evolutionary histories of CREs and their potential contributions to gene expression changes between species. By comparing mouse CREs to vertebrate genomes and chromatin accessibility profiles from the marsupial opossum, I identified a temporal decrease in CRE conservation, which is shared across cerebellar cell types. However, I also found differences in constraint between cell types, with microglia having the fastest evolving CREs in the mouse cerebellum. Finally, I used deep learning models to study the regulatory grammar of cerebellar cell types in human and mouse, showing that the sequence rules determining CRE activity are conserved across mammals. I then used these models to retrace the evolutionary changes leading to divergent CRE activity between species.

Collectively, my PhD work provides insights into the evolutionary dynamics of non-coding genes and regulatory elements, the processes associated with their conservation, and their contributions to the development and evolution of mammalian cell types and organs.

Zusammenfassung

Protein-kodierende Sequenzen machen nur 1-2% eines typischen Säugetiergenoms aus. Der verbleibende nicht-kodierende Bereich verbirgt Tausende genomische Elemente, von denen einige durch ihre DNA-Sequenz wirken, während andere in nicht-kodierende RNAs transkribiert werden. Viele gut charakterisierte nicht-kodierende Elemente sind an der Regulierung anderer Gene beteiligt, ein Prozess, der für die Entstehung verschiedener Zelltypen und Organe während der Entwicklung essentiell ist. Es wird angenommen, dass Veränderungen in der Expression konservierter Gene während der Entwicklung evolutionäre Innovationen in Form und Funktion hervorrufen. Daher geht man davon aus, dass nicht-kodierende genomische Elemente eine wichtige Rolle bei Entwicklungs- und Evolutionsprozessen spielen. Diese Elemente sind jedoch schwer zu identifizieren und zu charakterisieren, vor allem in Nicht-Modellorganismen. Deshalb wurde ihr Gesamtbeitrag zur Entwicklung und Evolution von Säugetierorganen bisher nur wenig untersucht. In dieser Dissertation habe ich daher zwei Hauptklassen von nicht-kodierenden Elementen umfangreicher untersucht: lange nicht-kodierende RNAs (lncRNAs) und *cis*-regulierende Elemente (CREs).

Im ersten Teil meiner Arbeit analysierte ich die Expressionsprofile von lncRNAs während der Entwicklung von sieben Organen bei sechs Säugetier- und einer Vogelspezies. Ich konnte zeigen, dass während der Organentwicklung, im Gegensatz zu proteinkodierenden Genen, nur wenige lncRNAs in reproduzierbar dynamischen Mustern transkribiert werden. Diese weisen mehrere Merkmale auf, die auf funktionelle Relevanz deuten, darunter eine erhöhte evolutionäre Konservierung und regulatorische Komplexität, was sie zu Kandidaten für eine weitere molekulare Charakterisierung macht. Anschließend habe ich diese lncRNAs auf der Grundlage ihrer Expressionsprofile mit bestimmten Genen und Funktionen verbunden. Meine Analysen ergaben auch Unterschiede in lncRNA Expression zwischen verschiedenen Organen und Entwicklungsstadien, wobei ich einen entwicklungsbedingten Übergang von breit transkribierten und konservierten lncRNAs hin zu einer zunehmenden Anzahl von spezies- und organspezifischen lncRNAs feststellte.

Im Anschluss an diese globalen Analysen konzentrierte ich mich dann auf eine neu identifizierte lncRNA beim Beuteltier Opossum, *Female Specific auf Chromosom X (FSX)*. Die breite, wahrscheinlich autonome und auf Weibchen begrenzte Expression von *FSX* lässt auf eine Rolle bei der X-Chromosom-Inaktivierung (XCI) bei Beuteltieren schließen. Ich zeigte, dass *FSX* viele Expressions- und Sequenzmerkmale mit einem bekannten Regulator der XCI bei Beuteltieren, die lncRNA *RSX*, teilt. Vergleiche mit anderen Beuteltieren ergaben, dass sowohl *RSX* als auch *FSX* im gemeinsamen Vorfahren der Beuteltiere entstanden und seitdem in den Genomen der Beuteltiere erhalten geblieben sind, wobei ihre breite und auf Weibchen begrenzte Transkription über mindestens 76 Millionen Jahre Evolution erhalten geblieben ist. Zusammengenommen haben meine Analysen *FSX* als einen neuen Kandidaten für die Regulierung der XCI der Beuteltiere herausgestellt.

Im dritten Teil dieser Arbeit analysierte ich CREs und ihre zelltypspezifischen Aktivitäten in der Kleinhirnentwicklung der Maus. Nach der Annotation von Kleinhirnzellen durch ihre Chromatinprofile, habe ich mutmaßliche CREs identifiziert und ihre Aktivität über Zelltypen und Entwicklungsstadien hinweg charakterisiert. Ich entdeckte zeitliche Veränderungen der CRE-Aktivität in Vorläuferzellen, die parallel in allen frühen Keimzonen auftreten, und ein Modell der Induktion des Zellschicksals durch gemeinsame Signale in der Entwicklung unterstützen. Durch die Untersuchung der Chromatinzugänglichkeitsdynamik während der neuronalen Differenzierung konnte ich außerdem eine graduelle Divergenz in den Regulierungsprogrammen der wichtigsten Kleinhirnneuronentypen feststellen.

Im letzten Teil untersuchte ich die Evolutionsgeschichte der CREs und ihren potenziellen Beitrag zu Veränderungen der Genexpression zwischen Säugetieren. Durch den Vergleich von CREs der Maus mit Wirbeltiergenomen und Chromatinzugänglichkeitsprofilen des Beuteltiers Opossum habe ich eine zeitliche Abnahme der Konservierung von CREs für alle Kleinhirnzelltypen festgestellt. Ich fand jedoch auch Unterschiede im Konservierungsgrad verschiedener Zelltypen, wobei Mikroglia die am schnellsten evolvierenden CREs im Kleinhirn der Maus aufweisen. Schließlich untersuchte ich mit Hilfe von Deep-Learning-Modellen die regulatorische Grammatik der Kleinhirnzelltypen von Mensch und Maus. Dadurch konnte ich zeigen, dass die Sequenzregeln, die die CRE-Aktivität bestimmen, zwischen Säugetieren konserviert sind. Anschließend verwendete ich diese Modelle, um die evolutionären Veränderungen nach-zuvollziehen, die zu einer unterschiedlichen CRE-Aktivität zwischen Spezies führen.

Insgesamt bietet meine Doktorarbeit Einblicke in die evolutionäre Dynamik von nicht-kodierenden Genen und regulatorischen Elementen, in die Prozesse, die mit ihrer Konservierung verbunden sind, und in ihren Beitrag zur Entwicklung und Evolution von Zelltypen und Organen bei Säugetieren.

Acknowledgements

Science is a group endeavor and this work would have never been possible without the help and support of many people.

First and foremost, I would like to thank my supervisor and mentor, Prof. Dr. Henrik Kaessmann, for giving me the opportunity to work in such an inspiring scientific environment. He believed in me and my ideas, allowed me to pursue my dreams and supported me in every possible way. He also gently pointed out my shortcomings and patiently helped me address them, always pushing me to become a better scientist and person. I couldn't have asked for more from a PhD supervisor.

I am also deeply grateful to two exceptional scientists and dear friends, Dr. Margarida Cardoso-Moreira and Dr. Mari Sepp. I feel blessed to have had the opportunity to work with them, which taught me so much, both scientifically and personally. They have both served as role models for me, inspiring me with their love for science and dedication to their work. Our long scientific discussions have been some of the most enjoyable and productive moments of my doctoral years.

Many more colleagues helped me in many ways during my PhD work. I am very thankful, in alphabetical order, to: Robert Frömel, Dr. Piyush Joshi, Dr. Lena Kutscher, Kevin Leiss, Dr. Evgeny Leushkin, Dr. Ray Marin, Dr. Konstantin Okonechnikov, Prof. Dr. Stefan Pfister, Philipp Schäfer, Julia Schmitt, Céline Schneider, Dr. Tania Studer, Nils Trost and Tetsuya Yamada. This work could not have been done without them.

During my PhD, I had the rare opportunity to spend three months visiting another lab. I am grateful to Prof. Dr. Stein Aerts and his lab members, in particular Dr. Nikolai Hecker, Ibrahim Taskiran and Carmen Bravo González-Blas, who welcomed me in their group and generously shared their expertise with me. I am also thankful to EMBO for funding my visit.

My deepest thanks to the members of my thesis advisory committee, Prof. Dr. Duncan Odom and Dr. Judith Zaugg for their critical comments and wholehearted support throughout these years. They were truly invested in helping me succeed and showed this in practice on multiple occasions. I would also like to thank them, as well as Dr. Ralph Grand, for critically reading this thesis and forming my examination committee.

I am also extremely thankful to all the past and present members of the Kaessmann lab for providing such a friendly and supportive environment, leading me to constantly prolong my stay in the group. I am very grateful to Evgeny and Francesco who helped me in my first steps in bioinformatics; to Noe, Nils, Bassi, Evgeny and Matthias (and all other guest members) for our epic D&D sessions; to Juan, Tania, Evgeny and Xuefei for not giving up on their hope to make me better at bouldering; and to Robert, Philipp, Tetsuya, Xuefei and Leti for some really fun volleyball games. I feel really lucky to have met all of them.

I would also like to thank Kathrin Hall and Kornelia Mack for their constant support with every administrative hurdle that came my way during my time in the group.

I was also very lucky to make friends out of the lab who have been essential in supporting me during my PhD. Many thanks to Spiros, Juan, Paris, Eva, Ellen, Bryan, Rafail, Kai, Elisa, Vincent, Saskia and many others with whom I shared so many fun memories in Heidelberg.

I am also very thankful to my family and all my friends in Greece. Despite the distance, I always felt their support and love. Being away from them has been hard but it also pushed me to make the most out of my time abroad. Special thanks to my parents, Alexandros and Sofia, who from a very young age inspired me to be curious, encouraged me to always ask more questions, and taught me to never give up on my dreams.

Finally, I am deeply grateful to my partner Iris Valent who has always been there for me, supporting me in my failures and celebrating my successes. Her super-power to be able to accept me for who I am while simultaneously inspiring me to be a better person kept me going through a pandemic, the breakdown of our institute, and my own personal insecurities. Nothing makes me more optimistic about the future than knowing that we will share the next adventure together.

Table of Contents

List of abbreviations

Commonly used abbreviations not defined in text

bHLH	basic helix–loop–helix
BLAST	Basic Local Alignment Search Tool
bp	base pair
CRISPR	clustered regularly interspaced short palindromic repeats
DNA	deoxyribonucleic acid
GABA	γ-Aminobutyric acid (gamma-aminobutyric acid)
kb	kilobase (1,000 bp)
Mb	megabase (1,000,000 bp)
nt	nucleotide
RNA	ribonucleic acid
UMI	unique molecular identifier

Abbreviations defined in text

ATAC-seq	Assay for Transposase-Accessible Chromatin with high-throughput sequencing
au	area under (the curve)
CCA	canonical correlation analysis
ChIP	chromatin immunoprecipitation
CPM	counts per million
CRE	*cis*-regulatory element
CS	Carnegie stage
DN	deep nuclei
E[X]	embryonic day [X] (e.g., E10 – embryonic day 10)
EGL	external granule cell layer
FDR	false discovery rate
GCL	granule cell layer
HCR	hybridization chain reaction
LINE	long interspersed nuclear element
lncRNA	long non-coding RNA
LSI	latent semantic indexing
LSTM	long short-term memory
LTR	long terminal repeats
ML	molecular layer
NMF	non-negative matrix factorization
mRNA	messenger RNA
P[X]	postnatal day [X] (e.g., P4 – postnatal day 4)
PCA	principal component analysis
PCL	Purkinje cell layer
pcRNA	positionally-conserved RNA
piRNA	PIWI-interacting RNA
PR	precision-recall
ROC	receiver operating characteristic
RPKM	reads per kilobase of transcript, per million mapped reads

rRNA	ribosomal RNA
scRNA-seq	single-cell RNA sequencing
SINE	short interspersed nuclear element
snATAC-seq	single-nucleus ATAC-sequencing
snRNA	small nuclear RNA
snRNA-seq	single-nucleus RNA-sequencing
snoRNA	small nucleolar RNA
TAD	topologically associating domain
TE	transposable element
TF	transcription factor
tRNA	transfer RNA
TSS	transcription start site
UBC	unipolar brush cell
UMAP	uniform manifold approximation and projection
UTR	untranslated region
VST	variance stabilizing transformation

1. Introduction

1.1 Non-coding regions in mammalian genomes

In 1957, four years after reporting the structure of the DNA double helix (*1*), Francis Crick described the "central dogma of biology", according to which genetic information is stored in DNA and eventually transferred into proteins through an RNA intermediate (*2*). Based on this view, DNA is tasked with the preservation of genetic information through DNA replication and RNA serves as a messenger between DNA and proteins, which are mainly responsible for the form and function of living organisms. However, only 1.5% of a typical mammalian genome corresponds to protein-coding gene exons (this fraction rises to 25% when considering introns), with the rest corresponding to DNA that is either not transcribed at all or that gives rise to RNA molecules that are not translated into proteins (*3*–*5*). While much of this vast non-coding space is likely devoid of function, a topic that has been fiercely debated over the last decade (*6*–*11*), it is by now established that several classes of non-coding elements play important roles in many biological processes. The goal of this dissertation was to identify and characterize non-coding elements in mammalian genomes and to characterize their contributions to organ development and evolution.

1.1.1 The non-coding genome harbors a diverse set of elements

The 98.5% of the human genome that does not encode for protein sequences is far from homogeneous. Around half of the genome is covered by repeats (*3*, *12*), most of which are derived from transposable elements (TEs). Long and short interspersed nuclear elements (LINEs; SINEs) account for 21% and 15% of the human genome respectively, with an additional 9% covered by long terminal repeats (LTRs) (*12*). DNA transposons and tandem repeats contribute an additional 6%, bringing the total fraction of repeats in the human genome to ~51% (*12*). Most of these repeats are no longer active and likely correspond to decaying evolutionary relics of previous expansions, reflecting inefficient removal by natural selection rather than a benefit for the host (*12*, *13*). However, some of these repeats have been shown to play important roles in various biological processes. These include direct contributions to the host's gene regulatory networks (*13*, *14*) or simply the expansion of the available genomic sequence from which new non-coding elements can arise *de novo* during evolution (*15*).

Functional non-coding elements — regardless of whether derived from unique or repetitive DNA — can be broadly classified into two groups: 1) non-coding transcripts, where the RNA molecule plays some role in a cellular process, and 2) *cis*-regulatory elements (CREs), where the function lies in the DNA sequence itself (**Figure 1.1**). However, even this distinction should not be considered absolute. Many CREs are transcribed into short RNAs that might contribute to their function (*16*) and transcription of the CRE into long and spliced RNAs is associated with increased regulatory activity (*17*, *18*). On the other hand, the functionality of many non-coding RNAs relies on the act of their transcription (i.e., the biochemical activity of their DNA locus) rather than the RNA molecule itself (*19*, *20*).

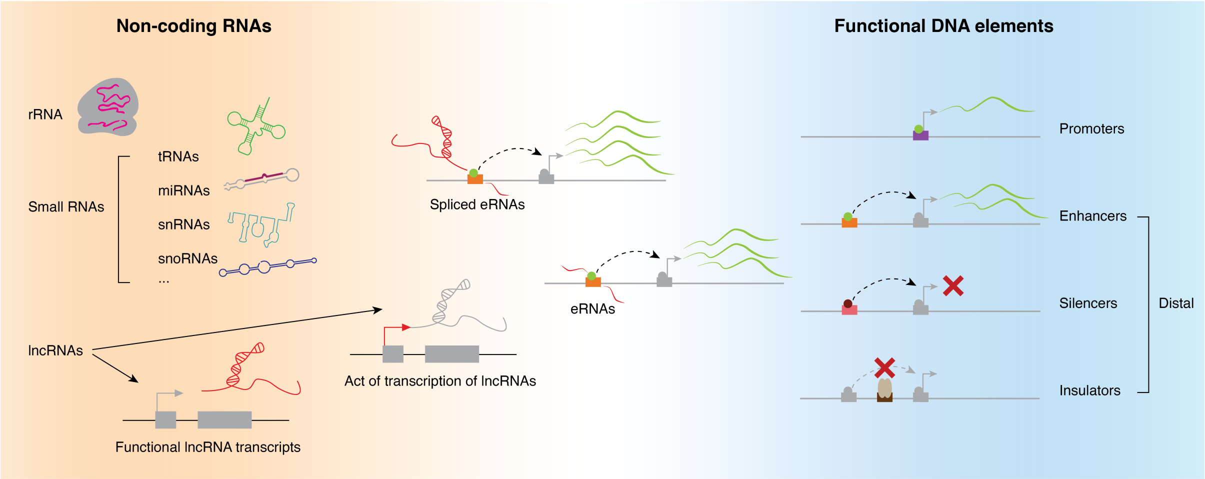


Figure 1.1: Major classes of non-coding elements in a typical mammalian genome. Transcribed enhancers (eRNAs) and lncRNAs functioning via the act of their transcription bridge the gap between non-coding transcripts (left) and CREs (right).

1.1.1.1 Major classes of non-coding transcripts

Ribosomal RNAs (rRNAs) are the most abundant non-coding transcripts, accounting for ~80% of all RNA molecules in a given cell, and are responsible for the catalytic activity of the ribosome during protein synthesis (*21*). Non-ribosomal non-coding transcripts can be further divided into long — spliced and longer than 200 nt — and small non-coding RNAs (*22*). Small non-coding RNAs are by far the better understood of the two groups. They include transcripts with housekeeping roles, such as transfer RNAs (tRNAs), which serve as an adapter between messenger RNA (mRNA) codons and amino acids during protein synthesis (*23*), small nuclear RNAs (snRNAs) that together with proteins catalyze the splicing of pre-mRNA in the nucleus (*24*), and small nucleolar RNAs (snoRNAs) that guide chemical modifications in other RNA molecules (*25*). Other groups of small RNAs are involved in the regulation of gene expression. These include microRNAs (miRNAs) and small interfering RNAs (siRNAs) which typically hybridize with the 3' untranslated region (UTR) of mRNA molecules to induce their degradation and/or translational repression, although they can also lead to direct transcriptional silencing in the nucleus (*26*, *27*). PIWI-interacting RNAs (piRNAs) are a specialized class of small RNAs that employ similar molecular mechanisms to specifically repress the transposition of TEs in the germline (*22*, *28*).

Despite this diversity in structure and function, small RNAs are much better understood than their longer counterparts, long non-coding RNAs (lncRNAs). These represent a large and highly heterogeneous group of non-coding transcripts (more than 95,000 human lncRNAs annotated in NONCODE v5) that do not belong to the classes described above. By convention, lncRNAs are defined as showing no coding potential, being long (longer than 200 nt) and spliced (*19*, *29*). However, even these three simple rules leave room for notable exceptions, such as the monoexonic *NORAD* (*30*, *31*) and the protein-coding gene *ASCC3* which can be repurposed into a functional non-coding transcript under specific conditions (*32*). Despite nearly two decades of intense study, only a handful of lncRNAs have been deeply characterized at the molecular level, and the differences observed in their mechanisms and functions leave little hope for generalizing the lessons learned from these "paradigms" (*33*). Furthermore, it is still unclear what fraction of mammalian lncRNAs could represent byproducts of transcriptional noise rather than having any functional consequences for the organism (*6*, *29*, *33*). To this end, the systematic characterization of lncRNAs at the genome scale can serve as an important first step to prioritize candidates for more detailed functional investigation (*20*, *29*). The first two parts of my thesis work focus on lncRNAs, which are discussed in greater detail below (1.1.2, 1.1.3).

1.1.1.2 Major classes of *cis*-regulatory elements

Non-coding elements that act via their DNA sequence, broadly referred to as CREs, also show considerable diversity, but share the same basic principle: They all contain one — or usually more — short DNA sequences that are recognized by DNA-binding proteins called transcription factors (TFs). TFs are recruited to these recognition sequences, often leading to further recruitment of additional factors through protein-protein interactions. Depending on the genomic position of the CRE, the kind of TFs that are recruited to its sequence, and the effect of this recruitment on gene expression, CREs can be classified into different groups. CREs overlapping with or proximal to a gene's transcription start site (TSS) typically recruit the core transcriptional machinery and are termed promoters (*34*, *35*). The rest of the CREs are collectively referred to as distal elements, and include enhancers, silencers and insulators. Enhancers are DNA elements that can activate gene expression across long distances, and independently from their orientation, by recruiting activating TFs which in turn come to close spatial proximity to the TSS (*36*, *37*). Silencers, considerably less understood than enhancers, work in a similar way to repress transcription by recruiting repressor TFs to their sequences

(*34*). Finally, insulators or boundary elements are involved in the regulation of the spatial architecture of the genome, recruiting architectural proteins such as CTCF and cohesin, which limit the interactions between genomic regions from opposite ends of the boundary (*34*, *38*). However, the distinction between these classes is not absolute. Promoters can act as enhancers, activating the expression of adjacent genes (*35*, *39*) and enhancers can be transcribed into typically short but occasionally longer RNAs (*16–18*, *40*). Additionally, the same CRE can recruit activating TFs in one context and repressing TFs in different contexts, shifting from an enhancer to a silencer (*41*, *42*). Even the same TF can have opposing effects on gene expression by interacting with different sets of co-factors (*43*). Ultimately, it is the nature of the collective TF machinery recruited to a CRE that determines its function.

The next four sections of this introduction focus on the non-coding elements examined during this work, lncRNAs (1.1.3, 1.1.4) and CREs — mainly enhancers and promoters (1.1.5), starting with the methods used to identify them at a genome-wide scale (1.1.2).

1.1.2 Genome-wide identification of non-coding elements

The systematic study of non-coding genomic elements is considerably more complicated than that of protein-coding genes. The simple rules underlying the genetic code behind protein synthesis allow the *in silico* identification of reasonably long open reading frames (ORFs) directly from the genomic sequence (*44*). Additionally, most protein-coding genes show high sequence constraint, allowing the extrapolation of findings from one species to homologous genes in other species (*44*). By contrast, the limited evolutionary conservation of most non-coding element classes (*45*, *46*) and our incomplete understanding of how their function is encoded in their sequence (*33*, *37*) severely limit the scope and efficiency of similar analyses. Instead, the identification of non-coding elements typically requires some genome-wide biochemical assay, which serves as a proxy for their activity.

In the case of non-coding transcripts, this assay typically involves the sequencing of RNA molecules (RNA-seq). The most popular approach (*47–49*) starts with the capture of polyadenylated RNA. RNA molecules are subsequently fragmented and reverse-transcribed into cDNA, which is then subjected to next-generation sequencing (*50*). Small modifications in the protocol can provide additional information about the strand of the original RNA molecule (*51*). This method allows the *de novo* identification of previously uncharacterized transcripts and the simultaneous quantification of their expression levels (*50*). However, the use of short reads limits the detection of the precise transcript boundaries and the characterization of differential isoform usage. Capturing the 5'-cap of RNA transcripts has been used to accurately define lncRNA transcription start sites (*52*), while recent methods that allow the sequencing of long RNA fragments (median length of 1-1.5 kb) have led to a more precise definition of lncRNA transcript models (*53*, *54*). However, short-read RNA-seq remains the most commonly used method for genome-wide transcriptomic profiling due to is cost efficiency and scalability.

The genome-wide identification of CREs is more complicated, as no single assay serves as a golden standard. Many studies have used chromatin immunoprecipitation followed by sequencing (ChIP-seq) to identify loci marked by specific histone modifications associated with promoter (e.g., H3K27ac and H3K4me3) or enhancer (e.g., H3K27ac and H3K4me1) activity (*35*, *46*, *55*). More recent methods such as CUT&RUN also make use of antibodies to target histones or TFs but don't require precipitation, leading to an improved signal-to-noise ratio (*56*). An alternative to antibody-based methods is to identify open chromatin regions, as TF binding on CREs leads to nucleosome displacement and increased chromatin accessibility (*57*). In such methods, open chromatin regions are identified based on how accessible they are to enzymes such as a DNase (DNase-hypersensitivity assay) or a transposase (Assay for Transposase-Accessible Chromatin; ATAC-seq) (*57*). ATAC-seq has become especially popular over the last years due to the simplicity of its protocol and its high sensitivity (*57*, *58*). However, it is important to acknowledge that open chromatin only serves as a proxy for regulatory activity, and that it provides no additional information regarding the class of the CRE (e.g., promoter, enhancer or silencer). Thus, inferring CRE activity with high confidence typically requires additional evidence from massively parallel reporter assays (MPRAs) or CRISPR interference/activation screens, which can assess the ability of the sequence to affect gene expression in a given biological context (*35*).

Genome-wide assays such as RNA-seq and ATAC-seq have revolutionized the systematic profiling of non-coding elements, and more generally the study of gene expression and gene regulation. However, a major limitation of these assays has been the need for large amounts of biological material acquired from thousands or even millions of cells (*59*). Thus, in the context of profiling heterogeneous samples, such as developing organs, these assays are only able to provide an ensemble view, averaged across many cell types and states that exist in different proportions. This limits the scope of the data in many ways (*59*), but is especially problematic for non-coding elements, such as lncRNAs or CREs, which are highly cell type-specific (*37*, *47*, *60*–*62*). Relying on bulk organ profiling, not only do we remain oblivious about the precise cell type in which a CRE is active, but we even risk to miss identifying it altogether, especially if it is active in a rare cell type.

However, recent improvements in sequencing technology have enabled the profiling of gene expression and chromatin accessibility in single cells (*59*). Single-cell methods benefit from advances in microfluidic technologies, which allow — at least part of the reaction — to take place within liquid droplets (or in less commonly used protocols, in individual wells), each containing a single cell or nucleus (*59*). Sequences arising from the same cell are marked with unique barcodes, allowing them to be grouped together after sequencing (*63*). The molecular profile of each cell can be subsequently used to associate that cell with a distinct cell type and state (*64*). Modifications of this basic protocol have enabled the profiling of gene expression (e.g., snRNA-seq) (*63*), chromatin accessibility (e.g., snATAC-seq) (*65–67*) and more recently the measurement of both modalities from the same nucleus (*68*, *69*).

1.1.3 Genomic and evolutionary features of lncRNAs

The large-scale investigation of lncRNAs through RNA-seq has provided important insights into their genomic features and evolutionary histories. As discussed above, lncRNAs represent a highly heterogeneous group defined as long (longer than 200 nt), spliced and non-coding transcripts (*19*). They resemble mRNAs in that they are typically capped and polyadenylated (*19*). However, lncRNA transcripts are shorter (*70*) and their splicing is less efficient (*71*). Their promoters are simpler than those of protein-coding genes and contain fewer TF binding sites (*71*). This is also reflected in their lower and more context-specific expression (*47*, *48*). The low expression of lncRNAs is often used as an argument against their functional relevance (*33*, *72*). However, expression levels of lncRNAs might be underestimated by bulk RNA-seq because of their high cell type-specificity (*61*). Additionally, even when lowly expressed, lncRNAs can become highly concentrated in individual subcellular compartments associated with specific biological processes (*72*). Another argument supporting the notion that many lncRNAs might correspond to transcriptional noise is that the majority of lncRNAs is exclusively expressed in the mammalian testis (*47*, *48*, *73*). Most of these testis-specific lncRNAs are likely byproducts of a pervasive chromatin environment that also allows the transcription of short intergenic sequences (*73*, *74*), although individual lncRNAs with important functions in spermatogenesis have also been described (*75*). Due to their heterogeneity, lncRNAs are often classified based on their position (overlapping, intergenic) and orientation (sense/antisense, bidirectional) compared to the nearest protein-coding gene (*76*). However, such a classification scheme offers little information regarding their potential functions (*76*).

Another major difference between lncRNAs and mRNA genes concerns their evolutionary conservation. As expected by the relaxed constraints associated with their non-coding nature, lncRNA sequences diverge much faster than those of protein-coding genes (*47–49*, *77*), although they are on average more constrained than random intergenic regions (*47*, *48*). Similarly, homologous lncRNAs show lower expression similarity compared to protein-coding genes (*48*, *49*, *77*). The rapid evolutionary turnover of lncRNAs is mainly facilitated by TE insertions (*45*, *78*, *79*), stabilization of short transcripts from bidirectional promoters or enhancers (*45*, *80*, *81*) and the *de novo* exaptation of previously non-transcribed regions (*45*). An additional source of lncRNAs is the pseudogenization of protein-coding genes and eventual acquisition of a new function by the transcript (*82*), with the most prominent case being the emergence of the regulator of X-chromosome inactivation, *XIST* in eutherian mammals (*83*). Emergence through gene duplication, the main mechanism underlying the origination of new protein-coding genes (*15*) appears to be rare based on the low within-species sequence similarity observed for lncRNAs (*45*).

1.1.4 The quest for functional lncRNAs

The holy-grail of lncRNA biology is to distinguish between functional non-coding transcripts and transcriptional noise (*6*, *33*). The last decade has witnessed a quest for the systematic identification of functional lncRNAs.

Evolutionary conservation represents one of the strongest lines of evidence for the functionality of a genomic locus (*11*, *29*, *33*, *45*). However, since lncRNAs don't encode for proteins, their sequences are subject to much less constraint than that observed for protein-coding genes, suggesting that even homologous lncRNAs with conserved functions between species can quickly diverge beyond the level of detectable sequence similarity (*45*, *84*). Additionally, posing conservation as a prerequisite for functionality *a priori* dismisses thousands of lineage-specific lncRNAs, some of which could be associated with evolutionary innovation (*45*). Analysis of within-species, population-level constraints could provide additional insights (*33*) but our incomplete understanding of how function is encoded in non-coding sequence (*85*) further complicates such endeavors. Thus, requiring evidence for evolutionary or population-level sequence constraint could underestimate the functional relevance of lncRNAs.

Another approach is to search for overlaps between lncRNA annotations with disease-associated genetic polymorphism (*52*). However, due to genetic linkage, the causal variant can be several kb away from the locus identified as significantly associated with the trait (*86*). Additionally, the causal variant might be associated with the function of a different overlapping genomic feature, such as a CRE (*33*). Thus, this approach represents a very relaxed definition of functionality.

The ultimate proof for lncRNA functionality requires extensive experimental investigation, typically by examining the phenotypic consequences upon perturbation of the lncRNA locus or its expression (*20*, *33*). The main limitation of this approach is its scale. To date, only a few lncRNAs have been deeply characterized at the molecular level. These include *XIST*, which is essential for X-chromosome inactivation in eutherian mammals (*87–90*), *HOTAIR*, which is involved in the control of the expression of the *HOXD* cluster (*91*, *92*), and *NORAD*, a critical regulator of genomic stability (*30*, *31*). Each of these lncRNAs has been intensely studied by several groups for many years, an approach that is not feasible for the interrogation of thousands of transcripts, especially under the expectation that many of them are completely devoid of any function. More recently, large-scale perturbation experiments of lncRNA loci have become feasible (*93*, *94*). However, even with this increase in the number of assayed loci, each experiment can only measure the effect of a perturbation in a specific context (typically a cell line) and for a limited set of phenotypes (most often cellular growth and proliferation). Thus, given the high context specificity of lncRNAs, an exhaustive functional investigation would require assaying a large number of conditions for a large number of potentially very subtle phenotypes.

Finally, expression features can also be used as a starting point to investigate lncRNA functionality (*47*, *48*). Gene expression profiling through RNA-seq can be applied at a genome-wide scale and across a much wider range of conditions compared to currently available perturbation assays. Although inadequate to prove functionality in the absence of additional data, gene expression features such as the reproducibility and stability of a lncRNA transcript (i.e., whether it can be found across multiple individuals and/or conditions) and its dynamic expression during developmental processes can be used to enrich for putatively functional lncRNAs (*29*, *33*). Additionally, the spatiotemporal profiles of lncRNAs can generate testable hypotheses about plausible functions (a process commonly referred to as “guilt-by-association”), reducing the number of putative phenotypes that need to be considered experimentally.

1.1.5 Molecular mechanisms and evolutionary histories of CREs

In contrast to lncRNAs, the main principles behind the action of CREs are much better understood. As described in 1.1.1.2, CREs contain short DNA sequences (motifs) that are recognized and bound by TFs, which typically recruit additional proteins and come in close proximity to the TSS, leading to the activation or repression of gene expression in that locus (*37*). However, the sequence basis of CRE functionality is far from understood. TF motifs are typically short (6-10 nt) and degenerate (i.e., TF binding can occur even without a perfect motif) (*36*). The same TF can recognize different motifs and the same motif can be bound by many TFs (often from the same family) (*95*). Additionally, many sequences containing a TF motif in the genome are not bound *in vivo*, whereas TF binding can be observed in sites that don't contain a canonical motif (*36*). These discrepancies can be attributed to co-operative interactions between TFs: the absence of a required partner can preclude binding despite the presence of a canonical motif whereas a TF can be recruited to sites without its motif through protein-protein interactions with a partner (*36*, *37*). Finally, most CREs show considerable flexibility in their sequence grammar (i.e., the distance, orientation and order of their TF motifs) (*37*) — although exceptions showing rigid syntax requirements have been described (*96*). This complexity, both in terms of individual motif instances and their interactions, has led to the "futility theorem" by Wasserman and Sandelin, stating that the vast majority (99.9%) of TF binding sites predicted from DNA sequence alone would be false positives (*97*).

However, recent developments in machine learning technologies, primarily utilizing deep learning models, provide new tools to approach this problem. Convolutional neural networks, most famous for their applications in image recognition (*98*), have been successfully used to predict chromatin accessibility (*99–102*), TF binding (*103*), enhancer activity (*104*) and even gene expression (*105*) from DNA sequence alone. More importantly, interpretation of these models through dedicated algorithms (*106–108*) allows the extraction of the sequence features contributing to the model predictions (i.e., the TF motifs), providing insights into CRE grammar. However, despite the great successes of these models, current implementations are still not accurate enough to be applied at a genome-wide scale, and typically have to be limited to pre-existing CRE annotations obtained through other assays (e.g., ATAC-seq, ChIP-seq).

Another challenge when studying CREs is that enhancers can act over great genomic distances (even up to 1 Mb), complicating the identification of their target genes (*109*, *110*). Despite spanning such large distances in the linear genome, enhancer-promoter interactions require them to come to close spatial proximity in the nucleus, although the precise molecular mechanisms underlying these processes are still under investigation (*109–111*). Such interactions are typically constrained within the same topologically associating domain (TAD). TADs are defined as self-interacting genomic regions showing a high degree of chromatin interactions within them and marked by boundary elements in both ends, which separate them from other adjacent TADs (*109, 110*). TAD boundaries are generally considered to be relatively stable across cell types, developmental stages and even across species (*112*), although recent studies suggest that they might be more dynamic than previously thought (*109, 113–116*).

As discussed above (1.1.1.2), the distinction between promoters and enhancers appears to be less absolute than previously thought, with many enhancers being transcribed (*16–18, 40*) and many promoters influencing the expression levels of adjacent genes (*35, 39*). However, at the level of population averages, promoters and enhancers differ substantially in their genomic and evolutionary properties. Promoters are overall more GC-rich (*35*), contain more direct binding sites for general TFs, show largely invariant chromatin accessibility across cell types and developmental stages and are more conserved during evolution (*46, 117*). By contrast, enhancers are more context-specific (*55*) and show a faster evolutionary turnover (*46, 117*). Most new enhancers emerge *de novo*, from previously inactive DNA (*46*) or by incorporating TE sequences (*14, 118, 119*). However, recent studies have also highlighted the repurposing of the spatiotemporal activity of pre-existing enhancers as another prominent mechanism of regulatory evolution (*119, 120*).

Similar to CREs, TF binding sites diverge rapidly during evolution (*121–124*). However, in stark contrast to their individual binding sites in the genome, the activity of TFs that are central to cell type-defining gene expression programs is largely conserved across vertebrates (*125*), as are their DNA binding domains and the sequence motifs recognized by them (*124, 126*). Thus, despite the rapid turnover of individual CREs, the sequence code determining their spatiotemporal activity is largely preserved during evolution.

1.2 Non-coding elements are central to developmental processes

During animal development, a single fertilized egg gives rise to hundreds of different cell types, such as neurons, muscle cells and the gut epithelium. Each of these cell types has a different form and set of functions, which are encoded in the genome (*64, 125*). Yet, with very few exceptions, such as lymphoid and meiotic germ cells, all cells in a single organism share the same DNA (*127*). Despite being based on the same genome, each cell type is marked by the expression of a distinct set of genes, which determines its unique properties (*64, 125*). However, many of the genes regulating the development of organs and cell types are actually involved in multiple process, i.e., they are pleiotropic (*128*). As an example, in mammals, the Sonic hedgehog (Shh) signaling pathway regulates processes as distinct as the development of the cerebellum, the formation of the limb bud, the patterning of the gut tube and the shaping of digits (*128*, *129*). Similarly, two central regulators of cell fate specification in the cerebellum, *Atoh1* and *Ptf1a* (*130*), are also involved in the development of the inner ear (*131*), skin (*132*) and intestine (*133*) (*Atoh1*), and of the pancreas (*134*) and the retina (*135*) (*Ptf1a*). Thus, rather than relying on a large number of context-specific genes, animals are able to generate a diversity of cell types and organs by reusing the same genes in different combinations (*125*, *128*).

This process depends on the precise spatiotemporal regulation of the expression of each gene, which — as discussed above — relies on non-coding elements, such as CREs and regulatory RNAs (*19*, *36*, *37*). It is the context-specific activity of these non-coding elements that facilitates a system of modular control of gene expression (*36*, *37*). Pleiotropic genes are often controlled by a large set of CREs, each tasked with regulating the expression in a specific cell type and developmental window (*36*, *37*) (**Figure 1.2**). The specific activity of that CRE is in turn controlled by the activity of the TFs whose motifs are present in the CRE sequence, which are themselves controlled by their own CREs (*36*, *37*, *125*, *128*). Ultimately, the cell type-specific activity of these interconnected gene regulatory networks is determined by initial asymmetries in the molecular composition of the early embryo and/or by differences in the cellular microenvironment (*136*) that become established and amplified as differential TF activity between germinal zones (*68*) and eventually cellular lineages (*137*, *138*). Over the course of development, these lineage-determining networks can be further refined by incorporating external signals (*139*, *140*).

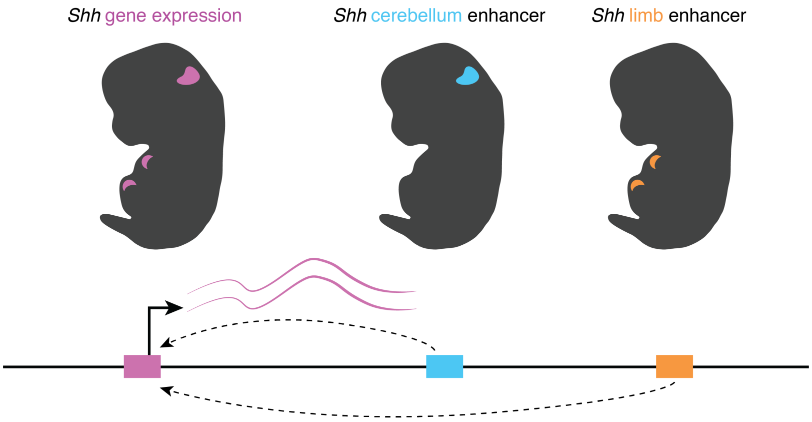


Figure 1.2: Modular control of developmental gene expression. The cell type-specific activities of multiple distal CREs (right), which can act over large distances in the linear genome, add up to the pleiotropic expression patterns of developmental genes (left). *Schematic illustration not reflecting the precise genomic positions and spatiotemporal activities of the illustrated genes and CREs.*

The modular nature of gene expression regulation can be best illustrated by investigating the phenotypic consequences of perturbations in non-coding regulatory sequences. For example, while the deletion of the *Shh* gene in mouse leads to multiple severe organ malformations and embryonic lethality (*141*), homozygous deletion of its enhancer ZRS/MFCS1 only leads to limbless but otherwise healthy mice (*142*). Similarly, coding mutations in *PTF1A* in humans affect both the pancreas and the cerebellum, whereas mutations in an enhancer region ~25 kb downstream of the gene only lead to pancreas agenesis without any neurological phenotype (*143*, *144*). Although this simple model of modular regulation of gene expression is complicated by enhancers showing synergistic interactions (*145*) or redundancy in their function (*146*, *147*), most CREs have been shown to only be active in a subset of organs (*119*, *148*, *149*), and developmental stages (*55*, *150*, *151*). Finally, some of the most exciting developmental processes, such as cell fate specification, differentiation and migration, take place at the level of single cells. Although recent studies examined the chromatin accessibility profiles of single cells in various contexts (*62*, *152–157*), a comprehensive characterization of the CRE dynamics of single cells during the entire development of a mammalian organ was missing at the time I conducted my thesis work.

1.3 Non-coding elements contribute to evolutionary innovation

Unlike the emergence of diverse cell types from the same genetic material during development, each species has its own genome that could — in theory — contain large numbers of genes specific to that species. In fact, Ernst Mayr, one of the fathers of the Modern Synthesis of Evolutionary Biology, predicted in 1963 that "the search for homologous genes is quite futile except in very close relatives" (*158*). However, the unexpected discovery that the same set of Homeobox (Hox) genes responsible for anterior-posterior patterning in the fly (*159*) are present and expressed in similar patterns in almost every bilateral animal examined (*160*) suggested that at least some genes could be retained for more than 600 million years. During the last two decades, the sequencing of many mammalian genomes revealed that the number of evolutionary conserved genes is surprisingly large. For example, 80% of the protein-coding genes in mouse have an ortholog in human (*4*). Additionally, these genes have retained a high sequence similarity in their coding sequence — e.g., 78% median amino acid similarity between human and mouse 1:1 orthologous protein-coding gene pairs (*4*). Thus, while the emergence of new genes (*15*, *161–163*), and mutations in protein-coding sequences (*164*, *165*), undoubtedly contribute to evolutionary innovation, most of the phenotypic differences observed between species are likely explained by changes in the spatiotemporal expression patterns of conserved protein-coding genes (*128*, *166*, *167*).

Evolutionary innovation through changes in the expression rather than the sequence of protein-coding genes also alleviates the constraints associated with their pleiotropic nature (*128*, *167*). A sequence change affecting the final protein product would influence all processes regulated by that gene and is less likely to be beneficial. By contrast, due to the modular nature of gene expression regulation, a mutation in a CRE can selectively modify the function of the target gene in a single context, leaving its remaining functions unaffected (*128*, *167*). The loss of limbs in snakes serves as an illustrative example **(Figure 1.3)**. A recent study showed that this loss is associated with mutations in the *Shh* enhancer ZRS (*168*) — the deletion of which leads to limbless but otherwise healthy mice, as discussed in 1.2. These mutations ablate enhancer activity and *Shh* expression in the prospective developing limb bud, potentially representing one of the (likely many) genomic changes that led to this evolutionary innovation (*168*). However, snakes are still able to develop other structures that depend on *Shh* function, such as the cerebellum.

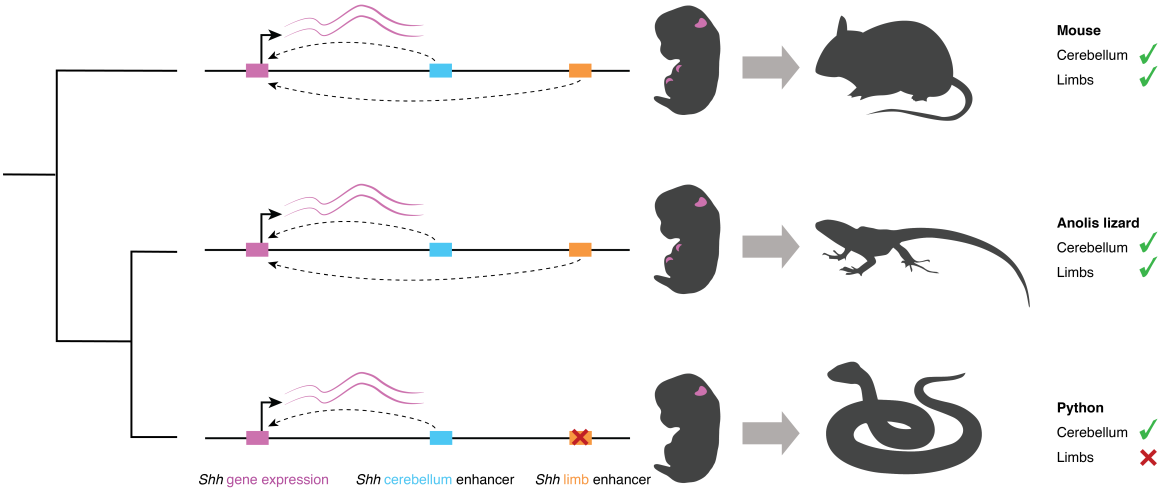


Figure 1.3: The modular control of gene expression facilitates context-specific evolutionary innovation. Evolutionary changes in cell type-specific CREs can affect the expression and function of a developmental gene in a specific context without interfering with other processes controlled by different CREs. *Schematic illustration not reflecting the precise genomic positions and spatiotemporal activities of the illustrated genes and CREs.*

During the last decade, large-scale transcriptomic studies have pursued the systematic identification of protein-coding genes with significant expression changes across mammalian species, which serve as candidates for driving phenotypic innovations (*5*, *167*). These include comparisons of gene expression in adult (*169–171*) and developing (*172–175*) organs. More recently, single-cell RNA-seq methods have enabled the investigation of these gene expression changes at the level of cell types (*74*, *176–179*). Although homologous organs from different species show higher transcriptomic similarity than different organs from the same species (*167*, *172*), expression divergence of conserved genes appears to be pervasive, with 50% of orthologs between human and mouse showing a radical expression change during the development of at least one organ (*172*). In parallel, comparative studies of gene regulation have identified thousands of lineage-specific or repurposed CREs (*46*, *119*, *120*, *180–182*). However, even though integrative evolutionary analyses of the transcriptome and epigenome have recently emerged (*117*, *119*, *157*), the systematic identification of CREs associated with gene expression evolution remains limited, primarily due to the challenges in assigning CREs to their target genes and in quantitatively understanding how regulatory input relates to transcriptional output.

1.4 Interplay between development and evolution

The two previous chapters illustrate some interesting parallels between development and evolution. Both processes describe a change in form over time, which in both cases is mainly driven by the usage of different combinations of the same gene set. In both processes, non-coding regulatory elements are central to regulating which genes will be expressed in a given cell type and stage, ultimately determining how they develop and evolve.

Besides these parallels, the two processes also directly influence each other. The previous chapter already provided an example of how small molecular differences in an early developmental stage (e.g., a difference in the activity of an enhancer in the prospective limb bud) can be amplified into a striking phenotypic change in the adult (e.g., the loss of limbs in snakes). More generally, evolutionary innovation can be facilitated by changes in where (heterotopy) and when (heterochrony) developmental genes are expressed (*128*). Thus, even if the phenotypic consequences of such evolutionary innovations are more prominent in adults, the identification of their molecular basis often requires the comparison of developmental processes (*128*). Additional examples include numerous evolutionary innovations in *Drosophila* species (*128*), but also the expansion of the human brain through the expression of human-specific genes (*161*, *163*) and overall delay of gene expression programs (*183*, *184*) during development.

But developmental processes also constrain evolution. Already during the 19th century, the comparative embryologist Karl Ernst von Baer observed temporal differences in the degree of morphological similarity between embryos from different species (*185*). Although the very first steps of development (fertilization and first divisions) show large differences even across mammals, embryos converge to a similar morphology around late gastrulation and early organogenesis, when most major organs are molecularly defined (*185*, *186*). From this stage on, also termed the phylotypic period, morphological differences between species become increasingly pronounced during development (*185*, *186*). These observations gave rise to an "hourglass model" of embryonic evolution, according to which evolutionary divergence is most prominent during the earliest and latest stages of animal development (*185*, *186*). Recent studies have also recapitulated these models at the level of protein-coding gene expression (*172*, *187*) and enhancer activity (*55*, *155*, *188*).

Collectively, development and evolution are closely intertwined. Early organ development — which coincides with the end of the phylotypic period — is expected to show the smallest degree of evolutionary divergence due to developmental constraints. However, these relatively few differences are likely to have the strongest impact on the phenotypic divergence observed between species in later stages.

1.5 Mechanisms of mammalian X-chromosome inactivation

The second part of my dissertation work focused on a newly identified lncRNA as a candidate for regulating X-chromosome inactivation in marsupials. This section provides a brief — and by no means exhaustive — introduction to mammalian sex chromosome evolution, the need for dosage compensation and the molecular innovations that emerged to tackle this problem.

In most therian mammals, including human, mouse and the marsupial opossum, sex is determined by an XY sex chromosome system, which emerged around 180 million years ago, after the split from egg-laying monotreme mammals (*189*–*191*). This process started with the emergence of the male determining gene *SRY* on an autosome in the ancestor of all therian mammals (*192*, *193*). It is hypothesized that the subsequent emergence of sexually antagonistic mutations (beneficial for one sex but deleterious for the other) and chromosomal inversions in the proximity of *SRY* eventually led to a local arrest of meiotic recombination between chromosomes containing *SRY* (proto-Y) and those that did not (proto-X) (*194*, *195*). Only a small part of chromosome Y, termed "pseudoautosomal region" has retained its ability to recombine with the X in extant therian mammals (*195*, *196*). The loss of recombination for proto-Y (proto-X could still recombine in females) resulted in reduced selection as deleterious mutations could "hitchhike" on adjacent beneficial mutations, eventually leading to many protein-coding genes being lost or migrating to the autosomes (*189*, *191*, *194*, *195*). In modern humans, chromosome

Y only contains 45 protein-coding genes, as opposed to 856 found on chromosome X *(source: Ensembl 107)*.

The loss of many genes from chromosome Y created a new problem: an imbalance in gene expression across sexes. Most genes on the X are present in two copies in females but only in one copy in males. To achieve equal gene expression levels between male and female individuals — although not necessarily matching the ancestral levels (*197*) — therian mammals evolved new mechanisms to inactivate one X chromosome in females (*198*). This process evolved independently in eutherian mammals (e.g., human and mouse) and marsupials (e.g., opossum) (*199*). The two mechanisms share similarities but also show important differences. In eutherians, X-chromosome inactivation depends on a lncRNA, *XIST*, which in females gets randomly activated in one of the two X chromosomes, marking that copy for inactivation (*198*). The precise mechanism differs between human and mouse but generally involves coating of the inactive X by *XIST* (*198*). Certain repeat regions on *XIST* are able to recruit proteins that lead to heterochromatin spreading and eventually to the condensation of the inactive X into a structure called the Barr body (*90, 198, 200*). Most of the genes found on chromosome X are silenced by this process, although some escape and remain expressed from both chromosomes (*198*).

X-chromosome inactivation in marsupials is far less studied than in eutherians and almost all our knowledge comes from the opossum *Monodelphis domestica*. Interestingly, this process also involves a lncRNA, *RSX*, which however is not homologous to *XIST* (*201*). Similar to *XIST, RSX* is quite long, enriched for repeats and can induce silencing of adjacent genes (*201*). However, unlike the random X-chromosome inactivation observed in eutherians, females in marsupials (or at least in opossums) specifically inactivate the paternal X through a yet unknown mechanism (*199*). Additionally, while the repressive histone mark H3K27me3 constitutively marks the eutherian Barr body, in marsupials it can mainly be observed during the late S and early G2 cell cycle phases (*202*). Collectively, although the eutherian and marsupial mammalian lineages share the same sex chromosome system, they have evolved independent, albeit to some degree analogous, strategies to deal with the challenges associated with that system (*199*).

1.6 Development and evolution of the mammalian cerebellum

The final two parts of my dissertation work focused on the developing mammalian cerebellum, as a case study of how non-coding regulatory elements guide developmental processes and evolutionary innovation in mammalian organs. This chapter provides a brief introduction into the functions, development and evolution of the cerebellum.

1.6.1 Functions, connectivity and development of the mammalian cerebellum

The cerebellum (Latin for "little brain") is a hindbrain structure primarily known for motor learning and coordination (*130, 203*). However, its role in other complex behaviors, such as language, memory and integration of sensory information, is being increasingly appreciated (*130, 203*). Accordingly, abnormalities in cerebellar function can lead to motor-related conditions, such as spinocerebellar ataxias, but also to disorders like autism and schizophrenia (*130, 203*). Additionally, the cerebellum harbors some of the deadliest pediatric brain tumors, including medulloblastoma, ependymoma and pilocytic astrocytoma (*204*). Thus, understanding the molecular processes underlying normal cerebellar development could help with the diagnosis and treatment of several human diseases.

The cerebellum is present in all jawed vertebrates and its main circuit has been conserved during evolution (*205*). This circuit is centered on the inhibitory Purkinje cells, characterized by their extensive network of dendrites that were first observed in 1899 by one of the fathers of neuroscience, Ramón y Cajal (*206*). Purkinje cells integrate signals that arrive to the cerebellum from other areas of the central nervous system and project outwards to the neurons of the cerebellar deep nuclei (DN) which are embedded in the white matter (*206*). Information is then transmitted from the DN neurons to the rest of the brain and spinal cord (*206*). Some of the afferent (incoming) signals directly reach the Purkinje cells while others are modulated by a diverse set of interneurons (*206*). These include the small but extremely abundant excitatory granule cells, which account for 50-80% of all neurons in a typical mammalian brain, unipolar brush cells (UBCs), and a series of inhibitory GABAergic interneurons (*206*). The spatial organization of these neuron types in the cerebellar cortex is highly structured and also conserved during amniote evolution (*205, 206*). Granule cells reside in the innermost layer, termed the granule cell layer (GCL) whereas their axons extend to the outermost layer, the molecular layer (ML), which also contains the Purkinje cell dendrites (*206*). In-between these two layers lies the Purkinje cell layer (PCL) containing the Purkinje cell somata, which are arranged in a single layer (monolayer) (*206*). UBCs are also found in the GCL, whereas all three layers are populated by morphologically and transcriptomically distinct subclasses of GABAergic interneurons (*206*). Non-neuronal cell types such as the microglia (the immune cells of the brain), the Bergmann glia (a specialized form of astrocytes in the PCL), parenchymal astrocytes and oligodendrocytes complete the cellular make-up of the mature cerebellum (*206*).

The emergence of these cell types during the development of the cerebellum depends on a spatially and temporally restricted mode of cell fate specification (*130, 205*). During embryonic development, two germinal zones, the ventricular zone and the rhombic lip give rise to GABAergic and glutamatergic neurons, respectively (*130, 205*). However, different neuron types are generated by each germinal zone in different developmental windows **(Figure 1.4)**. GABAergic DN neurons, Purkinje cells and GABAergic interneurons are sequentially derived from progenitors of the ventricular zone, whereas rhombic lip progenitors give rise to glutamatergic DN neurons, followed by granule cell progenitors and UBCs (*130, 205*). Some of the key developmental regulators of this process are already well established. The ventricular zone and rhombic lip germinal zones are marked by the expression of the TF genes *Atoh1* and *Ptf1a* respectively, which are also essential for their formation (*130, 205*). The temporal switch from Purkinje cells to GABAergic interneurons is mediated by a transition in ventricular zone progenitors from the expression of *Gsx1* to *Olig2* (*130, 205*).

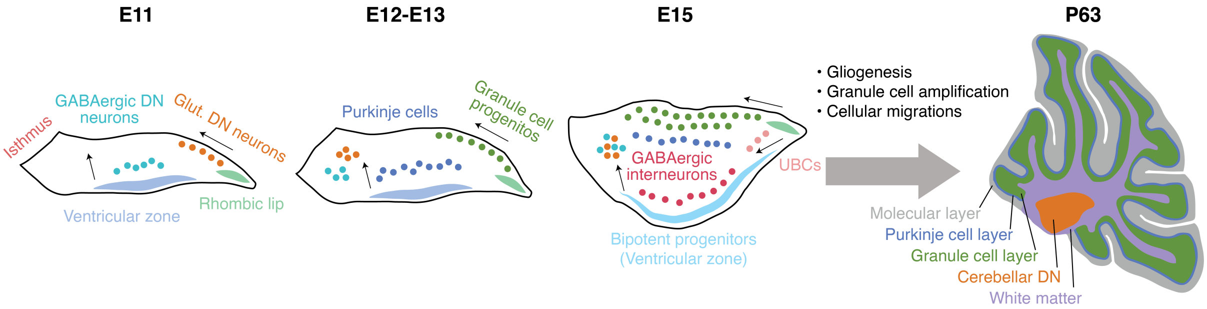


Figure 1.4: Spatiotemporal control of cell fate specification in the developing mammalian cerebellum. Two spatially defined germinal zones (ventricular zone and rhombic lip) generate distinct cell types during consecutive stages of embryonic development (left). Additional processes in perinatal and postnatal development, such as gliogenesis and cell migration, lead to the final form of the mature cerebellum (right). *Cerebelli icons (not at scale) were modified from an original design by Dr. Mari Sepp.*

Gliogenesis in the cerebellum mainly generates astrocytes as oligodendrocytes and microglia mostly migrate into the cerebellum from other regions (*130, 205*). The developing cerebellum hosts two progenitor populations that give rise to different astrocytes. Gliogenic progenitors detach from the ventricle, migrate to the PCL, and produce parenchymal astrocytes of the GCL and the Bergmann glia (*130, 207*). Another group of progenitors migrates to the prospective white matter, where they continue to give rise to GABAergic interneurons but also produce parenchymal astrocytes of the white matter (*130, 207, 208*). Due to their ability to generate both GABAergic interneurons and astrocytes, these progenitors are termed bipotent (*130, 208*).

In parallel to gliogenesis, the postnatal period of cerebellum development in mammals is marked by the secondary amplification of granule cell progenitors, a process dependent on *Shh* signaling (*130, 205*). Granule cell progenitors proliferate in a transient zone termed the external granule cell layer (EGL), while postmitotic differentiating granule cells gradually migrate inwards to the final location in the GCL (also referred to as internal granule cell layer in contrast to the EGL) (*130, 205*). This process extends to the second postnatal week in mice and up to two years after birth in humans (*130, 205*).

Although the development of the cerebellum is well described at the anatomical level, the identification of the major regulators underlying these developmental processes had until recently been limited to single-gene investigations (*130*). By the time this work was conducted, scRNA-seq studies had provided genome-wide insights into the spatiotemporal dynamics of gene expression associated with the development of the cerebellum (*179, 209–212*), but the regulatory basis of these processes remained largely unexplored.

1.6.2 Evolutionary innovations in the cerebellum

Despite the conservation of its main circuit, the cerebellum is also marked by a series of evolutionary innovations, varying greatly in its size and morphology across vertebrates (*205*). One major evolutionary shift is associated with differences in the size of the granule cell layer (*205, 213, 214*). The transient developmental structure of the EGL is absent outside of tetrapods and non-proliferative in amphibians (*213, 214*). Thus, the secondary amplification of granule cell progenitors leading to the extreme abundance of granule cells in the mature mammalian cerebellum appears to be an amniote-specific innovation, which was facilitated by a gain in the expression of the *Shh* gene in Purkinje cells (*214*).

Another major evolutionary innovation concerns the expansion in the number of cerebellar deep nuclei, from a single pair outside of amniotes, to two pairs in reptiles and birds, and three pairs in mammals (*215*). This was recently attributed to the duplication and subsequent divergence of an entire set of neuron types that were all present in the ancestral nucleus (*215*). Furthermore, the lateral cerebellar nucleus is markedly enlarged in humans, which was mediated by the expansion of a single class of excitatory neurons (*215*).

A third innovation is related to the overall size of the cerebellum. Most studies of mammalian brain evolution have focused on the neocortex, which is specific to mammals and has expanded greatly in primates (*216, 217*). However, cerebellar volume and neuron numbers appear to have expanded in parallel to the neocortex, keeping the ratio of cerebellar to cortical neurons (4:1) stable during mammalian evolution (*218, 219*). Furthermore, great apes, including humans, deviate from this pattern, showing an even greater expansion of the cerebellar volume compared to what would be predicted from their already expanded cerebral cortex (*220, 221*). This expansion might have been facilitated by the emergence of a subventricular zone of basal progenitors in primates (*222*), similar to what has been observed for the cortex (*216*, *217*). Additionally, the rhombic lip germinal zone has been expanded in both space and time in humans compared to macaques and mice, allowing for an increased production of neurons (*222*). Although the molecular basis of these expansions remains elusive, a large number of transcriptomic differences have been observed between the developing human and mouse cerebellum (*172*, *179*).

Collectively, thanks to its relatively simple and conserved cytoarchitecture, with unambiguously homologous and yet molecularly divergent cell types, the mammalian cerebellum is an ideal system to study how non-coding regulatory elements control gene expression programs during development and how changes in these programs can lead to evolutionary innovation.

2. Aims

This dissertation is comprised of four parts, each with its own scope and objectives. However, they all revolve around the following central aims:

- Identifying non-coding elements active during mammalian organ development.
- Characterizing the contributions of these non-coding elements during development by associating them with specific genes and processes based on their spatiotemporal activity.
- Investigating the evolutionary histories of non-coding elements and relating them to their activity across organs, cell types and developmental stages.
- Exploring how the interplay between development and evolution affects the non-coding fraction of mammalian genomes.
- Understanding how changes in the molecular profiles and/or proportions of cell types shape the developmental patterns observed at the level of whole organs.
- Identifying non-coding elements involved in evolutionary innovations, such as the emergence of new molecular mechanisms (3.1.2) or the divergent expression of conserved protein-coding genes (3.1.4).

3. Results

3.1 LncRNAs in mammalian organ development

The goal of this part of my thesis work was to explore the overall contribution of lncRNAs to the developmental programs of mammalian organs, to identify candidates for further experimental characterization and to provide insights into their putative functions. To this end, I utilized a large-scale RNA-seq dataset covering the development of seven major organs (forebrain/cerebrum — hereafter referred to as "brain" — cerebellum, heart, kidney, liver, ovary and testis) from early organogenesis to adulthood across six mammals (human, rhesus macaque, mouse, rat, rabbit, opossum) and a bird (chicken) (*172*). This project was supervised by Prof. Dr. Henrik Kaessmann and Dr. Margarida Cardoso-Moreira and the main findings were published in *Sarropoulos et al. 2019* (*223*). Initial exploratory analysis, primarily focusing on section 3.1.2, was performed in the framework of my MSc thesis (*224*). However, even the analyses presented in that section have been substantially revised and extended during my PhD work.

3.1.1 Atlases of lncRNAs expressed in mammalian organ development

Using this dataset, Dr. Ray Marin, a collaborator in this project, identified lncRNAs within each species as spliced transcripts that are longer than 200 nt and that show no evidence for protein-coding potential. He then estimated the sequence similarity between lncRNA annotations from different species and used a Markov clustering algorithm to reconstruct homologous lncRNA families **(Figure 3.1A)**. I subsequently assessed the validity of this approach by showing that older lncRNAs are found in more species **(Figure 3.1B-C)** and that they tend to remain in conserved synteny to a similar extent with protein-coding genes **(Figure 3.1D)**. Based on these analyses, we identified, in each species, a total of 16,784-33,419 lncRNAs which I classified based on their position and orientation compared to adjacent protein-coding genes. Different species show similar distributions across genomic classes with 32-48% of lncRNAs overlapping a protein-coding gene in antisense orientation **(Figure 3.2A)**.

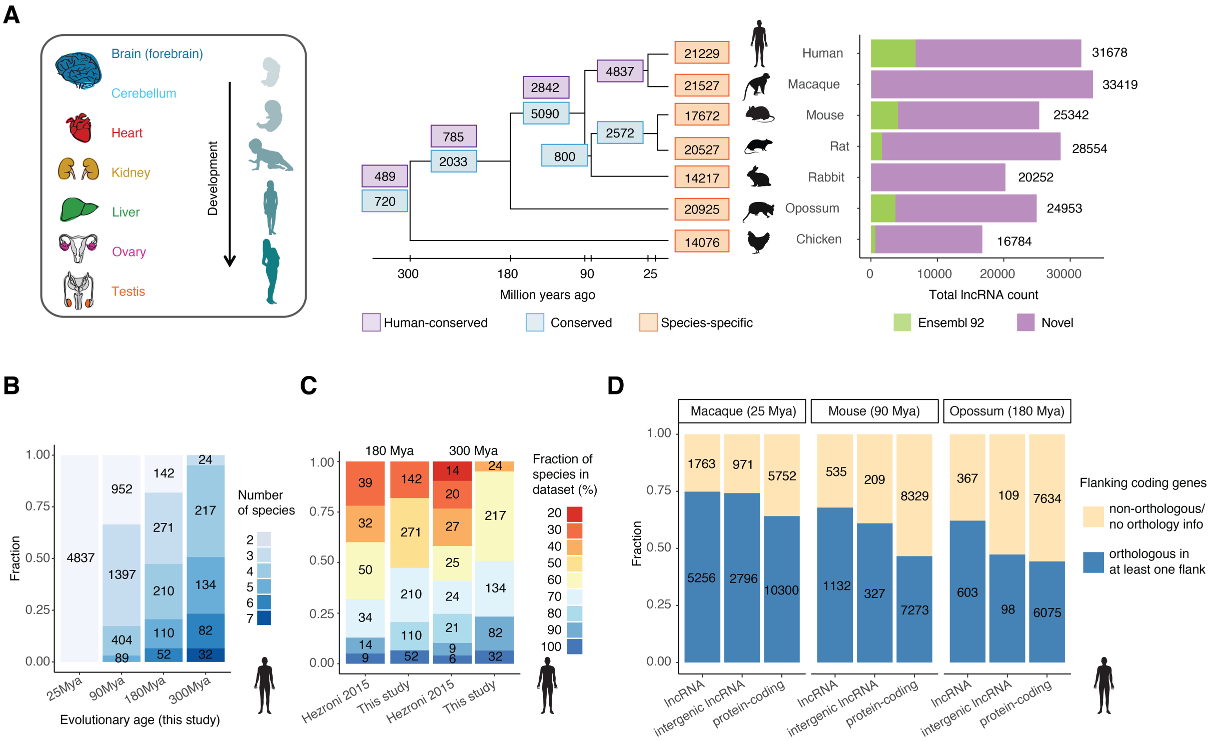


Figure 3.1: LncRNAs expressed during mammalian organ development. **(A)** Overview of the transcriptomic dataset (left) and number of lncRNAs identified in each species (right) and across phylogenetic groups (middle). Branches in the phylogenetic tree (middle) depict 1:1 orthologous lncRNA families, leaves correspond to species/lineage-specific lncRNAs. Barplots on the right show the overlap with Ensembl v92 annotations. **(B)** Number of species with a detected lncRNA member for human families of various evolutionary ages. **(C)** Fraction of species with a detected lncRNA member for human families conserved across mammals (180 Mya) and amniotes (300 Mya) in this and a previous study (*77*). **(D)** Fraction of lncRNAs and protein-coding gene orthologs found in conserved synteny with at least one protein-coding gene neighbor for increasing evolutionary distances. *Figure adapted from Sarropoulos et al. 2019* (*223*).

The spatial expression patterns of lncRNAs are similar across mammalian species, with most (35-60%) lncRNAs showing maximum expression in the testis **(Figure 3.2B)**, as previously observed (*47*, *48*, *77*). This disproportionate contribution of the testis has been attributed to the pervasive chromatin environment in the postmitotic spermatids, which also allows the transcription of various intergenic elements (*73*, *74*). Indeed, I showed that only the samples from sexually mature testis stand out in terms of lncRNA expression whereas earlier developmental stages resemble the other organs **(Figure 3.2C)**. Chicken shows an interesting difference compared to the mammalian species, expressing a large number of lncRNAs throughout ovary development, reaching similar levels to those observed for the mature testis, but eventually falling rapidly in adulthood **(Figure 3.2C)**.

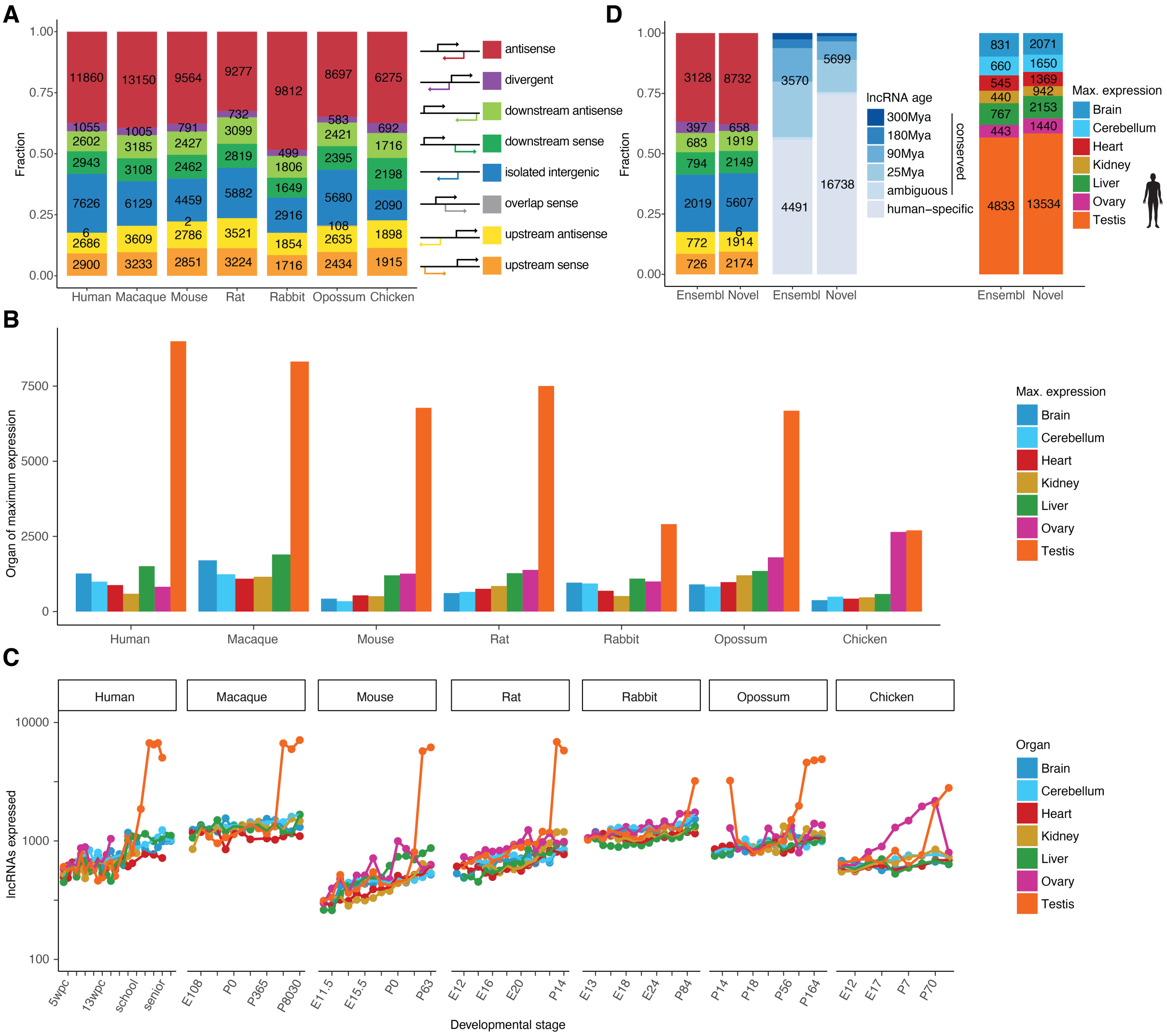


Figure 3.2: Genomic classification and spatiotemporal expression patterns of lncRNAs. (A) Distribution of lncRNAs among genomic classes in each species. **(B)** Organ of maximum expression for expressed lncRNAs (≥1 RPKM) in each species. **(C)** Number of lncRNAs expressed (≥1 RPKM) in each species during the development of each organ (in logarithmic scale). **(D)** Comparison of genomic classes (left), evolutionary age (middle) and organ of maximum expression (right) for known (Ensembl) and newly annotated (novel) human lncRNAs. *Figure adapted from Sarropoulos et al. 2019* (*223*).

The lncRNA repertoires identified during this work substantially expanded the annotations available in Ensembl at the time (*225*), even for well annotated species, such as human and mouse (**Figure 3.1A**). As expected, newly identified lncRNAs are enriched for species- and lineage-specific transcripts (**Figure 3.2D**). However, newly identified and previously annotated lncRNAs show similar distributions in their genomic classes and spatial expression patterns (**Figure 3.2D**), suggesting that these extensions of previous annotations are unbiased.

3.1.2 Developmentally dynamic lncRNAs are enriched for functional loci

A major goal in lncRNA research is to distinguish between transcripts that are likely to be functional and those that represent transcriptional noise (*6*, *29*, *33*). Several features have been suggested to prioritize functionally relevant lncRNAs, including robust expression levels, transcript stability, developmentally dynamic expression and reproducibility across biological replicates (*29*, *33*).

I thus sought to utilize the power of this transcriptomic dataset, which contains densely sampled time series and several biological replicates across multiple organs and species, to identify lncRNAs that meet these criteria, reasoning that this set would be enriched for lncRNAs associated with functions in organ development. Using a regression approach (*226*), which incorporates information about expression levels across replicates and developmental stages, I identified lncRNAs with significant differential expression through time (i.e., developmentally dynamic) for each organ and species **(Figure 3.3A)**. As the large number of lncRNAs expressed in the adult testis are thought to be associated with a pervasive chromatin environment, mature testis samples were excluded from this analysis. Collectively, I identified 3,929-7,491 dynamic lncRNAs in each species.

Following the identification of these developmentally dynamic lncRNAs, I used a series of metrics to assess whether these transcripts are indeed enriched for functional loci compared to the rest of lncRNAs. These analyses are presented in detail below.

3.1.2.1 Developmentally dynamic lncRNAs share features with protein-coding genes

The majority of protein-coding genes (73-90% depending on the species) but only a fraction of lncRNAs (16-38%) show developmentally dynamic expression in at least one organ. However, in contrast to the high variability in the numbers of non-dynamic lncRNAs, which scale with the size of each genome, the numbers of developmentally dynamic lncRNAs remain similar between species **(Figure 3.3B)**. Thus, unlike the total size of lncRNA repertoires which seems to evolve neutrally, the number of developmentally dynamic lncRNAs remains relatively stable during evolution. This resembles what is observed for protein-coding genes (14,041-17,090 dynamic in at least one organ depending on the species).

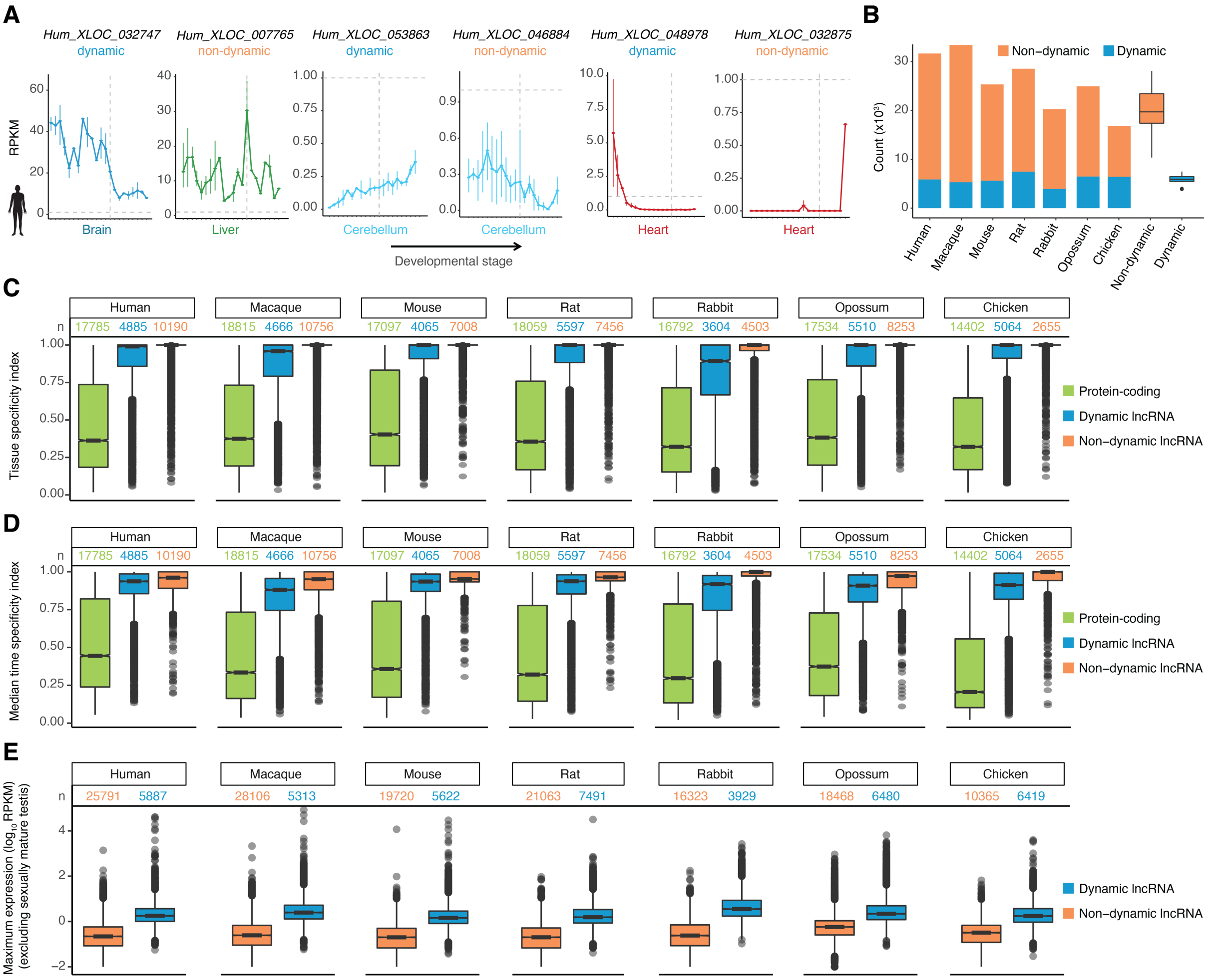


Figure 3.3: Features of developmentally dynamic lncRNA expression. **(A)** Examples of human developmentally dynamic (n = 5,887) and non-dynamic (n = 25,791) lncRNA expression profiles (mean expression; vertical bars represent the minimum and maximum values across replicates) for varying levels of maximum expression, replicate reproducibility and expression windows. The vertical dashed line represents birth; the horizontal dashed line marks 1 RPKM. **(B)** Number of non-dynamic and dynamic lncRNAs identified in each species. The box plots summarize the variability in the size of the repertoires across species (n = 7). **(C, D)** Tissue-specificity **(C)** and median time-specificity **(D)** of non-dynamic and dynamic lncRNAs, and protein-coding genes, across species. Tissue- and time-specificity indexes range from 0 (broad expression) to 1 (specific expression). All comparisons between non-dynamic and dynamic lncRNAs, and protein-coding genes are significant ($P < 10^{-15}$, two-sided Mann–Whitney *U*-test). **(E)** Maximum expression levels ($\log_{10}$(RPKM)) for developmentally dynamic and non-dynamic lncRNAs across species (excluding samples from the sexually mature testis). Developmentally dynamic lncRNAs are more highly expressed in all species ($P < 10^{-15}$, two-sided Mann–Whitney *U*-test). *Figure adapted from Sarropoulos et al. 2019* (*223*).

Developmentally dynamic lncRNAs also show similarities to protein-coding genes in terms of their expression patterns. Although most dynamic lncRNAs show significant differential expression in a single organ, in agreement with the overall high tissue specificity previously reported for lncRNAs (*47*), they show significantly broader spatiotemporal expression compared to non-dynamic lncRNAs (**Figure 3.3C-D**). Additionally, developmentally dynamic lncRNAs are expressed at higher levels (**Figure 3.3E**), and are on average longer and contain more exons compared to non-dynamic lncRNAs, suggesting selection for splice-sites and against premature polyadenylation sites (**Figure 3.4A-B**).

I next sought to examine whether the broader and more complex expression profiles of developmentally dynamic lncRNAs are also reflected in their transcriptional regulation. Compared to protein-coding genes, lncRNAs have been reported to be transcribed from simpler promoters that contain fewer TF binding sites (*71*). Estimating the number of distinct TFs bound on each promoter based on a collection of uniformly processed ChIP-seq experiments for more than 300 mouse TFs (*227*), I was able to recapitulate this observation (**Figure 3.4C**). However, significantly more TFs are bound to the promoters of dynamic than to those of non-dynamic lncRNAs, suggesting that the former show stronger and more complex transcriptional regulation (**Figure 3.4C**). In support of the relevance of this increased complexity for the development of the organs examined in this study, lncRNAs dynamic in distinct organs are more prominently bound by TFs that are specifically expressed or have well-established roles in the development of the respective organ (**Figure 3.4D**). For example, the TFs NKX2-5, MEF2D and GATA4, which all play important roles during heart development (*228*), predominantly bind to promoters of lncRNAs dynamic in the heart.

In sum, developmentally dynamic lncRNAs represent a subset of the total lncRNA repertoire of each species. This subset is enriched for features typically associated with protein-coding genes, such as stable numbers, higher and broader spatiotemporal expression, longer transcripts and more complex regulation.

3.1.2.2 Developmentally dynamic lncRNAs are enriched for older genes.

One of the strongest lines of evidence for the functional relevance of any genomic locus is its evolutionary conservation (*29*, *33*). I thus sought to test whether developmentally dynamic lncRNAs are enriched for older lncRNA groups. Indeed, the fraction of developmentally dynamic lncRNAs increases significantly across older groups. Whereas only 13% of human-specific lncRNAs are developmentally dynamic, this fraction rises to 40% when considering lncRNAs that are at least 300 million years old (**Figure 3.5A**). However, nearly half of all developmentally dynamic lncRNAs in humans are younger than 25 million years and not shared with any other species in this dataset. Thus, while the enrichment for evolutionary conserved genes supports the functional relevance of developmentally dynamic lncRNAs, the two metrics are not redundant to each other.

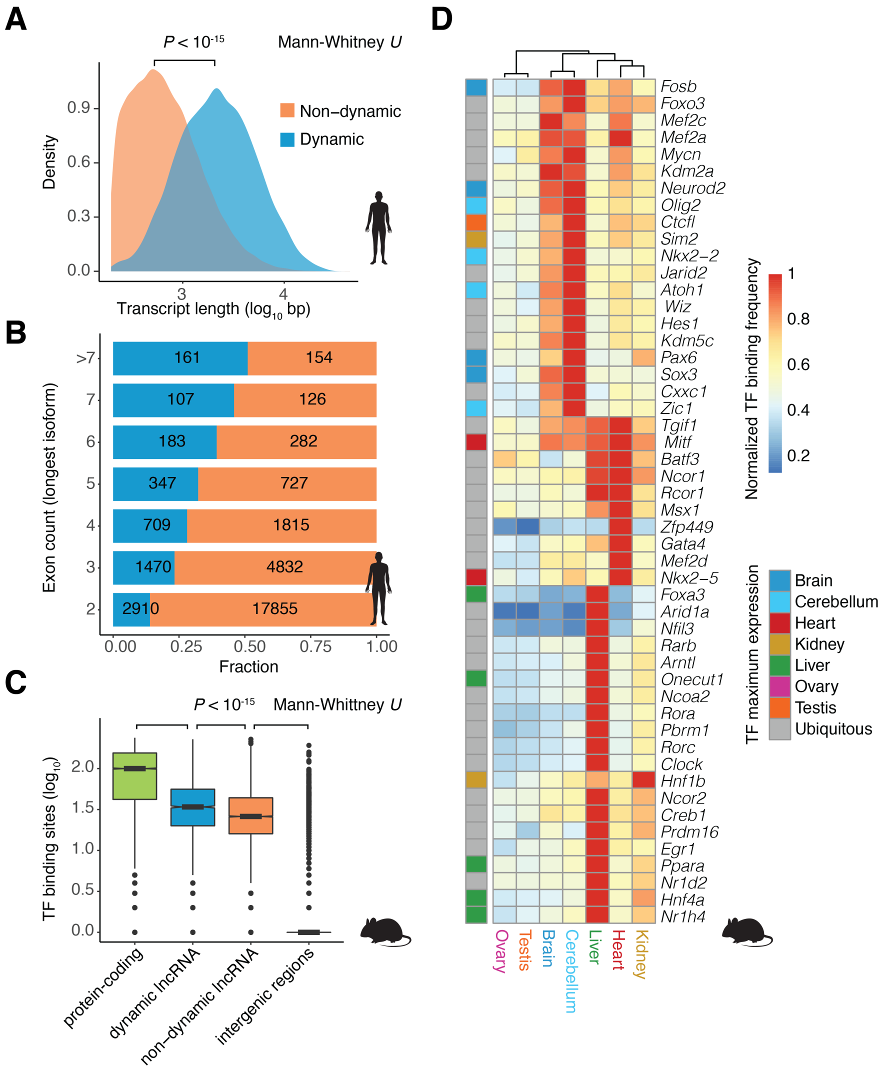


Figure 3.4: Transcript length and regulatory complexity of dynamic lncRNAs. (A) Density distribution of transcript length for non-dynamic (n = 25,791) and dynamic (n = 5,887) human lncRNAs. **(B)** Fraction of developmentally dynamic human lncRNAs among isoforms with an increasing number of exons. The number of exons is significantly higher for developmentally dynamic lncRNAs ($P < 10^{-15}$, two-sided Mann–Whitney U-test). **(C)** Number of TF-binding sites overlapping the promoters of protein-coding genes (n = 20,202), dynamic lncRNAs (n = 3,169) and non-dynamic lncRNAs (n = 11,818), and size-matched random intergenic regions (n = 20,202). **(D)** Normalized TF-binding frequency (heat map) of the 50 TFs with the highest binding variability across organs. Rows and columns are hierarchically clustered. The row annotation depicts the organ of maximum expression for organ-specific TFs. *Figure adapted from Sarropoulos et al. 2019* (223).

An important consideration for this analysis is the potential overlap of lncRNAs with other conserved genomic elements, such as protein-coding exons or *cis*-regulatory elements, which could lead to an overestimation of lncRNA evolutionary ages. This is especially important because developmentally dynamic lncRNAs are on average closer to protein-coding genes and are more likely to overlap with them **(Figure 3.5B)**. To control for this potentially confounding effect, I repeated the same analysis after excluding lncRNAs transcribed in antisense or divergent orientation to protein-coding genes, as well as intergenic lncRNAs overlapping previously reported transcribed enhancers (*229*). Developmentally dynamic lncRNAs remain significantly enriched for older lncRNA groups (**Figure 3.5C**).

Another important control for this analysis concerns lncRNA expression levels. As discussed above (3.1.2.1), developmentally dynamic lncRNAs are more highly expressed than non-dynamic ones (**Figure 3.3E**). Since high expression levels have been associated with increased evolutionary conservation (*49*), it was important to assess the added value of further requiring lncRNAs to be differentially expressed during organ development. To address this question, I generated an expression-matched set of dynamic and non-dynamic lncRNAs by selecting lncRNAs with similarly high maximum expression levels. The enrichment of older lncRNA groups for developmentally dynamic transcripts remained highly significant in this expression-matched set, demonstrating the importance of considering expression patterns across developmental series compared to single time points (**Figure 3.5D**).

Additionally, since the majority of dynamic lncRNAs are highly expressed, it is important to examine whether the few lowly expressed dynamic lncRNAs also display increased evolutionary conservation or whether this enrichment is driven exclusively by the highly expressed transcripts. In order to test these competing hypotheses, I generated a second set of expression-matched dynamic and non-dynamic lncRNAs. This time I focused exclusively on genes showing a maximum expression level between 0.25 and 0.75 RPKM, values often considered below the limit of detection for bulk sequencing studies (*172*). Only few dynamic lncRNAs fall in this expression range but they remain significantly enriched for evolutionary conservation compared to the expression-matched non-dynamic lncRNAs (**Figure 3.5E**). A potential explanation for the low expression levels observed for these dynamic lncRNAs could involve their specific expression in a cell type that is rare in the organs sampled in this study (*61*).

Overall, with these analyses I showed that developmentally dynamic lncRNAs are more evolutionary conserved than non-dynamic transcripts and that these differences cannot be explained by overlapping protein-coding genes or by differences in the maximum expression levels between the two classes of lncRNAs.

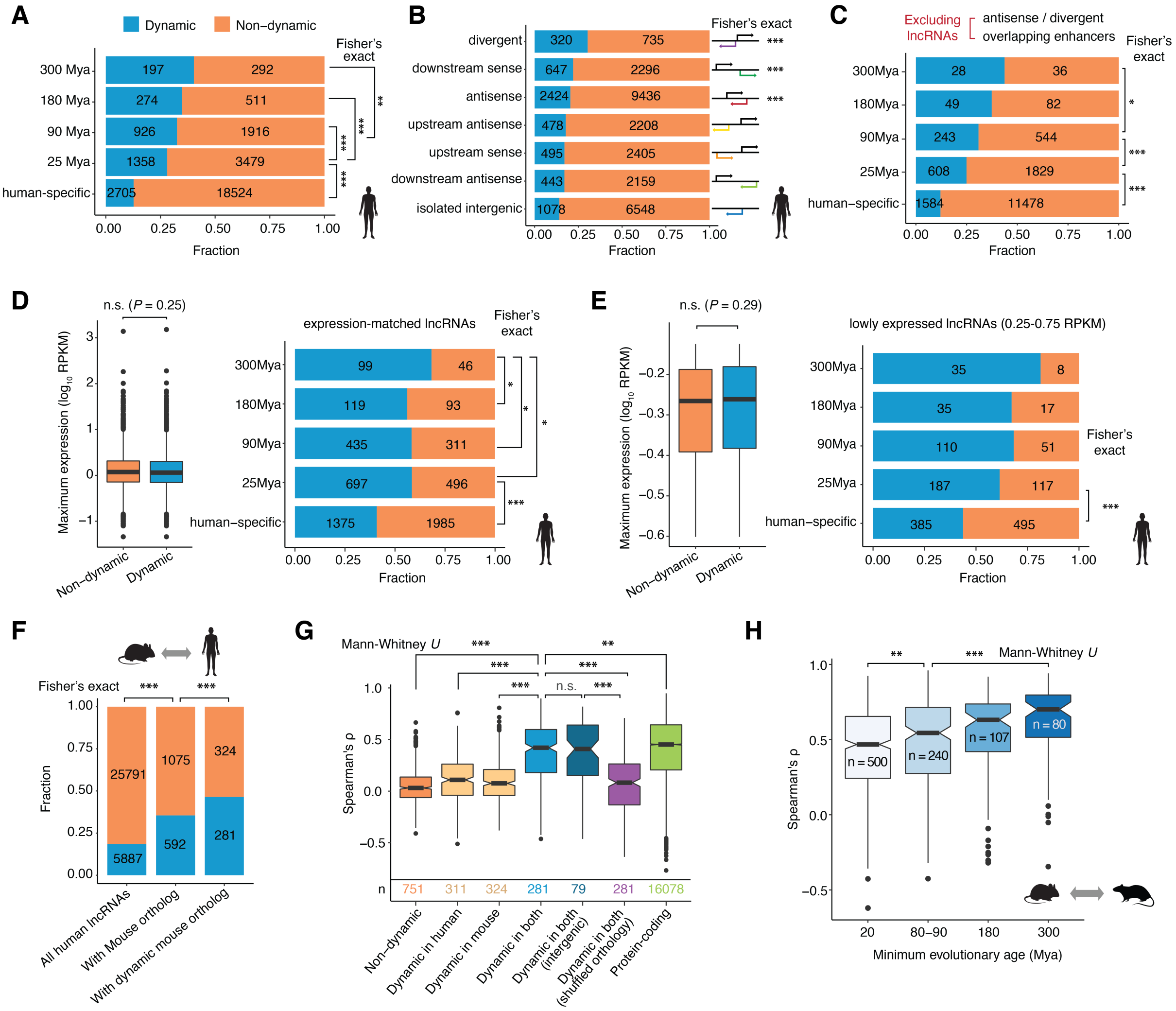


Figure 3.5: Sequence and expression conservation of dynamic lncRNAs. **(A)** Fraction of dynamic loci for human lncRNAs of different evolutionary ages. **(B)** Fraction of developmentally dynamic human lncRNAs (n = 5,887) for different genomic classes. Overrepresented classes were determined by comparing the fraction of dynamic lncRNAs in each class against all other classes. **(C)** Fraction of human lncRNAs that are intergenic, developmentally dynamic and that do not overlap enhancers (n = 16,481) among different age groups. **(D)** Generation of expression-matched dynamic (n = 2,906) and non-dynamic (n = 3,098) lncRNAs (left) and their distribution across different evolutionary age groups (right). **(E)** Generation of expression-matched dynamic (lowly expressed (0.25–0.75 RPKM) dynamic (n = 798) and non-dynamic (n = 717) lncRNAs (left) and their distribution across different evolutionary age groups (right). **(F)** Fraction of developmentally dynamic human lncRNAs (n = 5,887) with or without a mouse (dynamic or not) ortholog ($P < 10^{-15}$, Fisher's exact test). **(G)** Similarity of spatiotemporal expression (Spearman correlation coefficient between human and mouse organs/developmental stages) for 1:1 orthologs. **(H)** Expression similarity across matched organs and developmental stages for mouse and rat 1:1 orthologous lncRNAs that are dynamic in both species, for different evolutionary ages. *Figure adapted from Sarropoulos et al. 2019* (223).

3.1.2.3 Developmentally dynamic lncRNA expression is conserved across mammals

I next sought to investigate whether the higher conservation of developmentally dynamic lncRNAs at the sequence level is also reflected in the conservation of their expression. Comparing human and mouse, I observed that the likelihood that a conserved lncRNA is developmentally dynamic in one species is significantly higher when its ortholog is also dynamic (**Figure 3.5F**; $P < 10^{-15}$, hypergeometric test), suggesting that developmentally dynamic expression of at least a subset of lncRNAs has been preserved during evolution. To quantify the degree of spatiotemporal expression conservation between two orthologous lncRNAs, I estimated the Spearman's correlation coefficient between their expression profiles across corresponding organs and developmental stages. Developmental correspondences between human and mouse were established by Dr. Cardoso-Moreira in a separate study (*172*).

Strikingly, considering the overall rapid evolution of lncRNA expression, lncRNAs with developmentally dynamic expression in both species show almost as high expression similarity (median rho = 0.40) as protein-coding genes (median rho = 0.46; **Figure 3.5G**). This estimate is significantly higher than the similarity observed between non-dynamic lncRNAs, those dynamic in only one species, or pairs of dynamic lncRNAs with shuffled orthology relationships **(Figure 3.5G)**. To make sure that the observed effect is not due to the increased proximity of developmentally dynamic lncRNAs to protein-coding genes, I repeated the same analysis excluding antisense and divergent lncRNAs, observing similarly high levels of expression conservation **(Figure 3.5G)**.

Having identified a subset of lncRNAs with conserved expression, I next decided to test whether conservation levels are associated with the age of the lncRNAs, that is, whether older lncRNAs are subjected to stronger stabilizing constraints. To test this hypothesis across a wider phyletic range, I focused on lncRNAs dynamic in mouse and rat, which are separated by approximately 20 million years of evolution, the smallest distance in my dataset. I then stratified lncRNAs based on their inferred evolutionary age (20-300 million years), observing a significant increase in expression similarity with lncRNA age **(Figure 3.5H)**. Thus, expression profiles of lncRNAs that appeared earlier during tetrapod evolution, and have since been preserved, are subjected to stronger evolutionary constraints.

3.1.2.4 Molecular evidence supports the functional relevance of dynamic lncRNAs

The ultimate proof for the functionality of a genomic element comes from observing a phenotypic effect upon its perturbation. I thus sought to assess whether developmentally dynamic lncRNAs would be more likely to be associated with a phenotype based on prior studies. First, I examined a database of molecularly characterized, functional lncRNAs, lncRNAdb (*230*). The fraction of developmentally dynamic transcripts amongst lncRNAs annotated as functional in human is 76%, four times higher than that of all human lncRNAs, and almost as high as the fraction of protein-coding genes showing developmentally dynamic expression **(Figure 3.6A)**.

However, this observation may be, at least partially, due to ascertainment biases; for example, the preferred experimental characterization of broadly expressed and highly conserved lncRNAs — features associated with developmentally dynamic expression. I therefore also examined a set of lncRNAs that were recently associated with cell proliferation phenotypes based on an unbiased CRISPRi screen in human (*93*). It should be noted that this screen was performed in a very different physiological context (testing the effect of lncRNAs on cellular growth) and that most phenotype-associated lncRNAs showed an effect at most in a single cell type. However, despite this difference in context, I observed that lncRNAs shared between the screen library and my annotation were more likely to result in a phenotype upon perturbation **(Figure 3.6B)**. Moreover, developmentally dynamic lncRNAs showed a significant further enrichment amongst screen hits **(Figure 3.6B)**. Thus, the increased functional relevance of developmentally dynamic lncRNAs is further supported by molecular evidence.

In conclusion, a series of genomic, biochemical and evolutionary features, as well as direct evidence from perturbation assays, demonstrate that developmentally dynamic lncRNAs are enriched for functional loci and serve as candidates for further molecular investigation.

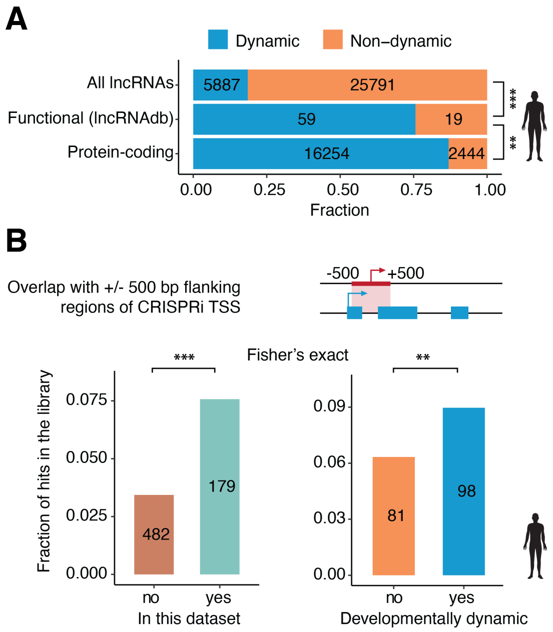


Figure 3.6: Molecular evidence for the functionality of dynamic lncRNAs. (A) Fraction of dynamic loci for human lncRNAs, functionally characterized lncRNAs and protein-coding genes. **(B)** Fraction of lncRNAs present in the CRISPRi screen library resulting in a significant growth phenotype (hits) in at least one cell line for lncRNAs present (n = 2,364) or absent (n = 14,037) in this lncRNA annotation and dynamic (n = 1,093) or non-dynamic (n = 1,277). *Figure adapted from Sarropoulos et al. 2019* (*223*).

3.1.3 Patterns of lncRNA expression during organ development

Having identified a high confidence set of lncRNAs that are less likely to represent transcriptional noise, I then sought to examine their expression patterns during the development of mammalian organs.

3.1.3.1 Periods of greater change of lncRNA expression

Mammalian organ development is not a smooth process, but instead marked by distinct periods during which large numbers of protein-coding genes change their expression levels (*172*). These periods are mostly associated with the establishment of organ identity during early development (e.g., between E11 and E12 in mouse) and the transition to mature organ-specific functions around birth (e.g., P3 and P14 in mouse). Aiming to understand how lncRNA expression relates to the patterns observed for protein-coding genes, I used both gene sets to identify differential expression across adjacent developmental stages.

The stages where developmentally dynamic lncRNAs show the greatest differential expression coincide with these previously identified peaks of developmental change **(Figure 3.7A, Supplementary Figure 1)**. To make sure that this concordance is not an artifact of chromatin changes associated with the expression of protein-coding genes, I repeated this analysis focusing on mouse intergenic lncRNAs, located more than 100 kb away from the closest protein-coding gene, observing similar patterns (**Figure 3.7B**). These observations suggest that sizeable sets of lncRNAs are differentially expressed during critical periods of organ development, in a likely autonomous manner, and potentially contributing to the developmental processes that take place during these periods.

3.1.3.2 Co-expression of lncRNAs with protein-coding genes

To further refine our understanding of the potential contributions of lncRNAs to mammalian organ development, I examined their co-expression with protein-coding genes, akin to previous "guilt by association" approaches (*48*, *231*). I focused this analysis on human and mouse, as functional annotations of protein-coding genes for the remaining species are sparser and to a large degree derived based on gene homology to mouse. I clustered protein-coding genes and lncRNAs with dynamic profiles in each organ and estimated the fraction of lncRNAs in each cluster. Although the precise gene ontology enrichment terms differ between organs, clusters with the highest lncRNA fraction consistently show similar developmental trajectories (**Figure 3.7C, Supplementary Figure 2**). For example, increasing expression during prenatal development with a postnatal expression drop is associated with developmental functions in most organs, as well as with immune responses in liver and kidney. Another cluster consistently characterized by a high fraction of lncRNAs is that associated with a steep increase in expression after birth, containing genes involved in organ-specific functions, such as ion transport for the nervous tissues and lipid metabolism for the liver. By contrast, lncRNAs contribute the least to clusters associated with housekeeping genes, supporting the notion that only few lncRNAs are involved in essential cellular functions (*29*). Overall, these transcriptome-wide co-expression analyses suggest that many dynamic lncRNAs could contribute to the development and adult physiology of mammalian organs.

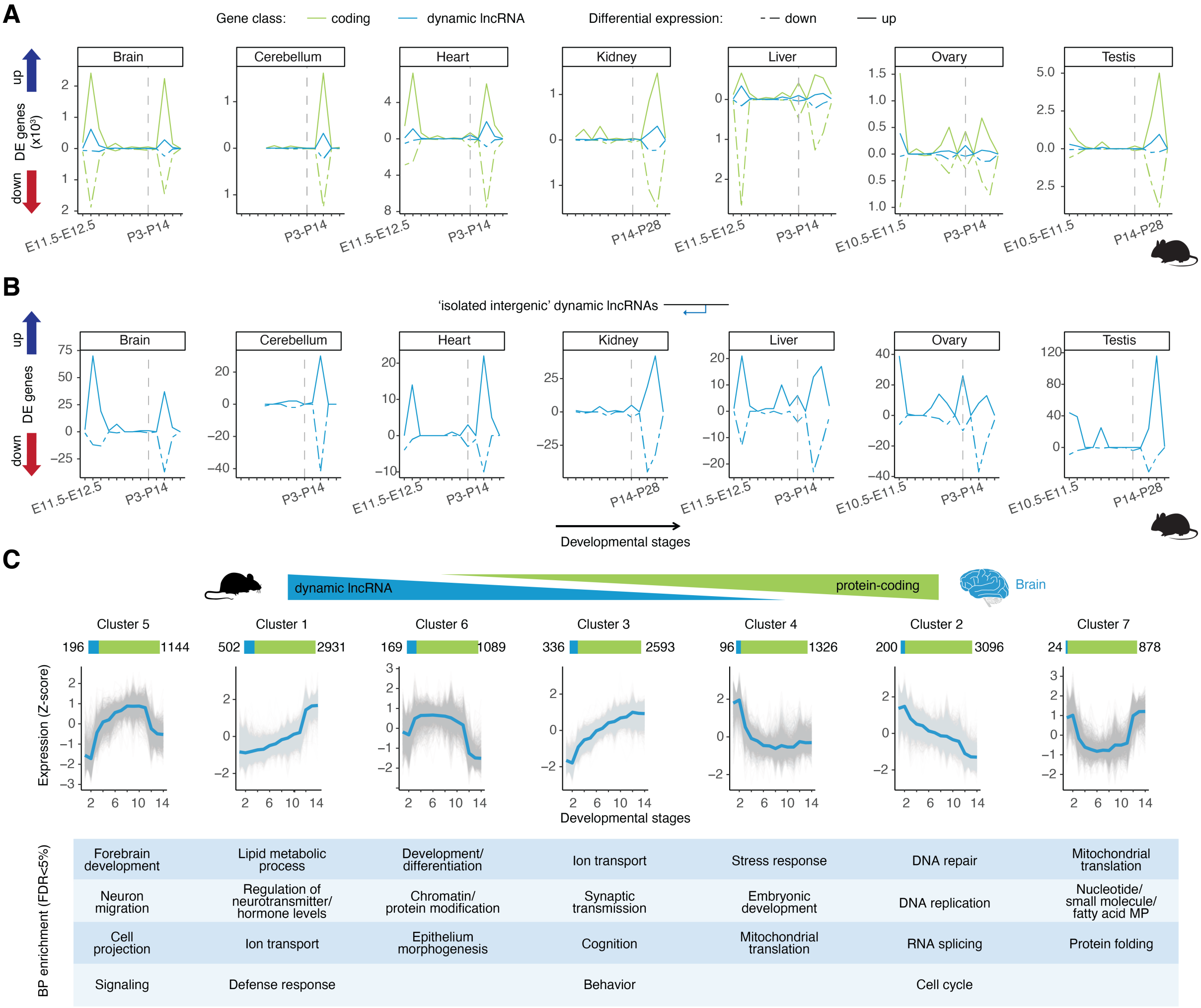


Figure 3.7: Expression patterns of dynamic lncRNAs in mammalian organ development. (A) Number of differentially expressed (DE) protein-coding genes and dynamic lncRNAs between adjacent developmental stages (additional species in **Supplementary Figure 1**). **(B)** Number of differentially expressed "isolated intergenic" (more than 100 kb from the closest protein-coding gene) dynamic lncRNAs between adjacent stages during mouse development. **(C)** Clusters of developmentally dynamic lncRNAs and protein-coding genes in the brain (n = 14,629 genes; additional organs in **Supplementary Figure 2**). Grey lines represent individual gene trajectories and solid lines posterior mean trajectories for each cluster. Clusters are arranged by decreasing fraction of lncRNAs. Enriched representative biological processes (BP; Benjamini–Hochberg adjusted $P < 0.05$, hypergeometric test) are shown for each cluster. *Figure adapted from Sarropoulos et al. 2019* (*223*).

3.1.4 Differences across organs and developmental stages

3.1.4.1 A dichotomy between developmental lncRNA expression and conservation

During early organogenesis, the protein-coding gene transcriptomes of different organs are highly similar to each other and also show the highest evolutionary conservation (*172*). As development progresses, the overall transcriptomes gradually diverge into distinct programs for each organ, with increasing contributions from evolutionarily younger protein-coding genes that lead to an overall decrease in conservation (*172*), in agreement with von Baer's predictions (see section 3.4). Motivated by these differences in the expression dynamics of protein-coding genes, I sought to investigate how lncRNA expression and conservation change during development.

In line with the high tissue-specificity of lncRNAs **(Figure 3.3C)**, I observed that the number of developmentally dynamic lncRNAs that are expressed in a given organ and stage increases during development — i.e., as organs become more distinct from each other **(Figure 3.8A)**. In mouse, the number of dynamic lncRNAs expressed in adult organs is approximately two times higher than in the respective organ during early development (E10-E11). However, contrary to the increase in the total number of expressed dynamic lncRNAs, evolutionary conservation declines. Specifically, the fraction of dynamic lncRNAs showing evidence for being selectively preserved during mammalian evolution (i.e., those with an age > 80 million years) overall decreases with time **(Figure 3.8B)**. In addition, the expression similarity between 1:1 lncRNA orthologs, which are dynamic in both human and mouse, also decreases during development **(Figure 3.8C)**. Thus, while the overall contribution of lncRNAs to the transcriptome of different organs increases during development, evolutionary conserved lncRNAs are mainly expressed in early stages of organ development.

3.1.4.3 LncRNA conservation is associated with pleiotropy and functional constraints

In protein-coding genes, evolutionary conservation is strongly associated with pleiotropy, i.e., the shared expression across cell types, organs and developmental stages (*172*). This increased evolutionary conservation is considered a consequence of pleiotropy imposing stronger functional constraints, meaning that changes in the sequence or expression of a gene are more likely to be deleterious for the organism if that gene is involved in multiple processes. Since both protein-coding genes and lncRNAs show higher evolutionary conservation during early organ development, I hypothesized that early expressed lncRNAs might also be more pleiotropic and associated with more severe phenotypic consequences.

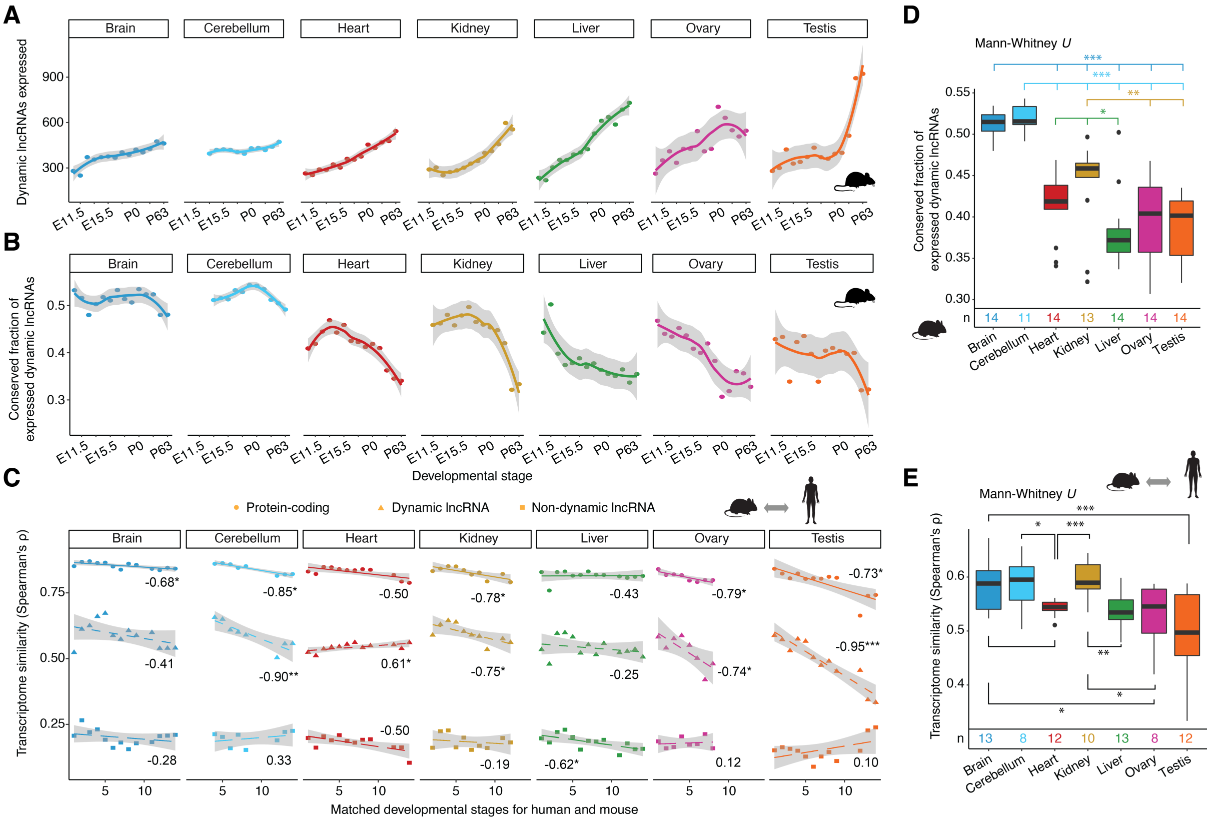


Figure 3.8: Evolutionary conservation of dynamic lncRNAs across organs and developmental stages. **(A)** Number of dynamic lncRNAs (n = 5,622) expressed and **(B)** fraction of those conserved (evolutionary age of at least 80 million years) during mouse organ development. Lines estimated through LOESS regression; 95% confidence interval shown in grey. **(C)** Expression similarity between human and mouse 1:1 orthologous protein-coding genes (n = 16,078), developmentally dynamic (n = 281) and non-dynamic (n = 1,386) lncRNAs across organs/developmental stages. Each point corresponds to the Spearman correlation coefficient of expression between human and mouse orthologs for matching samples. Lines and the 95% confidence interval (shaded regions) correspond to linear model predictions. Spearman correlation coefficients between expression similarity and developmental stage are given for each comparison. **(D)** Fraction of conserved (≥80 Mya) expressed dynamic lncRNAs from **(A)** summarized by organ. The line colour signifies the focal organ for each statistical comparison. **(E)** Expression similarity between dynamic human and mouse orthologous lncRNAs from **(B)** summarized by organ. *Figure adapted from Sarropoulos et al. 2019* (*223*).

To test this hypothesis, I revisited the clustering of lncRNAs based on their temporal expression profiles within each organ. I classified lncRNAs into early and late expressed, assigning clusters associated with more complex temporal patterns into a third group, termed “other”. Across all organs, early expressed lncRNAs show significantly lower tissue specificity indexes, that is they are more pleiotropic **(Figure 3.9A)**. To compare the functional impact of early versus late expressed lncRNAs, I reexamined the lncRNAdb catalog of functional lncRNAs and the CRISPRi screen data. Early expressed lncRNAs are significantly more likely to be listed as functional in lncRNAdb and to lead to a growth phenotype upon perturbation than those expressed late during development **(Figure 3.9B-C)**.

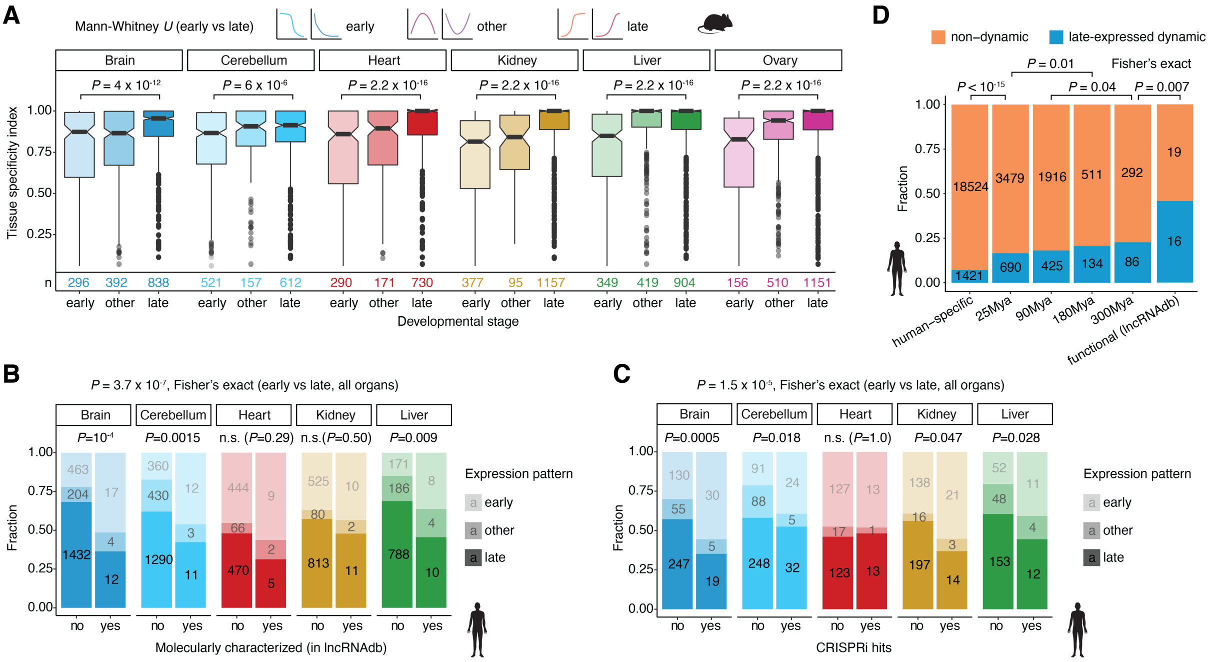


Figure 3.9: Pleiotropic and functional constraints of early- and late-expressed dynamic lncRNAs. **(A)** Tissue-specificity for mouse lncRNAs with different developmental trajectories. **(B, C)** Fraction of human lncRNAs with different developmental trajectories among functionally characterized lncRNAs (n = 59) **(B)** and CRISPRi growth screen hits (n = 98) **(C)**. **(D)** Fraction of late-expressed dynamic (n = 2,956) and non-dynamic (n = 25,791) lncRNAs for different age groups and functionally characterized human lncRNAs. *Figure adapted from Sarropoulos et al. 2019* (*223*).

Collectively, lncRNAs expressed in early stages of development are more pleiotropic and subject to stronger functional and evolutionary constraints. By contrast, late expressed lncRNAs, which remain enriched for older and functionally characterized transcripts compared to non-dynamic lncRNAs (**Figure 3.9D**), are characterized by higher tissue- and lineage-specificity, suggesting that each of them individually has a milder effect on developmental programs and phenotypes. However, the increase in the total number of lncRNAs expressed during development suggests overall sizeable contributions to organ-specific gene expression programs.

3.1.5 Co-expression of lncRNAs with adjacent protein-coding genes

Having established that developmentally dynamic lncRNAs are more likely to play roles in mammalian organ development, I next sought to investigate their potential mechanisms of action. Gene expression data alone are not sufficient for such inferences, which typically require extensive experimental scrutiny (*20*). However, transcriptomics data can be used to test whether lncRNA expression patterns are compatible with commonly proposed modes of action.

Perhaps the most frequently discussed mechanism of putative lncRNA function involves the regulation of the expression of their genomic neighbors, as has been observed for several well-characterized lncRNAs, such as *XIST* in eutherian mammals and *Airn* in mouse (*20*). LncRNAs acting in such a manner are expected to show correlated expression with their target genes, at least in a subset of biological contexts. I thus sought to assess the plausibility and potential extent of lncRNA-mediated *cis*-regulation in the context of organ development, focusing on developmentally dynamic lncRNAs.

For each human and mouse protein-coding gene, I calculated gene expression correlations (Pearson's *r*) to its nearest lncRNA and mRNA across all samples (excluding the sexually mature testis). I observed a significantly higher degree of expression correlation between dynamic lncRNAs and their adjacent protein-coding genes compared to mRNA-mRNA controls (**Figure 3.10A**; $P < 10^{-15}$, Wilcoxon's signed-rank test). Although the distance between two genes has an effect on the degree of their correlation, I noted an excess of positive correlations for lncRNA-mRNA pairs for distances up to 100 kb **(Figure 3.10B)**. In the absence of proper controls for bidirectional and antisense lncRNAs, as very few protein-coding genes are transcribed in such orientations, I repeated the analysis excluding such pairs. The difference between lncRNA-mRNA and mRNA-mRNA pairs remains significant (**Figure 3.10C**; $P < 10^{-15}$, Wilcoxon's signed-rank test).

To identify candidate co-expressed lncRNA-mRNA pairs, I required a significantly higher correlation coefficient compared to the respective mRNA-mRNA control pair. Additionally, I required all significant lncRNA-mRNA pairs to show a correlation greater than a cutoff (r>0.75) derived from comparing paralogous versus non-paralogous protein-coding genes **(Figure 3.11A-B)**. This additional comparison allowed me to infer a level of correlation that implies functional relatedness above the background expectation of sharing the same chromatin environment.

Protein-coding genes that are significantly more correlated with their neighboring lncRNA compared to the mRNA control are enriched for developmental genes involved in processes such as tissue morphogenesis and cell differentiation **(Figure 3.11C-D)**. This observation suggests that the previously reported enrichment of lncRNAs near developmental regulators (*231*) is likely of biological significance, rather than a mere artifact of the larger intergenic regions surrounding these genes (*232*). In agreement with this notion, the lncRNAs identified from this analysis are enriched for a set of "positionally conserved" lncRNAs that are linked to chromatin organization structures and are co-expressed with their adjacent developmental protein-coding gene (*233*) **(Figure 3.11E)**. Additionally, several of the identified lncRNAs, such as *GAS6-AS2* (*94*), *DEANR1* (*234*), *SSTR5-AS1* (*235*), *EMX2OS* (*236*) and *DLX1AS* (*233, 234*), have also been experimentally demonstrated to regulate their protein-coding gene neighbors.

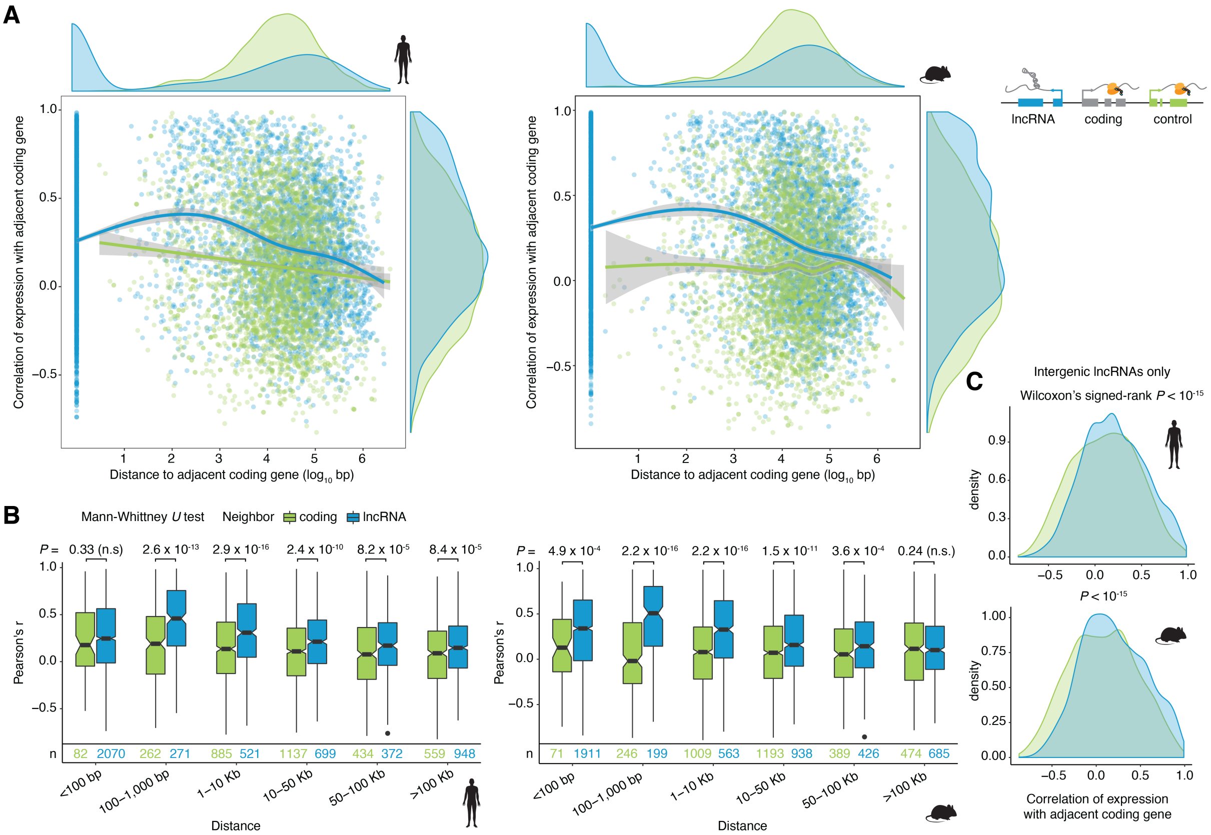


Figure 3.10: Co-expression of dynamic lncRNAs with adjacent protein-coding genes. (A) Relationship between distance and Pearson correlation of expression for lncRNA–mRNA (human = 4,881; mouse = 4,722) and mRNA–mRNA (human = 3,359; mouse = 3,382) pairs. Curves were estimated through LOESS regression and the 95% confidence interval is shown in grey. **(B)** Distribution of Pearson's *r* for lncRNA–mRNA and mRNA–mRNA pairs across different distance intervals. **(C)** Density distributions of Pearson's *r* between a protein-coding gene and its nearest dynamic lncRNA (human = 2,440; mouse = 2,549) and protein-coding gene (human = 1,606; mouse = 1,777) after excluding antisense and divergently transcribed lncRNAs. *Figure adapted from Sarropoulos et al. 2019* (*223*).

Of the 354 and 458 protein-coding genes co-expressed with a lncRNA in human and mouse respectively, 77 are shared between the two species **(Figure 3.11F)**, a significant enrichment compared to what would be expected by chance alone ($P < 10^{-15}$, hypergeometric test). In comparison to protein-coding genes co-expressed with a lncRNA in either species, those detected in both show an even stronger association with organ development (38% involved in the development of at least one organ; $P < 0.0002$, hypergeometric test). Thus, developmental regulators are often associated with adjacent co-expressed lncRNAs, a feature maintained throughout mammalian evolution.

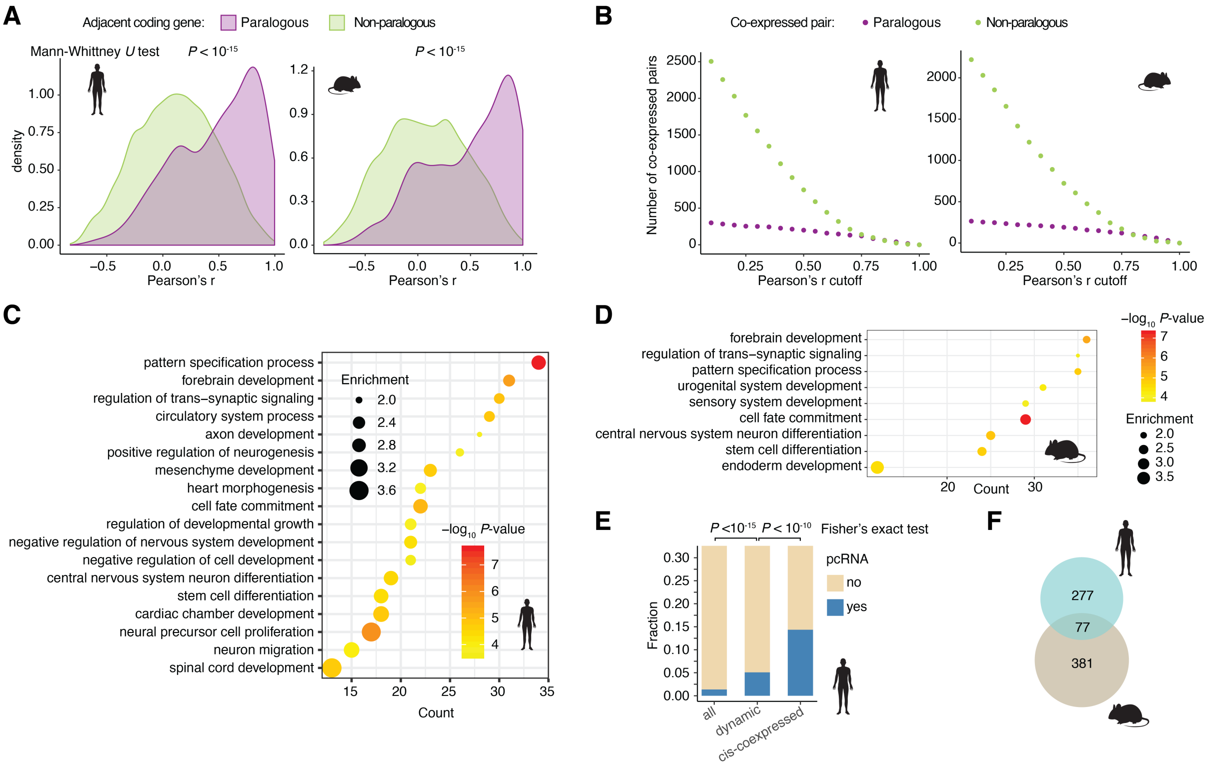


Figure 3.11: Identification of significantly co-expressed adjacent lncRNA-mRNA pairs. (A) Normalized density distribution of Pearson correlation coefficients (*r*) of spatiotemporal gene expression between adjacent paralogous (human = 267; mouse = 263) and non-paralogous (human = 3,359; mouse = 3,382) mRNA–mRNA pairs. **(B)** Number of paralogous and non-paralogous adjacent mRNA–mRNA pairs detected as co-expressed above a range of Pearson's *r* cutoffs. **(C, D)** Enriched biological processes among human **(C)** and mouse **(D)** protein-coding genes with significantly higher expression correlations with their adjacent dynamic lncRNA than with the control protein-coding gene (n = 358; Benjamini–Hochberg adjusted $P < 0.01$, hypergeometric test). **(E)** Fraction of positionally conserved lncRNAs (pcRNAs) among all lncRNAs (n = 31,678), developmentally dynamic lncRNAs (n = 5,887) and lncRNAs co-expressed with their adjacent protein-coding genes (n = 411). **(F)** Overlap between human and mouse protein-coding genes that have a significantly higher expression correlation (Pearson's *r*) with their adjacent dynamic lncRNA than with the control protein-coding gene. *Figure adapted from Sarropoulos et al. 2019* (*223*).

For a small number of protein-coding genes, I detected co-expression with multiple dynamic lncRNAs. An extreme case is that of the homeobox TF *POU3F3*, which is co-expressed with four lncRNAs in human and three in mouse (**Figure 3.12**). These include the lncRNAs *linc-Brn1a/Pantr1* and *linc-Brn1b/Pantr2*, which have been previously shown to regulate the levels of *Pou3f3* in various biological contexts (*233*, *237*, *238*). Two more dynamic lncRNAs show correlated expression with *POU3F3* in human: *LINC01114/PANCAT* and the newly identified *Hum_XLOC_027055*, which is transcribed from an experimentally validated enhancer element active in the hindbrain (*239*). Although all four human lncRNAs are overall correlated with *POU3F3*, each shows a distinct spatiotemporal expression profile, suggesting a modular mode of regulation. This case serves an example of the potential complexity of lncRNA-mediated gene regulatory networks, where different lncRNAs specifically contribute to the expression of a protein-coding gene in distinct tissues, developmental stages and — presumably — cell types.

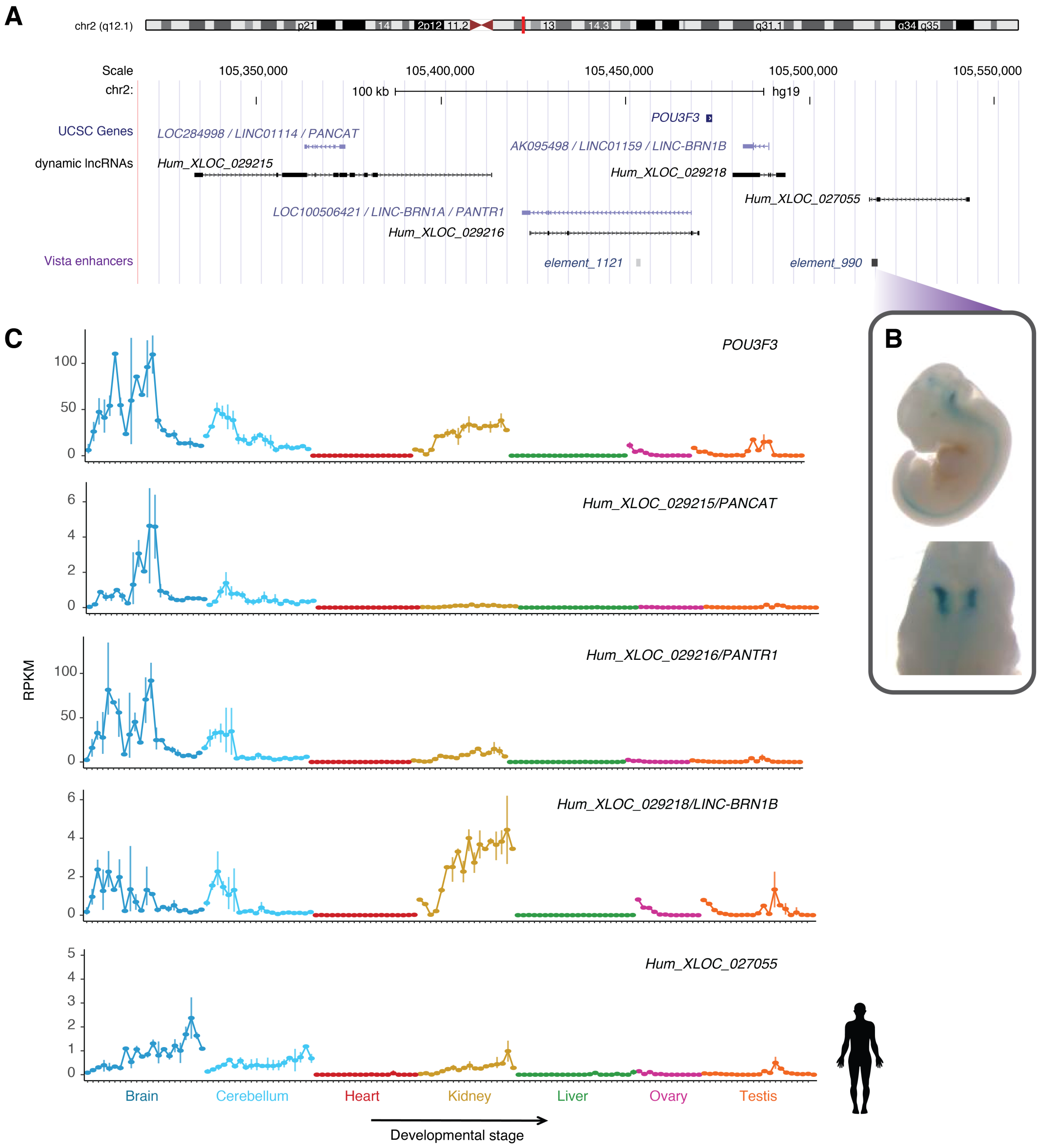


Figure 3.12: The protein-coding gene *POU3F3* shows correlated expression with multiple adjacent lncRNAs. **(A)** Genomic coordinates of *POU3F3* and adjacent co-expressed dynamic lncRNAs (adapted from UCSC genome browser, hg19). **(B)** *Hum_XLOC_027055* overlaps element_990, shown to have enhancer activity in the hindbrain (image from the VISTA enhancer browser). **(C)** Spatiotemporal expression profiles of *POU3F3* and adjacent co-expressed dynamic lncRNAs.

Collectively, in the first part of my thesis research I explored the expression profiles of lncRNAs during mammalian organ development, identifying candidates for further functional characterization and associating them with distinct biological processes and mechanisms of action. In the next chapter, I summarize the deeper characterization of a single lncRNA with a putative role in marsupial X chromosome inactivation.

3.2 A marsupial lncRNA associated with X-chromosome inactivation

The expression pattern of a gene can provide some information regarding its function, although such analyses are typically limited to its association with multiple and relatively broad terms (see section 3.1.3.2). A more detailed characterization of a gene's function requires extensive additional exploration at the cellular and molecular level. However, in some rare cases, gene expression patterns are so unique that they directly suggest an involvement in very specific biological processes.

In this project, which was carried out in collaboration with Dr. Mari Sepp and Dr. Tania Studer, and supervised by Dr. Margarida Cardoso-Moreira and Prof. Dr. Henrik Kaessmann, I investigated a newly identified opossum lncRNA, which we termed *Female Specific on chromosome X* (*FSX*), that shows such a unique and suggestive expression pattern. Dr. Tania Studer initially highlighted *FSX*, amongst other lncRNAs, as a candidate for being involved in X-chromosome inactivation (*240*). To test the plausibility of this hypothesis and to further characterize this lncRNA, I conducted all computational analyses described below. Dr. Mari Sepp performed the *in situ* hybridization experiments described in section 3.2.6.

3.2.1 *FSX* is a broadly expressed female-specific lncRNA on the X chromosome

Sex-biased expression of most genes is typically limited to only a single organ and the period following sexual maturity *(Rodríguez-Montes et al. in prep)*. Almost all of the genes with consistent female-biased expression across organs and developmental stages are found on the X-chromosome and involve protein-coding genes with a broadly expressed Y homolog (X gametologs) or components of the X chromosome inactivation machinery, such as *XIST* in eutherian mammals and *RSX* in opossum (*240*).

On the opossum chromosome X, six genes (four lncRNAs and two protein-coding genes) stand out as showing significantly higher expression in females compared to males across multiple organs and developmental stages **(Figure 3.13A)**. One of the four lncRNAs (*Opo_XLOC_045488*) corresponds to *RSX*, which has been previously shown to be involved in X-chromosome inactivation (*201*). Two of the remaining lncRNAs (*Opo_XLOC_044938* and *Opo_XLOC_045517*) are directly adjacent to the two female-biased protein-coding genes **(Figure 3.13B)**. This proximity, combined with an overall similarity in their spatiotemporal expression patterns **(Figure 3.13B, Supplementary Figure 3)** suggests that they could be involved in the regulation of these two protein-coding genes or be a byproduct of their expression.

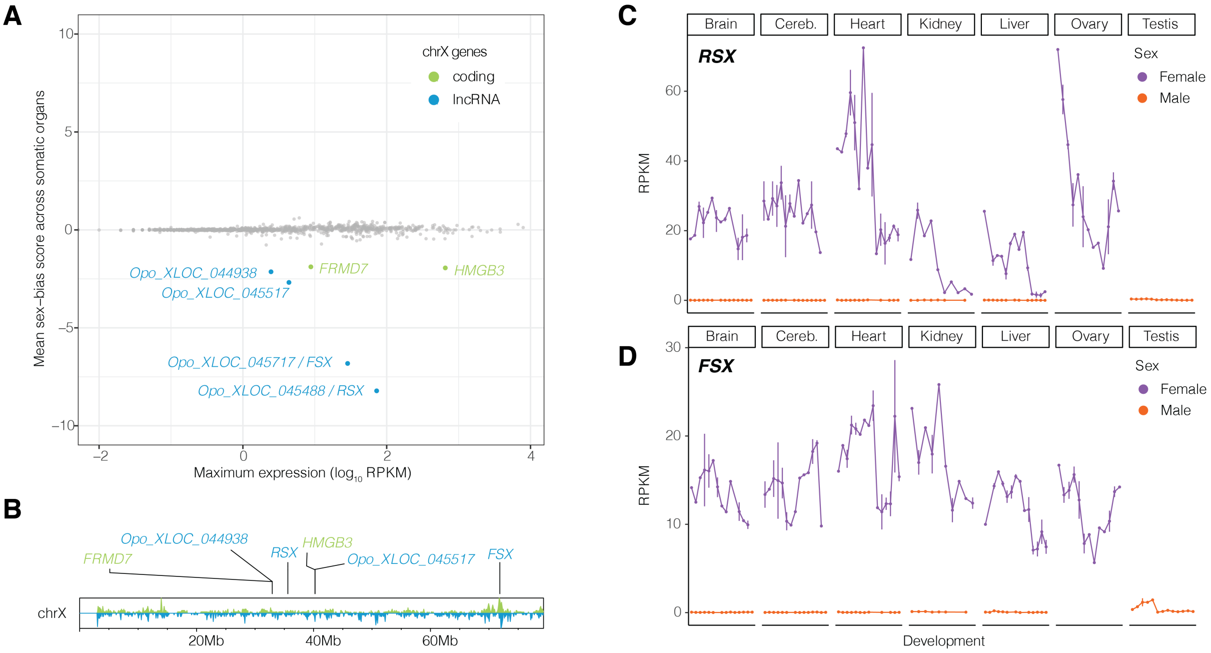


Figure 3.13: Developmental sex-biased expression of X-linked genes in the marsupial opossum. **(A)** Sex-bias score (mean across somatic organs) of X-linked protein-coding genes and lncRNAs. Genes with an absolute mean sex-bias score >= 1.5 are highlighted (green for protein-coding genes, blue for lncRNAs). **(B)** Position of genes with significant female-biased expression (from **A**) on the opossum X chromosome (coordinates from monDom5). The histograms show the gene density across the chromosome (green for protein-coding genes, blue for lncRNAs). **(C, D)** Expression profiles (RPKM) of *RSX* (**C**) and *FSX* (**D**) across developing organs and sexes.

However, the final female-biased lncRNA, *Opo_XLOC_045717* (from hereafter *FSX*) is located more than 30 Mb away from *RSX* or any other female-specific gene **(Figure 3.13B)**. Similar to *RSX*, it is highly expressed throughout the development of all female organs, and shows no expression in males, with the exception of barely detectable expression in early testis development **(Figure 3.13C-D)**. Although transcribed in antisense orientation to the protein-coding gene *WDR13*, there is no similarity between their expression profiles and *WDR13* shows no sex-biased expression **(Supplementary Figure 3)**. Collectively, with these analyses I was able to show that *FSX* is a female-specific, X-linked lncRNA that is broadly expressed in a likely autonomous manner. This expression pattern is reminiscent of that observed for *RSX* and *XIST* raising the possibility that *FSX* might also be involved in X-chromosome inactivation.

3.2.2 *FSX* expression starts at the onset of X chromosome inactivation

A recent study used scRNA-seq to map the onset of X chromosome inactivation in opossum around embryonic day 3.5 (*241*), ten days earlier than the earliest sample in the dataset used above (E13). As expected, *RSX* transcription also occurs around the same time (E3.5) and is followed by rapid silencing of the paternal X by E4.5 (*241*). Thus, if *FSX* is associated with X chromosome inactivation, it would be reasonable to expect its expression to start around the same time as that of *RSX*.

To test this hypothesis, I reanalyzed the scRNA-seq data to also measure the expression of *FSX* (which was not included in the transcriptome annotation used by the original study). In a striking similarity to *RSX*, *FSX* transcription is also restricted to female cells, starting at E3.5 and reaching high expression levels by E4.5 **(Figure 3.14)**. As an additional control, neither of the other four female-biased genes on the X **(Supplementary Figure 4A)**, nor any other gene in the opossum genome **(Supplementary Figure 4B)**, shows such high resemblance to the expression pattern of *RSX*.

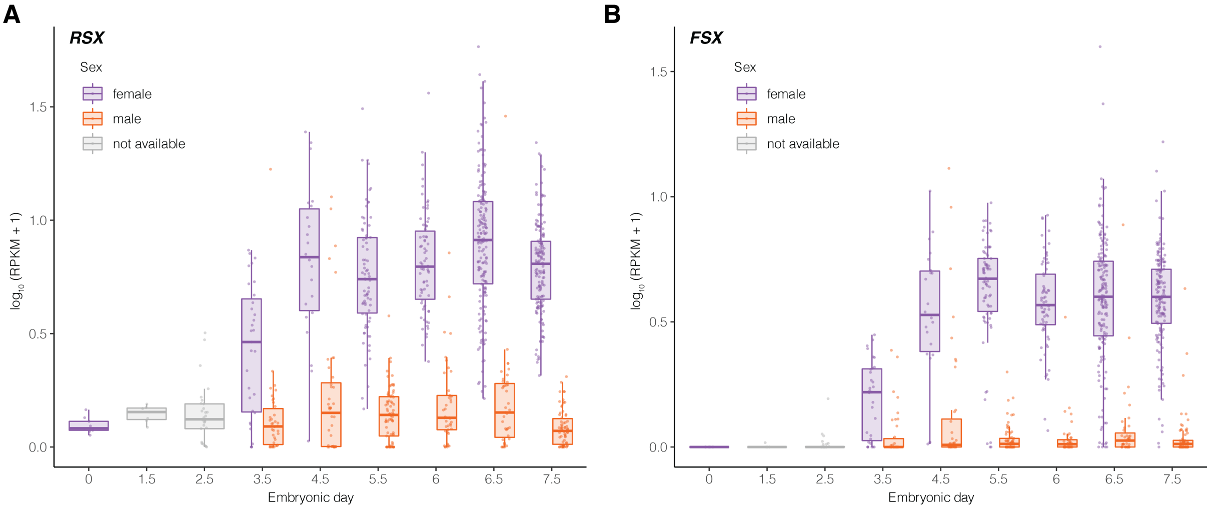


Figure 3.14: *RSX* and *FSX* expression at the onset of X-chromosome inactivation. Expression profiles (RPKM) of *RSX* **(A)** and *FSX* **(B)** in single-cells from early opossum embryos.

3.2.3 *FSX* shares features with other XCI-associated lncRNAs

Given the high similarity observed in the expression patterns of *RSX* and *FSX*, I sought to investigate whether the two lncRNAs also show sequence similarity. This would be compatible with a duplication event producing two copies of the same lncRNA or could suggest technical artifacts, such as the spurious mapping of *RSX* reads to the *FSX* transcript. However, using both a local (BLAST) and an optimal global (EMBOSS Needle) sequence alignment tool, I was unable to recover any significant sequence similarity between the two lncRNAs **(Figure 3.15A)**. This result lends further support to the notion that *FSX* expression is autonomous. Although — in theory — the sequences of related lncRNAs could diverge beyond detection of any sequence similarity, the large genomic distance between the two loci (35 Mb) further suggests that the two lncRNAs are most likely to have emerged independently.

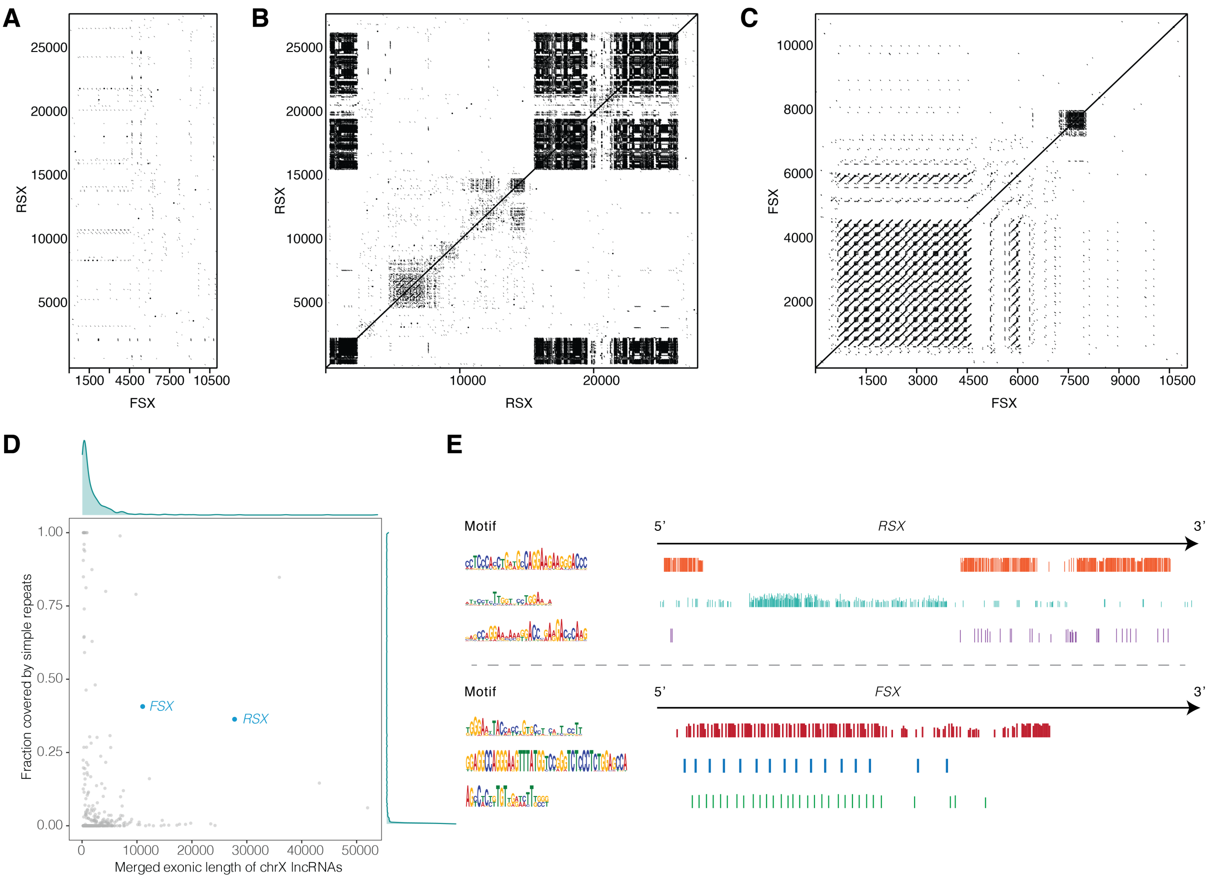


Figure 3.15: *RSX* and *FSX* show no sequence similarity but are both enriched for simple repeats. **(A)** Dotplot illustrating sequence similarity between the aggregated non-redundant exonic sequence of *FSX* (x-axis) and *RSX* (y-axis). Sequence similarity is averaged across 10 bp-windows; regions with a similarity score ≥ 50 are shown in black. **(B, C)** Dotplots illustrating sequence similarity within the *RSX* **(B)** and *FSX* **(C)** transcripts (i.e., sequence repetition). Sequence similarity is averaged across 10 bp-windows; regions with a similarity score ≥ 50 are shown in black. **(D)** Length of the aggregated non-redundant exonic sequences (x-axis) and its fraction covered by UCSC-annotated simple repeats (y-axis) for X-linked lncRNAs. Density plots (top and right) show the distribution of these metrics. *FSX* and *RSX* are both outliers in both metrics. **(E)** Sequence (left) and position (right) of the three most significantly enriched short (<50 bp) repeats for *RSX* (top) and *FSX* (bottom).

Marsupial *RSX* and eutherian *XIST* also share no sequence similarity but are marked by some distinctive features that are relatively uncommon for lncRNA transcripts (*201*, *242*). Both are quite long (25 and 17 kb respectively) and are enriched for simple — albeit different for each lncRNA — repeats (*201*). In the case of *XIST,* these repeats have been shown to be required for interactions with different proteins that in turn lead to the silencing of the inactive X-chromosome (*90*, *200*). I thus decided to assess whether *FSX* would also show these two features. Both *RSX* and *FSX* show significant similarity within their own sequences, suggesting the presence of repetitive sequences **(Figure 3.15B-C)**. Additionally, similarly to *RSX*, *FSX* is also quite long (11 kb), ranking amongst the longest 5% of opossum lncRNAs found on the X chromosome. More than 30% of both lncRNA transcripts is covered by simple repeats **(Figure 3.15D)**, an observation very uncommon amongst similarly long lncRNAs (only two more X-linked opossum lncRNAs show higher fraction of repeats in the same length range; **Figure 3.15D**). Although the core repeated motifs differ between *RSX* and *FSX* — as they also do between *RSX* and *XIST*, they are both GC-rich (**Figure 3.15E**).

Taken together, these analyses show that even though *RSX* and *FSX* are not homologous to each other, their transcripts share features that are also found in *XIST* and are thought to be associated with a role in X-chromosome inactivation.

3.2.4 Conservation of *FSX* sequence and synteny across marsupials

Since X-chromosome inactivation emerged independently in marsupials, its main regulators are expected to be shared across marsupial genomes but not found in eutherians or monotremes. Thus, I sought to investigate the evolutionary history of *FSX* and, for comparison, *RSX*. First, I tested whether I could detect any significant sequence similarity for these two lncRNAs in a selection of marsupial, eutherian and monotreme genomes. All marsupial genomes have regions showing significant similarity to both *RSX* and *FSX* **(Figure 3.16)**. By contrast, the only significant hit I detected outside of marsupials was for a part of *RSX*'s repetitive region, which shows some similarity to a region in the monotreme X1 (note that the ancestral therian X in monotremes is the autosomal chromosome 6). However, this part of the opossum *RSX* sequence is not shared with other marsupials and is thus more likely a false alignment hit or a secondary incorporation into the *RSX* transcript that occurred after the split between opossums and other marsupial lineages around 80 million years ago.

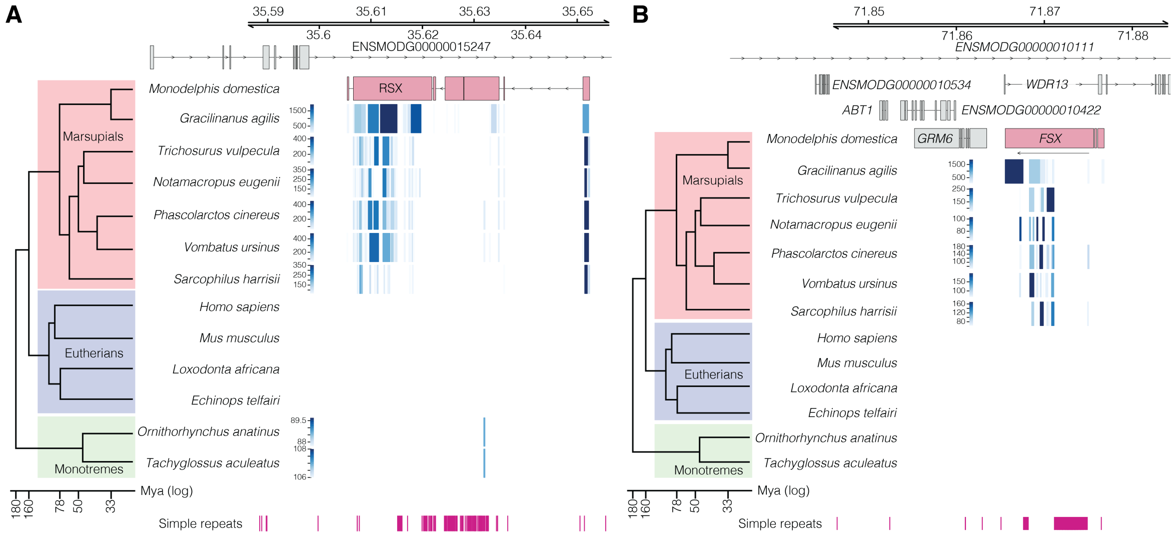


Figure 3.16: Conservation patterns of *RSX* (A) and *FSX* (B) sequences across marsupial, eutherian and monotreme genomes. The transcript models show the aggregated non-redundant exonic sequences in opossum. Regions showing significant (E-value < 10^{-5}, identity ≥ 10%, length ≥ 50 nt) sequence similarity, as determined by BLAST, in other species are color-coded based on the alignment score (bitscore). The *RSX* region showing significant sequence similarity in monotremes does not represent a reciprocally unique alignment between the species. Evolutionary relationships between the species are shown in the phylogenetic tree on the left (time in million years, log-scaled). The position of UCSC-annotated simple repeats is shown at the bottom of the genome browser.

Having identified the genomic loci that show significant sequence similarity to *RSX* and *FSX* in different marsupials, I next asked whether these loci were consistently found on chromosome X, as would be expected for the main regulators of the X-inactivation machinery. This can be easily addressed for species with chromosome-level assemblies, such as the Tasmanian devil and common brushtail possum (where I found that *FSX* indeed localizes on the X), but is more challenging when assemblies are fragmented into thousands of scaffolds. Thus, to answer this question systematically, I resorted to assessing the overall conservation of the genomic neighborhood around these two lncRNAs. If the conserved regions were consistently found around the same genes, it would be reasonable to assume that this region was still on chromosome X. Indeed, despite the challenges posed by the incomplete annotation of marsupial genomes, I was able to show that both *RSX* and *FSX* are consistently found around the same genes, in the same order and relative orientation (**Figure 3.17**).

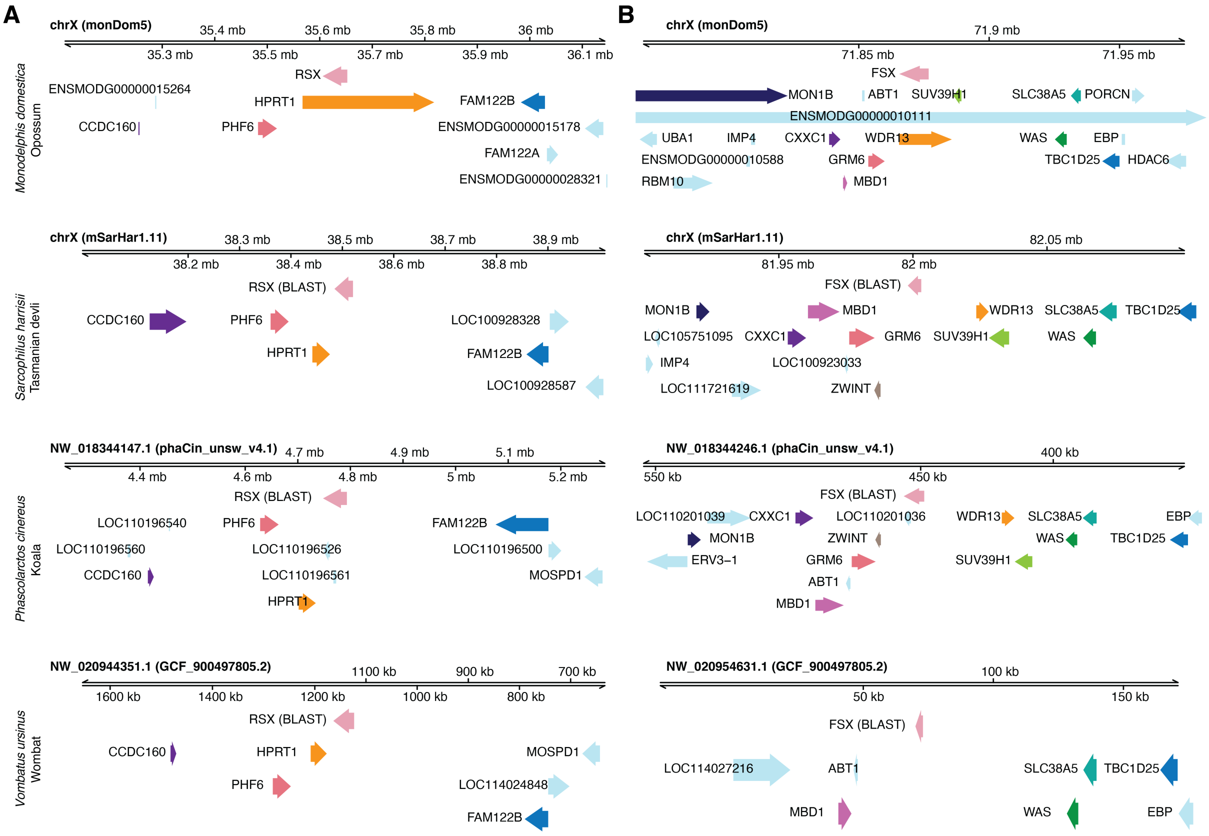


Figure 3.17: Conservation of genomic neighborhood around *RSX* (A) and *FSX* (B) sequences across marsupial genomes. The genome browser displays 500 kb (*RSX*) and 100 kb (*FSX*) from each side of the lncRNA genes (in opossum) and their aligned sequences (in the other species). A smaller window is shown for *FSX* due to the higher gene density of the locus. Some scaffolds have been inverted to show *RSX* and *FSX* in the negative strand, for consistency with the opossum genome. Orthologous genes are consistently color-coded across species.

Collectively, both *RSX* and *FSX* appear to have emerged in the last common ancestor of all living marsupials and to have been preserved on chromosome X ever since, in line with a possible role in X-chromosome inactivation.

3.2.5 *FSX* female-specific expression is conserved across 76 million years

If *FSX* is indeed involved in marsupial X-chromosome inactivation, its female-specific expression should also be preserved across marsupial species. Unfortunately, the availability of transcriptomics datasets, especially covering both sexes, for marsupials is very limited. However, a previous study has generated bulk RNA-seq data for several organs from two koala individuals (*243*). Koalas, as all Australian marsupials, have diverged from *Monodelphis* around 76 million years ago, a distance comparable to that between human and mouse (90 million years; *source timetree.org*). Despite a limited overlap in the sampled tissues, important differences in the sampling methodology, and the unavoidable confounding of differences between sexes and individuals in the absence of biological replicates (*243*), I sought to test whether the *FSX*-aligned region in koala would be supported by RNA-seq reads and whether it would show female-specific expression. Strikingly, both *RSX* and *FSX* regions were highly transcribed in all female tissues and showed no expression in the male samples (**Figure 3.18**). Thus, the broad and female-specific expression observed for both lncRNAs in opossum has been preserved for at least 76 million years of marsupial evolution.

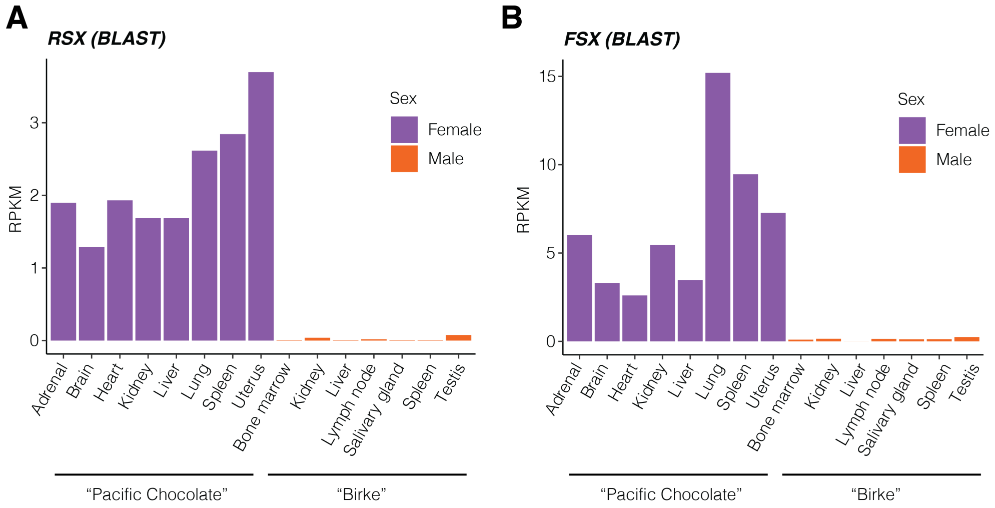


Figure 3.18: Conservation of female-specific expression in koala. (**A, B**) Expression (in RPKM) quantified across the regions of the koala genome showing significant sequence similarity to *RSX* (**A**) and *FSX* (**B**) across organs and individuals from different sexes.

3.2.6 *FSX* localizes on the inactive X chromosome

All results from the analyses presented above are concordant with *FSX* being involved in X-chromosome inactivation. To gain further support for this hypothesis at the cellular level, we decided to use *in situ* hybridization to examine the subcellular localization of the opossum *FSX* RNAs. Both *XIST* and *RSX* have been previously shown to localize on the eutherian and marsupial inactive X respectively, which in both lineages forms a condensed heterochromatic structure also known as the "Barr body". Tissue preparations, stainings and microscopy presented in this section were performed by Dr. Mari Sepp. I contributed by designing the probes and by participating in the planning and interpretation of the experiments.

The hybridization chain reaction (HCR) experiments confirmed that both *FSX* and *RSX* are exclusively expressed in female cells and further showed that both lncRNAs are predominantly nuclear (**Figure 3.19A**). Although the molecules detected for *FSX* are consistently fewer compared to *RSX*, both lncRNAs consistently co-localize with the Barr body, which was detected by co-staining for the heterochromatin-associated histone mark H3K27me2/3 (**Figure 3.19B**). As an additional control, *FSX* transcripts are separated from the transcription sites of the escaper gene *HMGB3*, which were identified by targeting intronic sequences in nascent *HMGB3* transcripts **(Figure 3.19C)**. Thus, *FSX* transcripts are located away from the active X, as well as from euchromatic regions of the inactive X that facilitate the transcription of escaper genes. Collectively, the nuclear localization of *FSX* and its spatial proximity to the Barr body lend further support to the hypothesis that it is involved in marsupial X-chromosome inactivation. On the other hand, the lower stoicheiometry compared to the *XIST*-like *RSX* poses additional questions regarding the potential mode of action of *FSX* in this process.

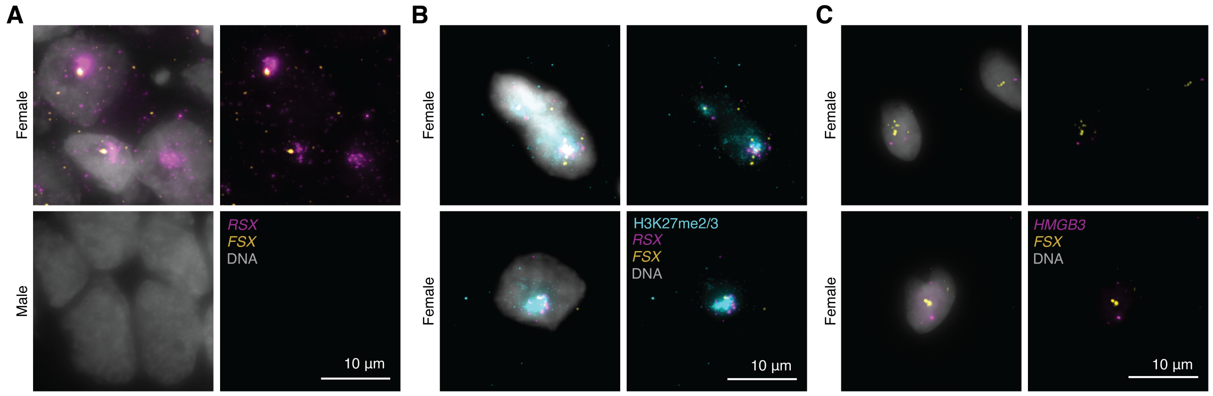


Figure 3.19: Subcellular localization of *FSX* in opossum. (A) Tissue sections from female (top) and male (bottom) opossum brain samples stained with *RSX* and *FSX* HCR probes. **(B)** Co-staining of isolated female opossum nuclei for *RSX* and *FSX* transcripts (HCR) and H3K27me2/3 (immunostaining). **(C)** HCR staining in isolated female opossum nuclei for *FSX* and an intron of *HMGB3* marking sites of active transcription on the active and inactive X-chromosomes. *All images show maximum projections across Z-stacks, with (left) and without (right) Hoechst staining for DNA. Experiments and microscopy were performed by Dr. Mari Sepp.*

Addressing these questions requires additional experiments, such as perturbing the expression of *FSX* and *RSX*, both individually and in combination, and assessing the effects of these perturbations on the expression of X-linked genes. These experiments are currently being designed and, when performed, will definitively assess the involvement of *FSX* in marsupial X-chromosome inactivation and shed more light into its precise role and its potential interaction with *RSX*.

3.3 Cell type-specific gene regulation in the developing mouse cerebellum

The goal of this third part of my thesis work was to identify putative *cis*-regulatory elements (CREs) across cell types in the developing mouse cerebellum and to characterize their chromatin accessibility dynamics during cell fate specification and differentiation. To this end, I analyzed a snATAC-seq dataset of ~90,000 cells covering 11 stages of mouse cerebellum development, from early neurogenesis to adulthood (**Figure 3.20A**). All data presented here were generated by Dr. Mari Sepp and Robert Frömel. All computational analyses are the product of my own work, under the supervision of Prof. Dr. Henrik Kaessmann, Prof. Dr. Stefan Pfister and Dr. Margarida Cardoso-Moreira. The findings described in this part were published in *Sarropoulos, Sepp et al. 2021* (*244*).

3.3.1 Quality control and cell type annotation

The first step of this project was to assess the quality of the dataset, identify cells and annotate them based on their cell type and state. At the level of individual snATAC-seq libraries, I observed the expected periodicity in the fragment length distribution and an enrichment of insertions around annotated TSSs, demonstrating a high signal-to-noise ratio (**Supplementary Figure 5**). I then proceeded to detect barcodes corresponding to high-quality cells and to remove putative doublets, identifying a total of 91,922 cells. For each cell, I quantified its chromatin accessibility profile across the genome, tiled into 500 bp-wide windows (*245*). Based on these estimates, I projected the cells into a low-dimensional embedding using an iterative latent semantic indexing (LSI) procedure, a technique commonly used in natural language processing (*62*, *245*). To facilitate the visualization of the data, this embedding can be further condensed into a two-dimensional projection using Uniform Manifold Approximation and Projection (UMAP). Cells sharing similar chromatin accessibility profiles appear close to each other in such a projection. Reassuringly, biological replicates (i.e., samples from different individuals but the same developmental stage) show high similarity in the UMAP embedding, further supporting the quality of the dataset (**Figure 3.20B**).

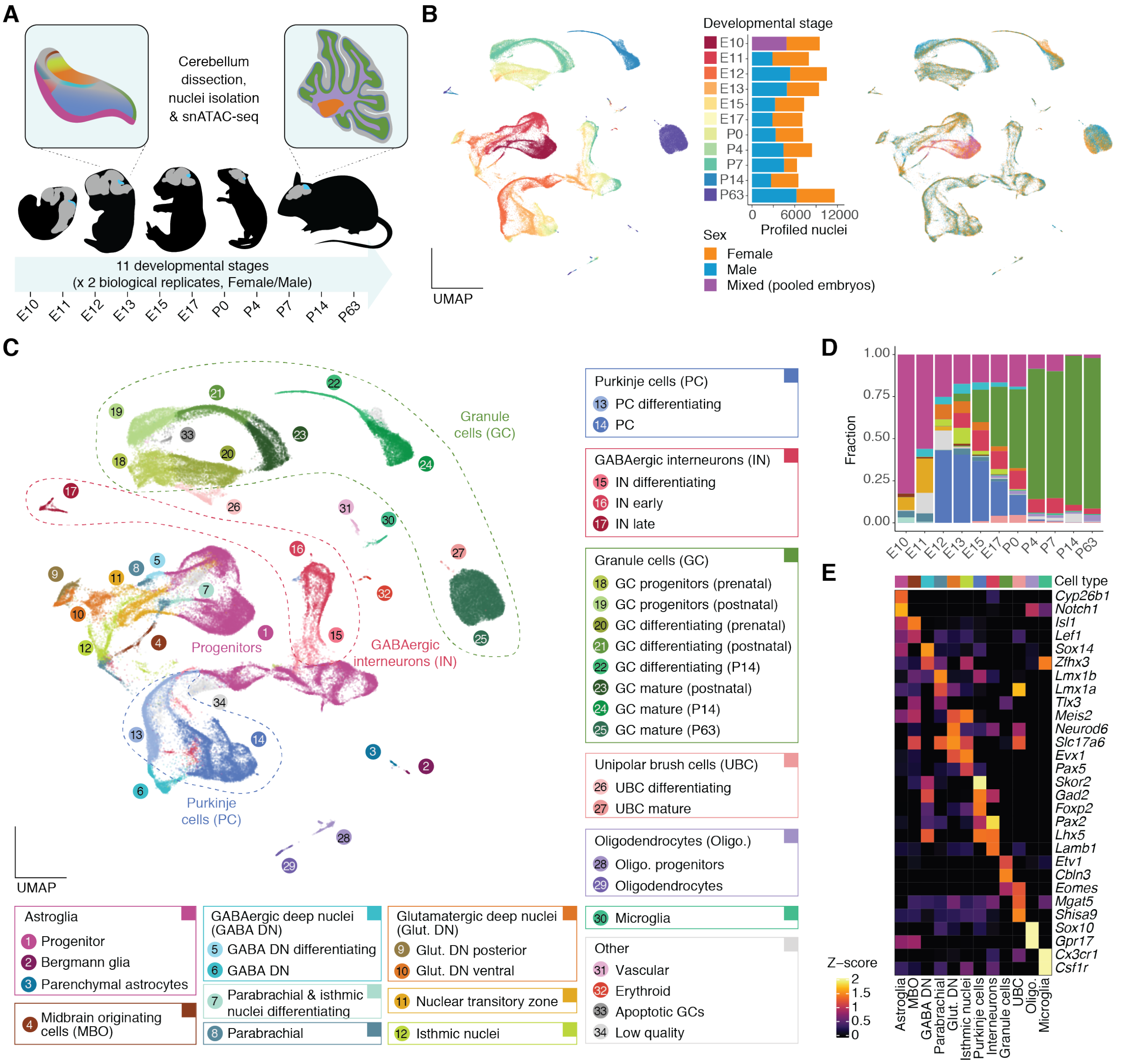


Figure 3.20: A snATAC-seq atlas of mouse cerebellum development. **(A)** Schematic overview of the dataset. Representative mouse silhouettes are shown for E11, E13, E17, P4 and P63 (brain in grey, cerebellum in cyan). The insets show the location of selected cell types in the cerebellum (colors are as in **C**). **(B, C)** UMAP projection of 91,922 cells colored by developmental stage (**B**, left) or sex (**B**, right), or cell type and state (**C**). Barplots in B show the number of profiled cells per stage and sex (each sex corresponding to one sample). In **C**, cell states or subtypes (numbered circles) are grouped into broad cell types (rectangles). **(D)** Proportions of broad cell types across developmental stages. **(E)** Activity scores of genes used for the annotation of broad cell types (Z-score, capped to 0-2). Broad cell type colors for **D** and **E** are as in **C**. *Figure reproduced from Sarropoulos et al. 2021* (*244*).

To annotate the identified cells based on their cell type and state, I clustered them into groups of cells with similar accessibility profiles. Aggregating the chromatin accessibility signal around a gene to infer “gene scores”, as a proxy for gene expression (*245*), I was then able to link these clusters to known cerebellar cell types and states. This process involved manual investigation of the available literature (*130*, *210*, *211*, *246*, *247*) and gene expression databases (*248*) and was greatly assisted by Dr. Mari Sepp and Kevin Leiss. In total, we identified 12 broad cell types and 42 subtypes / cell states **(Figure 3.20C-E)**.

To assess the quality of this annotation, as well as the overall utility of gene score estimates to approximate gene expression, I compared the snATAC-seq data to a previously published scRNA-seq dataset of mouse cerebellum development (*210*). I integrated the two datasets using canonical correlation analysis (CCA) as implemented in Seurat (*249*) and used the similarity to the scRNA-seq dataset to predict cell type labels and to impute RNA expression values for the cells in the snATAC-seq dataset. I then compared these predictions to our cell type annotation, observing an overall good agreement, despite differences in dissections and sampled developmental stages between the two studies **(Figure 3.21A)**.

Some discrepancies between the two annotations (e.g., cells annotated as "GABAergic deep nuclei neurons" in the snATAC-seq dataset matching "excitatory cerebellar nuclei neurons" in the scRNA-seq dataset) could be explained by scRNA-seq clusters containing mixtures of similar cell types. After reanalyzing the scRNA-seq data, I could observe that around half of all cells annotated as "excitatory cerebellar nuclei neurons" at E14 are in fact positive for markers of GABAergic deep nuclei neurons, thus explaining the mismatch between labels **(Figure 3.21B)**. Similarly, more than half of the scRNA-seq cells annotated as "unipolar brush cells" (UBCs) at P0 are negative for UBC markers, such as *Lmx1a*, and show expression of genes highly expressed in differentiating granule cells **(Figure 3.21C)**. Unsurprisingly, differentiating granule cells in the snATAC-seq dataset show high overlap with UBCs in the scRNA-seq dataset. Despite these discrepancies, the majority of cell type labels agree well between the two annotations **(Figure 3.21A)**. Similarly, in most developmental stages, gene score estimates are highly correlated to the RNA expression values imputed from the integration with the scRNA-seq data, demonstrating the utility of using gene scores as a proxy for gene expression **(Figure 3.21D)**.

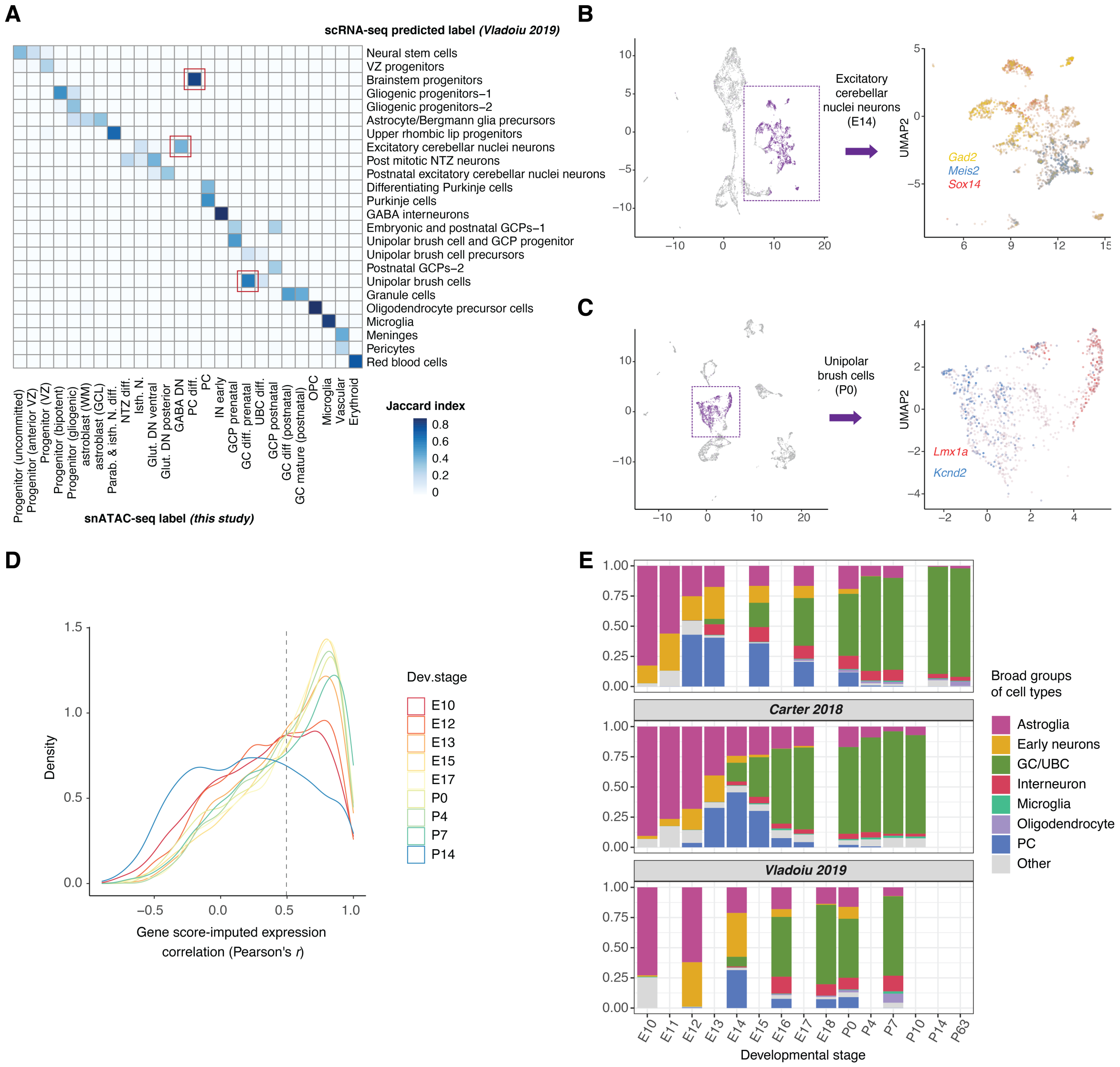


Figure 3.21: Comparison with scRNA-seq data. (**A**) Jaccard similarity index between cell type labels from this study (columns) and predicted labels after integration with scRNA-seq data (rows). Only labels with a similarity index of at least 0.15 with at least one other group are shown. The red rectangles mark unexpected matches analyzed further in **B**, **C**. (**B, C**) UMAP projections of 6,068 cells from E14 (**B**) and 4,809 cells from P0 (**C**) cerebellum profiled by scRNA-seq. Cells annotated by the authors as excitatory cerebellar nuclei neurons (**B**) or UBCs (**C**) are marked in purple. Only a fraction of these cells are positive for the corresponding marker genes. (**D**) Per gene correlation (Pearson's *r*) between gene score and imputed expression (after integration with scRNA-seq data) for highly variable genes in the scRNA-seq data. The vertical line indicates the median correlation coefficient across developmental stages. (**E**) Proportions of broad cell type groups across developmental stages in our snATAC-seq atlas (top), and scRNA-seq atlases by *Carter et al. 2018* (middle) and *Vladoiu et al. 2019* (bottom). Cell types were grouped into broad categories to facilitate comparisons between the three studies that differ in annotation strategies and resolution. *Figure adapted from Sarropoulos et al. 2021* (*244*).

In agreement with previous studies (*209*, *210*), most cells in the earliest developmental stages (E10-E11) are neural progenitors. Additional non-dividing cell populations, the most abundant of which are glutamatergic and GABAergic deep nuclei neurons, account for the remaining cells in these stages. E12 marks the appearance of a large number of differentiating Purkinje cells, which remain the most abundant cell type until E15. From E13 onwards, GABAergic interneurons and granule cells also appear and gradually outnumber Purkinje cells. While the fraction of GABAergic interneurons remains relatively stable until P7, the granule cell population expands rapidly, becoming the most abundant cell type by E17 and accounting for ~90% of all cells in the cerebellum in the last two stages (P14, P63). Small numbers of glial cells (astroglia, which includes parenchymal astrocytes and Bergmann glia, oligodendrocytes and microglia) are also traceable in postnatal stages. Collectively, cell type proportions in this dataset follow similar dynamics to those previously reported by scRNA-seq studies of the developing mouse cerebellum **(Figure 3.20D, Figure 3.21E)**.

Altogether, with these analyses I was able to assess the quality of the mouse snATAC-seq dataset, identify and annotate cells, and then validate these annotations based on previous studies (*209, 210*). This part serves as the foundation for the following sections.

3.3.2 Identification and characterization of CREs in the developing cerebellum

I next sought to identify open chromatin regions, as a proxy for putative CREs, across cell types and developmental stages. To be able to detect cell type-specific CREs without being biased by the proportions of each cell type in the organ, I aggregated chromatin accessibility profiles across cells from the same cluster (i.e., cell type and state) and sample (i.e., individual) and used MACS2 (*245, 250*) to call peaks of open chromatin **(Figure 3.22A)**. To ensure the reproducibility of the identified CRE annotation, I required each peak to be called in at least two samples for the same cluster (i.e., to be called in at least two replicates of a given cell type/state; **Figure 3.22A**). Peak annotations for each cluster were then merged in an iterative way into a non-redundant annotation of 499,146 peaks. To further exclude "noisy" peaks primarily originating from very abundant cell types (and thus large sequencing depth), I implemented an additional filtering step, requiring peaks to be accessible in at least 5% of cells in at least one cluster **(Figure 3.22A)**. This led to a total of 261,642 high-confidence putative CREs, which I used for all subsequent analyses **(Figure 3.22A-B)**.

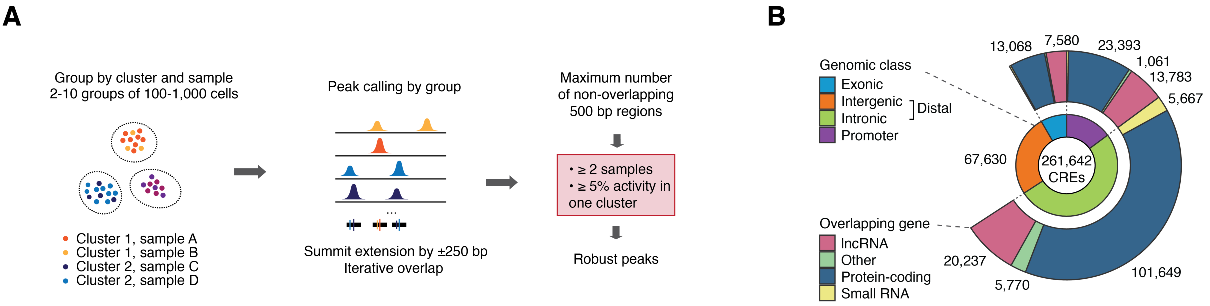


Figure 3.22: Identification of putative CREs in the mouse cerebellum. **(A)** Schematic representation of the procedure followed for the identification and filtering of putative CREs. **(B)** Genomic features of 261,642 putative CREs. Inner circle: genomic class; outer circle: biotype of the overlapping gene. *Figure adapted from Sarropoulos et al. 2021* (*244*).

I then characterized CREs based on their position compared to annotated protein-coding and non-coding genes **(Figure 3.22B)**. I found that most CREs are intronic (51%) or intergenic (26%), which I collectively refer to as "distal". Promoters (defined as being within -2,000/+100 bp from a TSS) account for 15% of the identified CREs, with the remaining 8% overlapping exonic regions. Although exonic CREs have been suggested to function in similar ways to other distal elements (*251*), their overlap with exons, especially in the case of protein-coding genes, imposes additional constraints on their sequence, which can affect subsequent sequence-based analyses. I thus decided to be cautious and treat them as a separate group.

To assess the quality of this CRE annotation, I examined a series of available resources. I first considered genomic annotations of regulatory activity across different tissues and developmental stages based on data from the ENCODE project (*150*). Regions annotated as "strong enhancers" in developing brain tissues show a high recall (70%-80%) in the cerebellar CREs, with the highest enrichment observed for the hindbrain **(Figure 3.23A)**. By contrast, cerebellar CREs are depleted of heterochromatin-associated regions in the hindbrain compared to other organs **(Figure 3.23B)**. Cerebellar CREs also show the highest recall (97%) of experimentally validated hindbrain enhancers (*239*), followed by other neuronal tissues **(Figure 3.23C)**. Similarly, by comparing to a collection of transcribed enhancers (eRNAs) across multiple conditions (*252*) and to a snATAC-seq atlas of adult mouse organs (*62*), I found that the CREs I identified in this study consistently show the highest overlap with elements active in the cerebellum **(Figure 3.23D-E)**. Collectively, these analyses support the high quality and reproducibility of the identified CREs.

A major challenge in the study of distal CREs is to associate them with their target genes, which can often be found hundreds of kilobases away in the linear genome (*36*, *37*). Single-cell datasets can be used to tackle this problem by offering statistical power to estimate correlations between

CRE accessibility and the expression of nearby genes. Although correlations by themselves are not sufficient to infer causal regulatory interactions, they offer significant improvement compared to simply considering the nearest gene (*245*, *253*). I thus decided to use this dataset to assign distal CREs to their putative targets based on correlations with their promoter accessibility and gene score (as a proxy for gene expression).

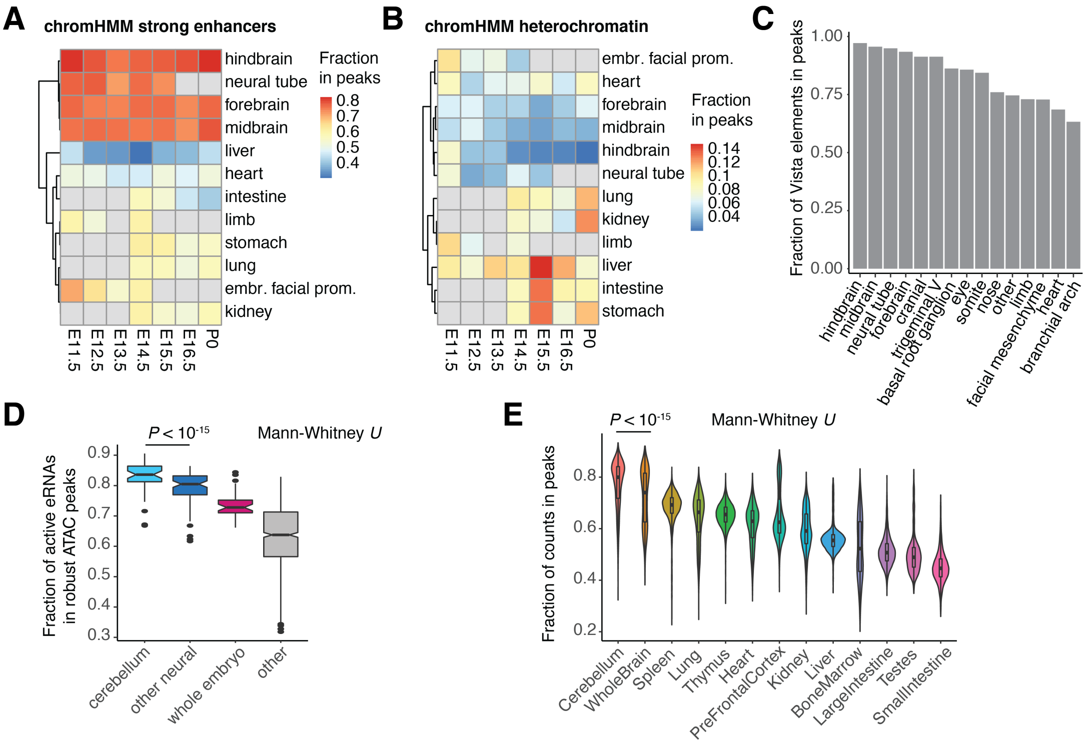


Figure 3.23: Comparison of identified CREs to other datasets. (**A, B**) Fraction of chromHMM predicted strong enhancers (**A**) and heterochromatin (**B**) across a series of tissues and developmental stages recalled in the robust CRE set from this study. Gray values: data not available. (**C**) Fraction of experimentally validated enhancers in mouse embryonic tissues overlapping robust CREs from this study. Tissues are ordered by decreasing fraction. (**D**) Fraction of expressed eRNAs recalled in the robust CRE set from this study across samples from the developing cerebellum, other neural tissues, whole embryo development and all other mouse samples. (**E**) Per-cell fraction of fragments in regions overlapping robust CREs from this study across different organs in the adult mouse. Tissues are ordered by decreasing median fraction across cells. *Figure adapted from Sarropoulos et al. 2021* (*244*).

After assigning a total of 32,792 distal CREs to 5,766 putative target genes (**Figure 3.24A**), I sought to assess the confidence of these assignments. By considering a dataset of chromatin interactions in neural progenitors (*115*), I showed that the identified CRE-gene pairs were more likely to share the same TAD (**Figure 3.24B**). Additionally, CRE-gene pairs show significantly higher correlations in the expression of promoter-derived transcripts (i.e., genes) and transcribed enhancers (eRNAs) in samples from the developing cerebellum (*252*) (**Figure**

3.24C). Distal CREs assigned to a target gene are also more likely to overlap a computationally predicted (*150*) or experimentally validated (*239*) hindbrain enhancer (**Figure 3.24D-E**). Taken together, these results show that my CRE-gene assignment approach enriches for *bona fide* regulatory interactions.

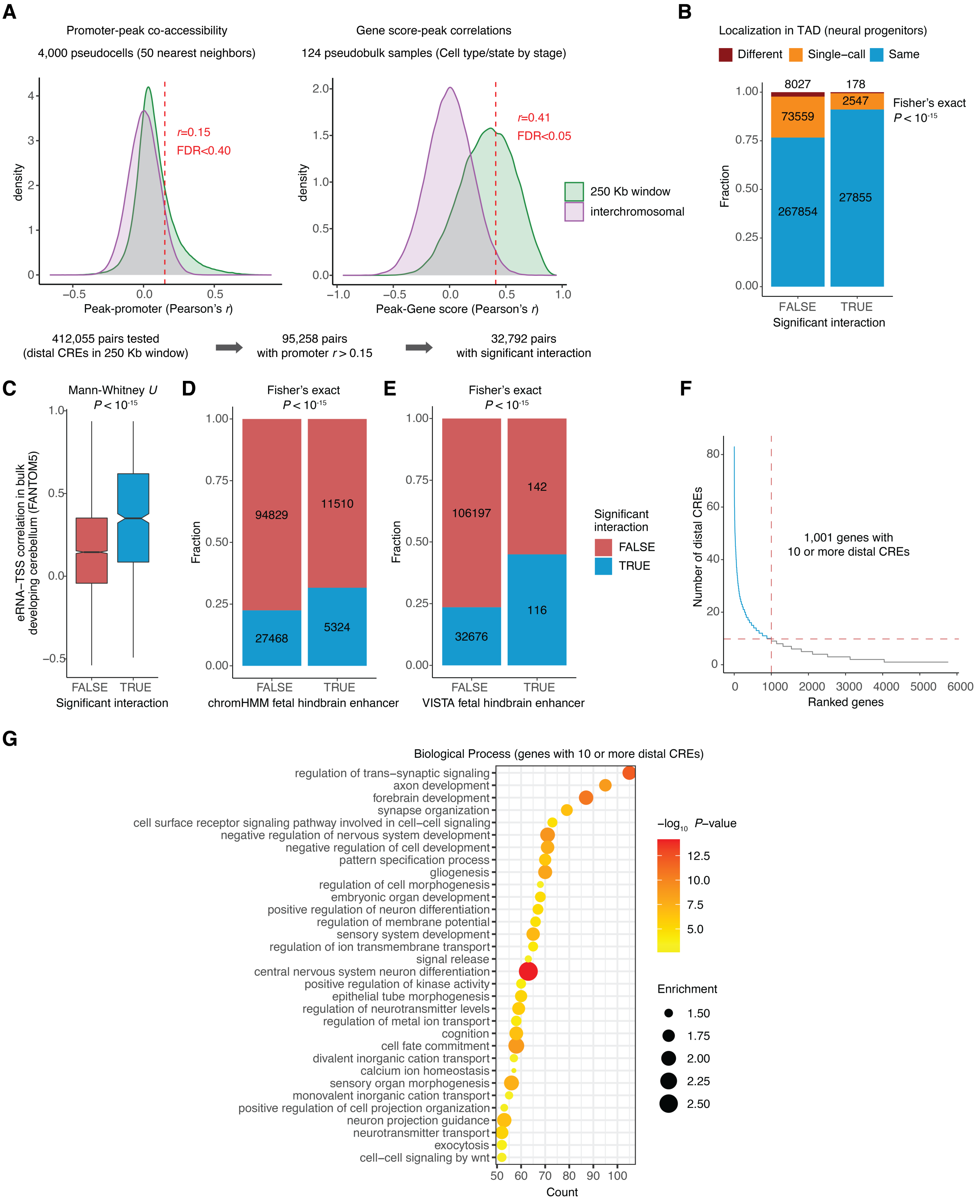


Figure 3.24: Assignment of CREs to putative target genes *(see next page for caption)*.

Figure 3.24: Assignment of CREs to putative target genes. (**A**) Correlation-based assignment of distal CREs to putative target genes. Left: Pearson's correlation coefficients between distal CRE and promoter accessibility in 250 kb windows (green) or across different chromosomes (purple). Right: Pearson's correlation coefficients between distal CRE accessibility and gene score for CRE-gene pairs with a promoter-peak correlation of $r \geq 0.15$ (left, maximum correlation for genes with multiple promoters) in 250 kb windows (green) or across different chromosomes (purple). Interchromosomal correlations were used to construct a null distribution and significant CRE-gene interactions were identified with $r \geq 0.41$ (BH adjusted $P < 0.05$). (**B**) Fraction of distal CRE-promoter pairs in the same (blue) or different (red) topologically associating domain (TAD) in mouse neural progenitors stratified by significance of interaction between CRE-promoter pairs. Single-call (orange): only one region was assigned to a TAD. (**C**) Pearson's correlation coefficients between the expression of eRNAs overlapping distal CREs from this study, and promoter associated RNAs across cerebellum development stratified by significance of interaction between CRE-promoter pairs. (**D, E**) Fraction of distal CREs assigned to a gene for elements overlapping putative enhancers in hindbrain development (**D**) or experimentally validated enhancers in the E11 hindbrain (**E**). (**F**) Genes ranked in decreasing order by the number of distal CREs assigned to them. (**G**) Biological process enrichment for genes associated with 10 or more distal CREs (from **F**). The x-axis indicates the number of genes associated with each term, the size and color of the dots show the effect and significance of the enrichment based on a hypergeometric test. *Figure adapted from Sarropoulos et al. 2021* (*244*).

3.3.3 CRE activity is shaped by both cell type and developmental stage

I next sought to determine the major patterns of spatiotemporal CRE activity in the developing mouse cerebellum. To this end, I aggregated the accessibility of CREs across cell types and developmental stages, scaled each CRE to a maximum value of 1 and performed a two-step clustering procedure (k-means, followed by hierarchical clustering) to identify 26 CRE clusters. Most of these CRE clusters are specific to a single cell type and developmental window, highlighting the overall high context-specificity of CRE activity (**Figure 3.25A**). CREs in cell type-specific clusters are close to genes associated with relevant gene ontology terms, such as myelination for oligodendrocytes and immune response for microglia. Similarly, cell type-specific CREs are also enriched for motifs of TFs known to be active in the respective cell types (e.g., SOX2 in progenitor cells, ATOH1 in cell types derived from the rhombic lip, SOX10 for oligodendrocytes, PU.1 for microglia). In contrast to these cell type-specific CREs, I also identified clusters of CREs that are active in multiple cell types. These include the early-born neurons (clusters 2, 11), glial populations (cluster 18) and late-born cell types (cluster 14). Finally, one group of CREs (cluster 12) shows constitutive activity throughout the dataset. Most (67%) of the CREs in this cluster are promoters (**Figure 3.25B**). Furthermore, more than 50% of the remaining distal CREs from cluster 12 contain a CTCF motif (**Figure 3.25C**), suggesting that they might be involved in the regulation of chromatin architecture. For example, many of these CREs could correspond to TAD boundaries, which are also known to show low variance across biological contexts (*152*). Collectively, these results show that despite the high specificity observed for most CREs, sizeable sets also show pleiotropic activity, in agreement with previous reports (*149*, *229*, *254*).

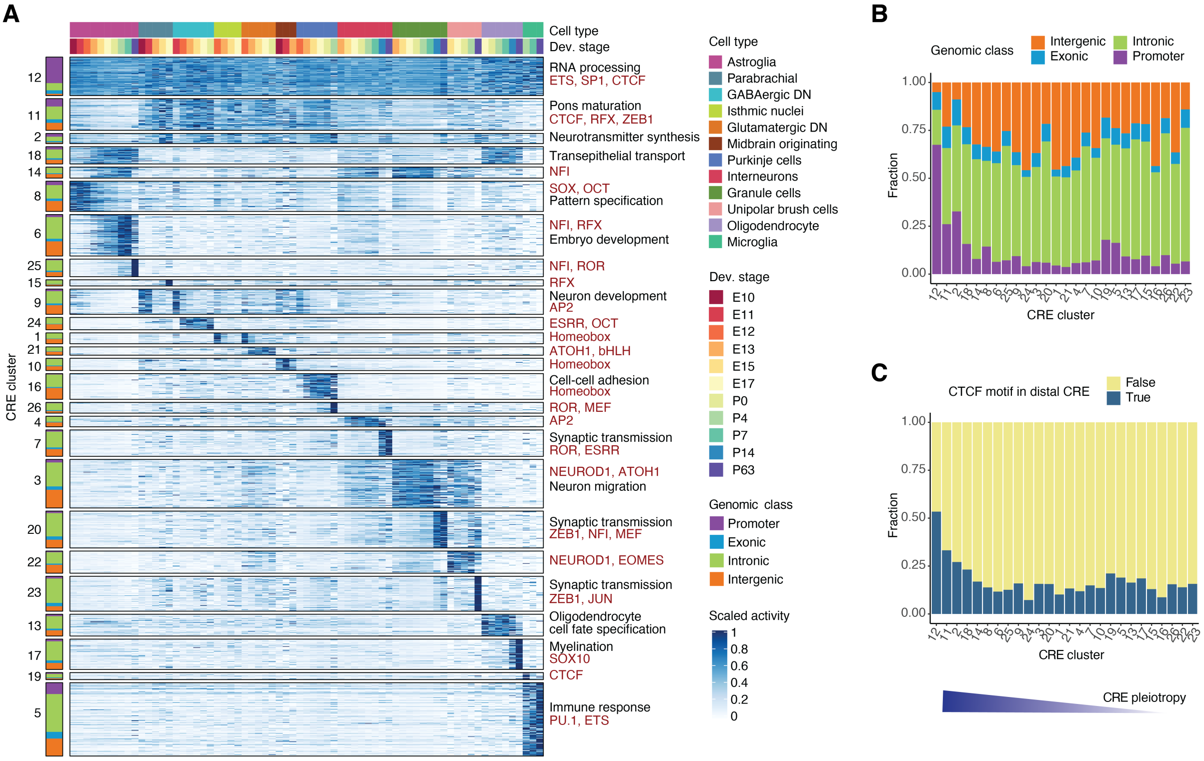


Figure 3.25: Spatiotemporal patterns of CRE activity in cerebellum development. **(A)** Clusters of CRE activity across cell types and developmental stages. CREs are grouped by activity cluster (k-means followed by hierarchical clustering) and genomic class (left). CRE clusters are arranged in decreasing order of pleiotropy (here: mean activity across rows) and then by cell type and developmental stage with maximum activity. Right: Representative enrichments (BH adjusted $P < 0.05$; hypergeometric test) for biological processes of adjacent genes (black) and motifs for TFs or TF families (red). 50,000 CREs confidently assigned to their cluster were chosen randomly for visualization. **(B)** Fraction of genomic classes across clusters of CREs. **(C)** Fraction of distal CREs overlapping at least one CTCF motif across clusters of CREs. *Figure adapted from Sarropoulos et al. 2021* (*244*).

The analysis described above suggests that both cell type and developmental stage contribute to the global chromatin accessibility patterns **(Figure 3.25A)**. The developmental effect on chromatin accessibility is also supported by the UMAP projection of the entire dataset **(Figure 3.20B-C)**. Cells from the same cell type and developmental stage consistently group together and apart from cells assigned to different cell types. However, cells from the same cell type but different developmental stages can also show additional separation in this embedding, which is especially prominent for specific developmental periods (e.g., E13-P0 granule cell progenitors are highly similar to each other but separated from P4-P7 granule cell progenitors).

To better understand these temporal differences, I performed a differential accessibility analysis between pre- and postnatal granule cell progenitors, identifying a total of 3,988 CREs with increasing or decreasing accessibility around birth **(Figure 3.26)**. Notably, this period also coincides with greater changes at the transcriptional level as observed for protein-coding genes (*172*) and lncRNAs **(Figure 3.7A)**. It is also concurrent with the Shh-signaling-mediated expansion of the granule cell progenitor pool around birth (*130*). CREs with increasing accessibility after birth are enriched for motifs of the NFI transcription factors **(Figure 3.26A)**, which are known transcriptional regulators of late-born neural cell types (*255*), as well as GLI2, one of the major TFs activated upon Shh-signaling (*130*). By contrast, CREs with decreasing accessibility are enriched for motifs marking embryonic progenitor cells, such as those of the SOX and MEIS TF families **(Figure 3.26A)**. Thus, the developmental differences I observed between corresponding cell types are likely to be of biological significance.

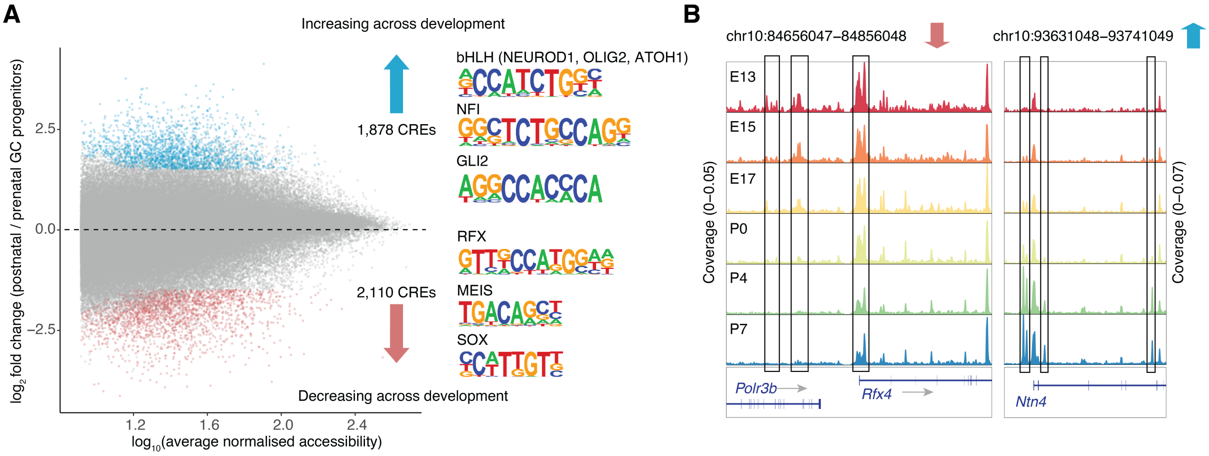


Figure 3.26: Differential accessibility between pre- and postnatal granule cell progenitors. (A) MA plot of differentially accessible CREs between pre- and postnatal granule cell progenitors. CREs with significantly increased (blue) and decreased (red) accessibility in postnatal compared to prenatal granule cell progenitors were identified with an absolute $\log_2$ fold-change of at least 1.5 and a BH adjusted *P*-value < 0.05. TF motifs with highest enrichment based on hypergeometric tests are shown for each group. **(B)** Examples of aggregated accessibility profiles (scaled by the total number of fragments in each group) across granule cell progenitors of different developmental stages for CREs with decreasing (left) and increasing (right) accessibility during development. *Figure adapted from Sarropoulos et al. 2021* (*244*).

3.3.4 Chromatin accessibility dynamics of cerebellar progenitors

The highly dynamic cellular composition of the developing cerebellum is largely due to the generation of distinct cell types from cerebellar progenitors in a spatially and temporally restricted manner (see section 1.6.1). For example, ventricular zone progenitors give rise to GABAergic deep nuclei neurons at E10-E11, shift to Purkinje cells around E12 and eventually to GABAergic interneurons from E15 onwards (*130*). In this section, I sought to investigate whether these shifts in cell fate specification were associated with spatiotemporal heterogeneity in the chromatin accessibility profiles of cerebellar progenitor cells.

3.3.4.1 Identification of known and novel subtypes of cerebellar progenitors

To identify subtypes of cerebellar progenitors, I subclustered cells from the astroglial lineage (i.e., progenitor cells and mature astrocytes) and used gene scores to associate clusters with distinct progenitor cell populations. Through this process, I was able to identify all major germinal zones in the developing cerebellum. However, differences in CRE accessibility between progenitor types are overall subtle and without sharp boundaries within a given stage **(Figure 3.27A-D)**.

Early (E10-E12) progenitor populations consist of previously described isthmic (*En1*, *Pax5*), ventricular zone (*Dll1*, *Ptf1α*) and rhombic lip (*Cdon*, *Atoh1*) progenitors **(Figure 3.27D)**. Additionally, I identified a set of progenitor cells that show no apparent commitment to a specific cell fate (herein: "Uncommitted"), which could correspond to a recently described population of *Sox2*$^+$ progenitors that can generate both excitatory and inhibitory neuron types (*256*). In addition, I also detected some early *Gsx1*$^+$ cells, which Dr. Mari Sepp was able to trace to the anterior ventricular zone (i.e., the border between the ventricular zone and the isthmus). We thus termed these cells "anterior ventricular zone progenitors" **(Figure 3.27A-D)**.

Gsx1 is a known marker of a progenitor population described as "bipotent" because of its ability to give rise to both GABAergic interneurons and parenchymal astrocytes (*207, 257*). This population is thought to emerge through a temporal shift of *Olig2*-expressing ventricular zone progenitors, which give rise to Purkinje cells, towards *Gsx1*-expressing cells around E13 (*130, 257*). Indeed, starting from E13 onwards, I was able to identify these bipotent progenitors, at the expense of the ventricular zone progenitors from earlier stages **(Figure 3.27D)**. Since *Gsx1* is a shared marker between bipotent progenitors and the anterior ventricular zone progenitors found in earlier stages, I sought to investigate whether the latter could represent a population that is already primed for the bipotent fate. Indeed, the two progenitor populations share additional markers, such as *Ndnf*, *Robo1* and *Wnt8b* **(Figure 3.27E)**. Furthermore, amongst early (E10-E12) groups, the anterior ventricular zone population shows the highest similarity in chromatin accessibility to E13-E15 bipotent progenitors **(Figure 3.27F)**. Thus, besides the previously described temporal transition of *Olig2*$^+$ Purkinje-generating progenitors (*130, 257*), my analyses identified an additional, molecularly distinct population in the early anterior ventricular zone, which is already primed to acquire the bipotent progenitor identity.

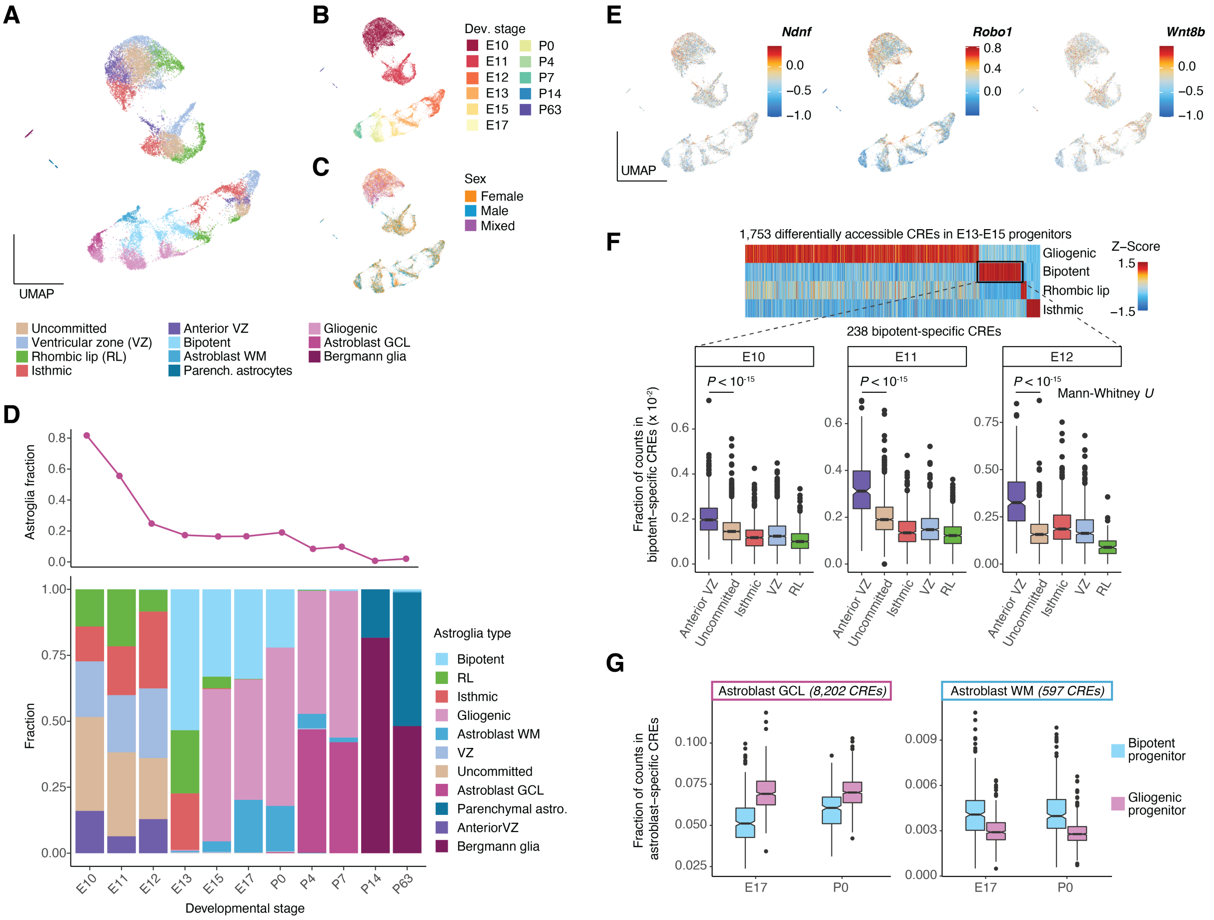


Figure 3.27: Characterization of cerebellar progenitor subtypes. **(A, B, C)** UMAP projections of 21,830 astroglia cells (progenitors and astrocytes) colored by subtype **(A)**, developmental stage **(B)** and sex **(C)**. **(D)** Relative abundance of astroglia types (bottom) and overall fraction in the cerebellum (top) across developmental stages. **(E)** Gene scores for marker genes that are shared between anterior ventricular zone and bipotent progenitors (counts per 10,000 fragments, capped at 10th and 99th quantiles and $\log_{10}$ transformed). **(F)** Comparison of bipotent progenitors to earlier populations. Top: Activity profiles (Z-score) of progenitor type-specific CREs in E13-E15. Bottom: Fraction of fragments per cell in CREs specific to bipotent progenitors across progenitor types and developmental stages. **(G)** Comparison of granule cell layer and white matter astroblasts to E17 and P0 bipotent and gliogenic progenitors. Boxplots show the fraction of fragments per cell in CREs specific to astroblast populations across progenitor types and developmental stages. *Figure adapted from Sarropoulos et al. 2021* (*244*).

E15 is marked by the appearance of another population, gliogenic progenitors (*Slc1a3*, *Grm3*). Bipotent and gliogenic progenitors are thought to give rise to two distinct parenchymal astrocyte populations, located in the white matter and granule cell layer, respectively (*207, 208*). In line with this, in perinatal stages (E17-P7) I identified two groups of astroblasts, which Dr. Mari Sepp traced to the white matter (*Slc6a11*, *Olig2*, *Kcnd2*) and the granule cell layer (*Aqp4*, *Tekt5*). To assess the potential lineage relationships of these two astroblast groups based on molecular similarity to the earlier progenitor populations, I identified differentially accessible CREs between them and then quantified their accessibility in progenitor cells from earlier stages. CREs specific to astroblasts of the white matter show higher accessibility in bipotent progenitors, whereas gliogenic progenitors show higher activity in CREs specific to astroblasts of the granule cell layer (**Figure 3.27G**). Thus, the progenitor and astrocyte groups identified here match previous descriptions and ontogenetic relationships.

3.3.4.1 Temporal changes in CRE accessibility are shared between progenitor groups

The subclustering analysis of the astroglial lineage **(Figure 3.27A-C)** revealed an unexpected result: progenitor cells primarily cluster by developmental stage rather than progenitor type, an observation most prominent in early development (E10-E12). This is despite the dense temporal sampling of our dataset, and even though I was able to identify the same progenitor types across consecutive developmental stages, often based on the same marker genes. This result is unlikely to be attributed to batch effects, as biological replicates show very high similarity **(Figure 3.27B)** and the separation in the UMAP recapitulates the progression of developmental time **(Figure 3.27C)**. Additionally, cells from adjacent developmental stages around and after birth (e.g., E17-P0, P4-P7) also cluster together in this embedding **(Figure 3.27C)**. Finally, I was able to observe the same clustering pattern when performing hierarchical clustering based on aggregated chromatin profiles across progenitor groups of the same cell type and developmental stage **(Figure 3.28A)**. Thus, this clustering pattern of cerebellar progenitor cells is likely explained by major temporal changes in CRE accessibility, which are more prominent during early development, a period coinciding with the sequential generation of distinct neuronal cell types from the same germinal zone (see section 1.6.1).

To identify the CREs that drive this clustering pattern, I performed the two-step clustering procedure described in section 3.3.3 to group CREs into 12 clusters based on their activity across progenitor types and developmental stages **(Figure 3.28B)**. Most CRE clusters can be classified as time-variant (i.e., their accessibility is primarily determined by developmental stage rather than by germinal zone). Thus, the clustering of progenitor cells by developmental stage **(Figure 3.27A-C, Figure 3.28A)** is primarily explained by large differences in CRE activity during development, which are shared between germinal zones. This suggests that concordant changes in cell fate specification (e.g., ventricular zone progenitors transitioning from Purkinje cells to GABAergic interneurons around E13 while rhombic lip progenitors shift from glutamatergic DN to granule cell progenitors) could be attributed to the same temporal cues. Such a major transition involves the shift from early CREs (E10-E12), enriched for nuclear receptors and SOX motifs and associated with chromatin silencing genes, towards CREs enriched for NFI motifs and adjacent to genes involved in signaling and cell adhesion **(Figure 3.28B)**. Additional groups of progenitor type-variant CREs are enriched for the motifs of known transcriptional regulators (e.g., ATOH1 for rhombic lip, PTF1A for ventricular zone; **Figure 3.28B**).

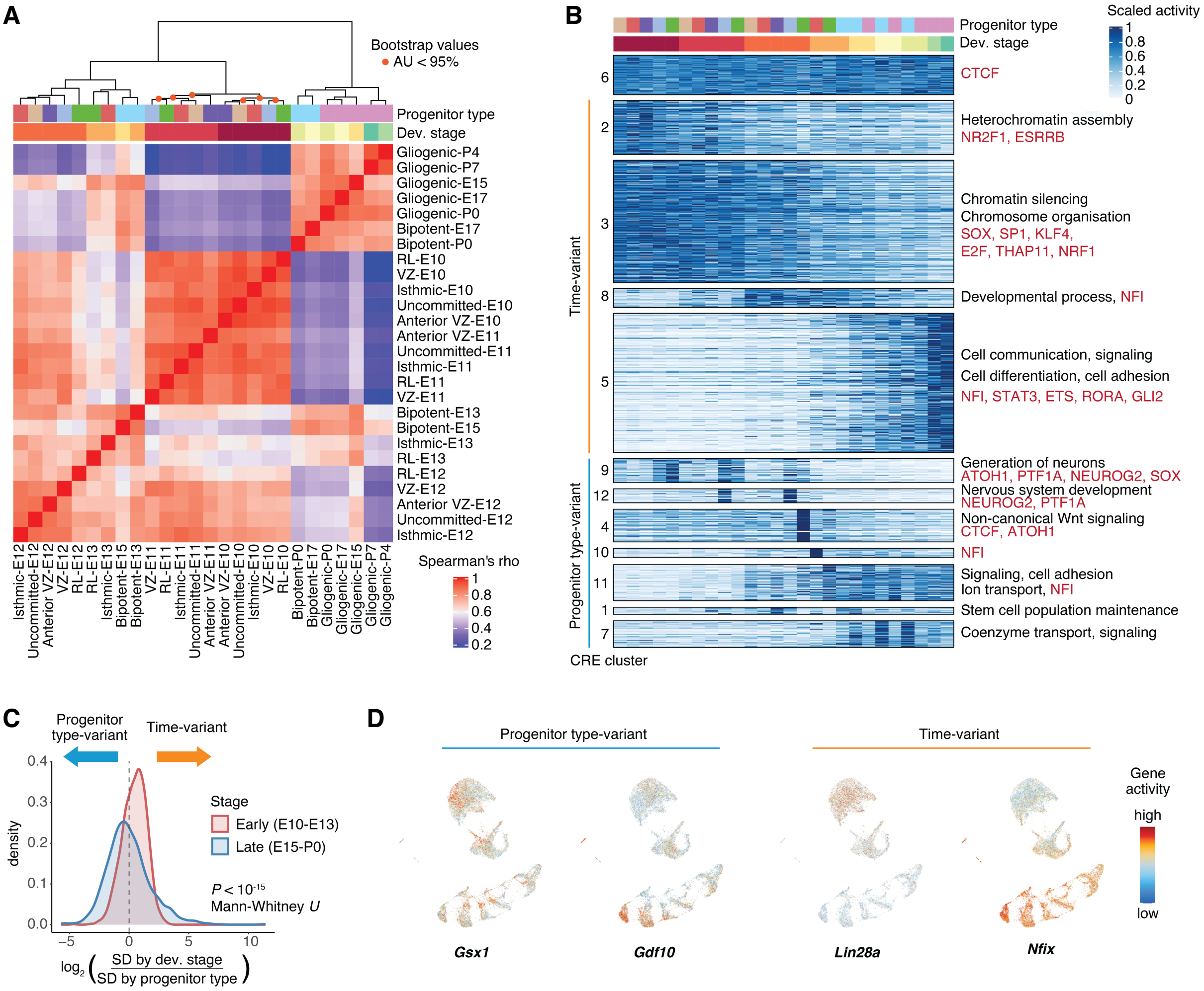
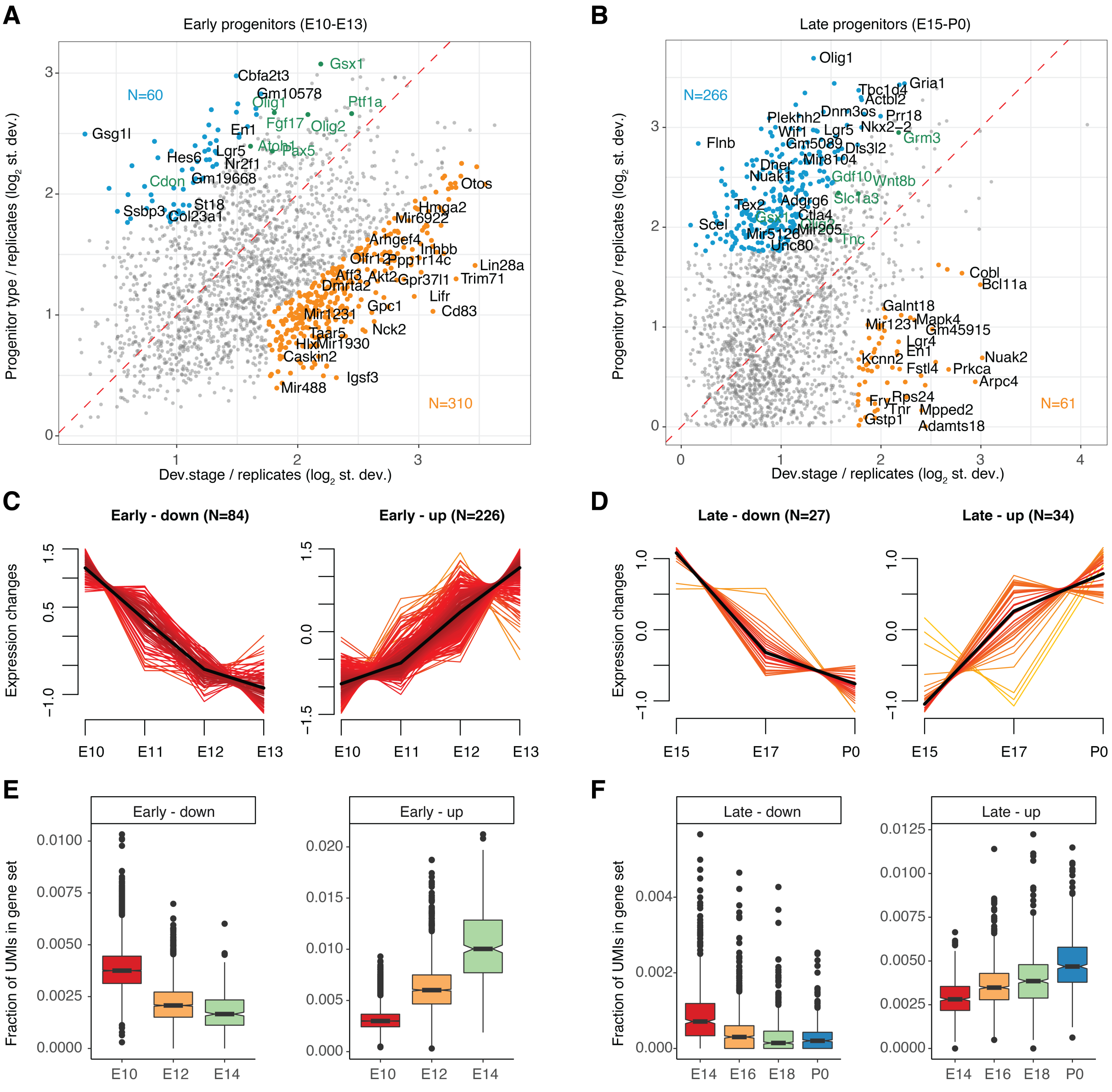


Figure 3.28: Spatiotemporal heterogeneity in cerebellar progenitor populations. **(A)** Hierarchical clustering based on Spearman's correlation coefficients in CRE accessibility across progenitor types and developmental stages. Orange dots indicate nodes with approximately unbiased (AU) probability values < 95%. **(B)** Clusters of CRE activity across progenitor types and developmental stages. CREs are grouped by activity cluster (k-means followed by hierarchical clustering). Right: Representative enrichments (BH adjusted $P < 0.05$; hypergeometric test) for biological processes of adjacent genes (black) and motifs for TFs or TF families (red). 25,000 CREs confidently assigned to their cluster were chosen randomly for visualization. **(C)** Density distributions for the $\log_2$-ratio of gene score standard deviation (SD) across developmental stages and progenitor types for the 2,000 genes with the highest variance in early and late progenitor populations. **(D)** Gene scores (capped at 10th and 99th quantiles and $\log_{10}$ transformed) for genes with high variance across progenitor types (left) or developmental stages (right). *Figure adapted from Sarropoulos et al. 2021* (*244*).

I next sought to assess whether these temporal changes in CRE accessibility also lead to concordant differences in gene expression. To this end, I first estimated the variance in gene scores (as a proxy for gene expression) for each gene across progenitor types and developmental stages (**Figure 3.28C, Figure 3.29A-B**). In agreement with my observations regarding CRE accessibility, the majority of the 2,000 most highly variable genes in early developmental stages (E10-E13) show higher variance across developmental stages than between germinal zones (**Figure 3.28C**). Such temporally variant genes include the pluripotency factor *Lin28a*, which is expressed in all E10 progenitor cells and shows a gradually decreasing activity during later stages (**Figure 3.28D**). By contrast, the gene score of the TF *Nfix* gradually increases across all progenitor types from E12 on (**Figure 3.28D**). Notably, as discussed above, NFI motifs are enriched amongst CREs with a shared increase in accessibility across progenitor types (**Figure 3.28B**; clusters 5 and 8). On the other hand, variance in gene scores by progenitor type is higher in late progenitor populations (E15-P0), further corroborating my observation that temporal differences are strongest in early cerebellar development (**Figure 3.28C, Figure 3.29B**). Genes with high variance across progenitor types include known marker genes for specific progenitor types, such as *Gsx1* for bipotent progenitors and *Gdf10* for gliogenic progenitors (**Figure 3.28D**).

To directly examine the expression patterns of these temporally-variant genes, which I identified based on their gene scores (i.e., gene expression inferred by chromatin accessibility), I utilized a previously published scRNA-seq dataset (*210*). I used soft clustering to identify genes with increasing or decreasing gene score activity during development (**Figure 3.29C-D**) and then estimated the aggregated expression of each group across progenitor cells from different developmental stages in the scRNA-seq data. Temporally variant genes show significant differences in their expression across developmental stages, matching the direction predicted by their gene scores (**Figure 3.29E-F**). Thus, the strong temporal differences I observed in the chromatin accessibility profiles of cerebellar progenitors, which are most prominent during early development and shared between germinal zones, also lead to developmental changes in gene expression.



scRNA-seq data from *Vladoiu et al, 2019*

scRNA-seq data from *Vladoiu et al, 2019*

Figure 3.29: Identification of temporally-variant genes in cerebellar progenitors. **(A, B)** Log$_2$ ratio of standard deviations across developmental stages (x-axis) and progenitor types (y-axis) over mean standard deviation between replicates, for the 2,000 genes with highest gene score variance for early (E10-E13; **A**) and late (E15-P0; **B**) progenitor populations. Temporally-variant (orange) genes and germinal zone-variant (blue) genes were identified with a log$_2$ ratio of at least 1.25 compared to both other standard deviations. Marker genes used for the identification of progenitor types are shown in green. **(C, D)** Z-score scaled gene score for temporally-variant genes (orange in **A, B**) for early (E10-E13; **C**) and late (E15-P0; **D**) progenitor populations with decreasing (left) and increasing (right) activity, as determined by fuzzy clustering. The number of genes in each cluster is shown on top. **(E, F)** Fraction of UMIs per progenitor cell in temporally-variant gene clusters from **C, D.** Data and annotations are from *Vladoiu et al. 2019* (*210*). Different y-ranges were used across gene sets to facilitate temporal comparisons within each group as the overall fraction of UMIs depends on the number of genes in each set. *Figure adapted from Sarropoulos et al. 2021* (*244*).

3.3.5 CRE activity during neuronal differentiation

Following cell fate specification, immature neurons need to acquire their final location, form and function. This often involves their migration to a different position in the brain, as well as a series of morphological changes, such as the growth of an axon and dendrites, and eventually the formation of synapses (*130*). Collectively, these processes are referred to as neuronal differentiation and maturation (*130*). To study the dynamics of chromatin accessibility during this process, I focused on the three most abundant neuron types in our dataset, granule cells, Purkinje cells and GABAergic interneurons. Since cells from different developmental stages show differences in CRE accessibility, which are independent of differentiation (**Figure 3.30A-C**), I used a batch-correction method (*258*) to explicitly remove developmental signal and align these cells along their differentiation trajectories, which I modelled using diffusion pseudotime (*259*) (**Figure 3.30A-C**).

Different neuron types vary in their differentiation dynamics and the shape of their trajectories. For Punkinje cells, most of the change in their chromatin accessibility profiles occurs during E12 and E13 (**Figure 3.30B**), in line with their stage-restricted generation (*130*). By contrast, granule cells are characterized by protracted differentiation with cells from multiple stages (E15-P14) showing a large spread in their pseudotime distribution (**Figure 3.30A**). GABAergic interneurons also show protracted differentiation (E13-P7) but in contrast to the largely homogeneous granule cell population, they are stratified into distinct temporally-specified subtypes (**Figure 3.30C-E**). These include early-born interneurons (*Zfhx4, Slit2*) which can be detected from E13 to E15, mid-born Golgi cells (*Chrm2*) that can be found in small numbers throughout development (E13-P63), and Purkinje cell layer interneurons (*Nxph1, Klhl1*) which are most abundant between E17 and P7 (**Figure 3.30D-E**). Late postnatal stages (P14, P63) are dominated by molecular layer interneurons of type 1 (*Sorcs3, Grm8*) and 2 (*Nxph1, Pvalb*; **Figure 3.30C-E**).

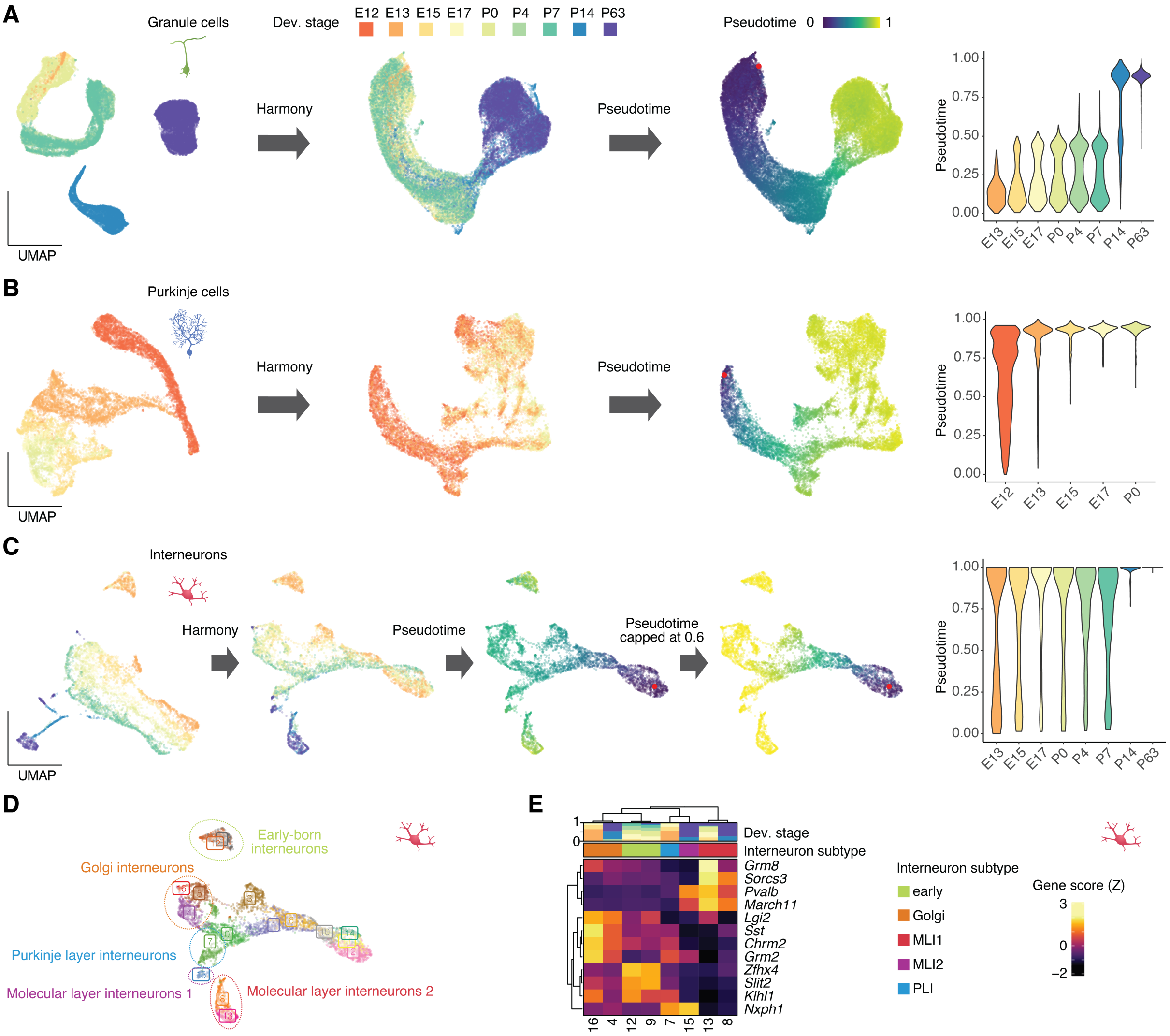


Figure 3.30: Differentiation dynamics of cerebellar neuron types. **(A, B, C)** UMAP projections and distribution of pseudotime values across developmental stages for 35,153 granule cells **(A)**, 13,214 Purkinje cells **(B)** and 5,113 interneurons **(C)** before (left) and after (middle and right) Harmony-alignment across developmental stages. Cells are colored by developmental stage (left and middle) and pseudotime value (right). For interneurons **(C)**, pseudotime was capped at 0.6 and rescaled to eliminate differences between temporally specified subtypes (see **D, E**). **(D)** UMAP projection of 5,113 Harmony-aligned interneurons colored by cluster. Temporally specified interneuron subtypes are shown in circles. **(E)** Gene score activity (Z-score) of marker genes for mature interneuron clusters (as in **D**). Subtype annotation and relative contribution of developmental stages per cluster are shown above the heatmap. *Figure adapted from Sarropoulos et al. 2021* (244).

Despite these differences in the dynamics of differentiation, I observed considerable similarity in the molecular processes occurring in matched differentiation states between neuron types (**Figure 3.31A**). CREs with high accessibility in early differentiation are enriched near genes associated with pattern specification, and are gradually replaced by CREs proximal to transcriptional and developmental regulators (**Figure 3.31A**). In later stages, chromatin accessibility increases near genes involved in axon guidance, migration and synapse assembly, whereas CREs active in mature neurons are close to genes associated with neurotransmitter secretion (**Figure 3.31A)**. Besides highlighting the similarity in the differentiation processes between different cerebellar neuron types, this analysis also supports the utility of pseudotime to model differentiation trajectories, as these processes recapitulate the well-established events leading to the formation of mature cerebellar neurons (*130*).

In addition to the convergence in biological processes, I observed that different neuron types also share a large number of dynamic genes (43% of protein-coding genes with dynamic activity across pseudotime are also dynamic in at least another neuron type; **Figure 3.31B**), supporting the existence of a core gene expression program central to neuronal differentiation (*260*). By contrast, only 20% of dynamic CREs are shared (i.e., pleiotropic) across neuron types (**Figure 3.31B)**, suggesting that the same target genes are often activated by distinct CREs in different neuron types, in line with the high context-specificity of most CREs (*36*, *229*). However, pleiotropic CREs are often active across matched stages of differentiation in different cell types (**Figure 3.31C-D**; early-early: $P < 10^{-15}$, late-late: $P < 10^{-15}$, early-late: $P > 0.99$, hypergeometric test). Additionally, with the exception of interneurons, pleiotropic CREs are enriched amongst clusters active in early stages of differentiation (**Figure 3.31E)**, suggesting a gradual divergence in the chromatin accessibility profiles of different neuron types during differentiation. This can also be seen in a principal component analysis (PCA), in which CRE accessibility profiles of different neuron types show higher similarity in early differentiation states (**Figure 3.31F)**.

Collectively, in this section I revealed similarities and differences in the differentiation processes of the three major cerebellar neuron types. I also identified a subset of CREs that are shared across multiple cell types and which are more prevalent during early differentiation. More generally, in this part of my thesis work I explored the chromatin accessibility dynamics of cell types during the development of the mouse cerebellum. These analyses set the stage for the evolutionary comparisons that follow in the next and final part of my dissertation.

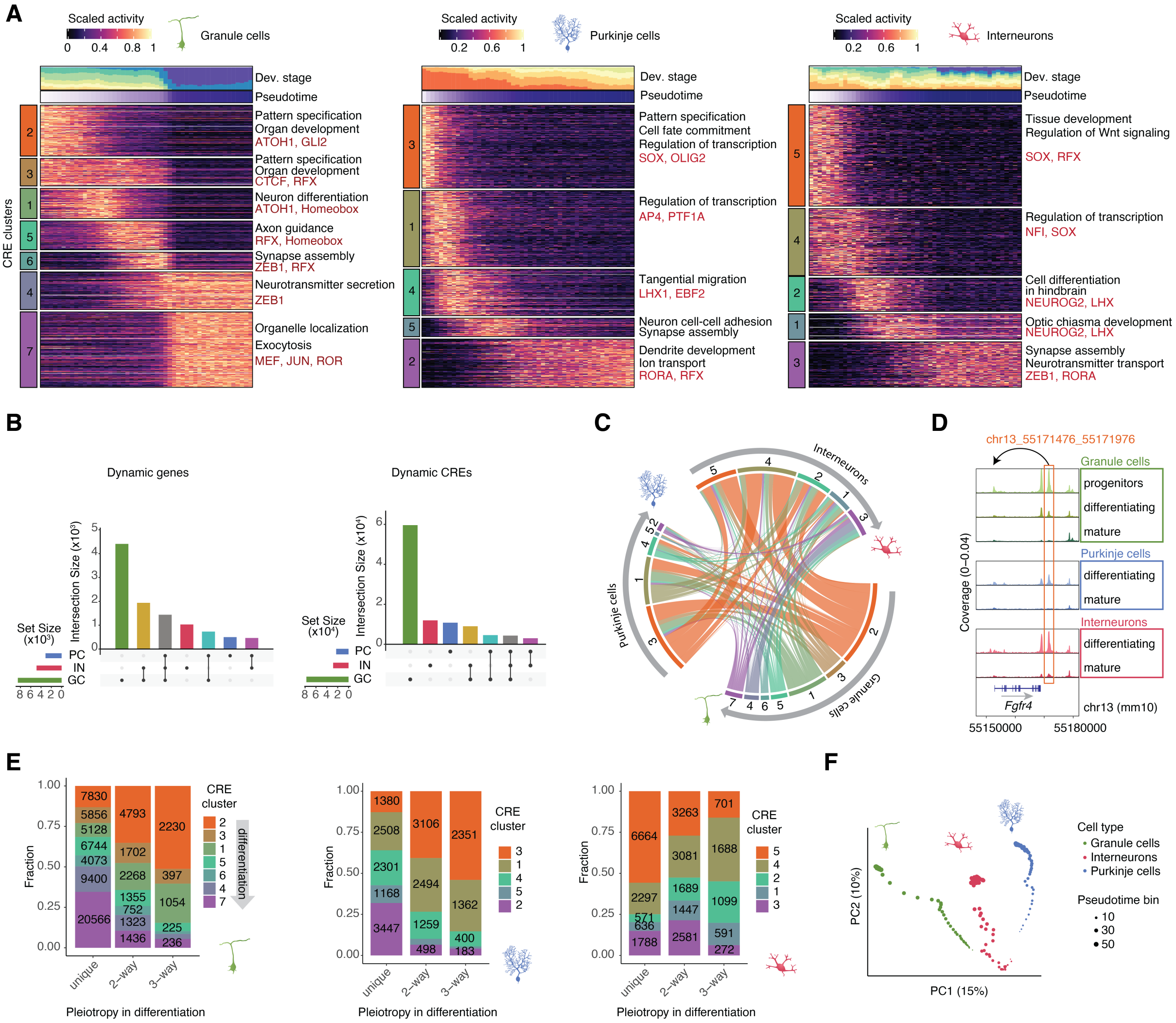


Figure 3.31: CRE activity and pleiotropy during neuronal differentiation. **(A)** Z-score scaled activity of dynamic CREs during granule cell (left), Purkinje cell (middle) and GABAergic interneuron (right) differentiation, averaged across 50 bins of increasing pseudotime ranks. Top: Contribution of developmental stages and mean pseudotime value for each bin. Right: Representative enrichments (BH adjusted $P < 0.05$; hypergeometric test) for biological processes of adjacent genes (black) and TF motifs (red). **(B)** Upset plots of intersections between genes (left) and CREs (right) with dynamic activity in differentiating granule cells, Purkinje cells and interneurons. Connected dots mark overlapping sets. Horizontal bars show the total number of dynamic genes (left) and CREs (right) per cell type. **(C)** Overlap between activity clusters for CREs dynamic in two or more neurons (pleiotropic). For each neuron type (outer sector) CRE clusters (as in **A**) are ordered from early (orange) to late (violet) activity during differentiation. Each node connects the activity clusters of two different neuron types for the same CRE. **(D)** Example of a pleiotropic intergenic CRE, assigned to *Fgfr4*. Accessibility profiles for each cell type and state were aggregated across cells from all developmental stages and scaled by the total number of fragments in each group. **(E)** Fraction of CRE clusters (as in **A**) across CREs dynamic in a single neuron type (unique) or shared across two or three cell types, for granule cells (left), Purkinje cells (middle) and GABAergic interneurons (right). **(F)** Principal component analysis of CRE accessibility during granule cell, Purkinje cell and interneuron differentiation. Percentage values show the proportion of variance explained by each component. *Figure adapted from Sarropoulos et al. 2021* (244).

3.4 The evolution of gene regulation in the mammalian cerebellum

This part of my thesis research focused on the evolution of CREs that are active in cell types of the developing mammalian cerebellum, starting from the identification of conserved CRE sequences and gradually transitioning towards the search for CREs associated with evolutionary innovation. The first two sections are based on the same mouse snATAC-seq dataset described in 3.3. The final four sections incorporate additional snATAC-seq datasets of the developing cerebellum in opossum (3.4.3) and human (3.4.4-6) produced by Dr. Mari Sepp with technical support from Julia Schmidt and Celine Schneider, as well as additional publicly available data. The findings presented in the first three sections were published in *Sarropoulos, Sepp et al. 2021* (*244*), whereas the final three sections describe unpublished work. The development of the deep learning models discussed in the last two sections was carried out in collaboration with Prof. Dr. Stein Aerts at the University of Leuven and his lab members, Dr. Nikolai Hecker, Ibrahim Taskiran and Carmen Bravo González-Blas. Unless stated otherwise below, all computational analyses are the product of my own work, under the supervision of Prof. Dr. Henrik Kaessmann, Prof. Dr. Stein Aerts, Prof. Dr. Stefan Pfister and Dr. Margarida Cardoso-Moreira.

3.4.1 Evolutionary dynamics of CRE sequences in cerebellum development

First, I focused on the evolution of CRE sequences and its relation to their spatiotemporal activity. Previous studies have reported a decrease in the conservation of protein-coding gene expression (*172*, *187*, *188*) and of enhancer sequences (*55*) during mammalian organ development, a pattern that I showed to also extend to lncRNA expression (**Figure 3.8B-C**). However, since all aforementioned studies examined whole organs, it remained unclear whether these patterns are driven by the higher evolutionary conservation of CREs active in cell types that are highly abundant in early stages of organ development (e.g., progenitors, deep nuclei neurons) or whether there are temporal differences in the conservation of CREs within cell types.

To assess the relative contributions of these two scenarios, I estimated the sequence constraint (phastCons scores) and the minimum age (i.e., when a CRE first appeared during evolution) based on syntenic sequence alignments between mouse and 16 other vertebrates at various phylogenetic distances (**Figure 3.32A**). To avoid biases associated with the higher evolutionary conservation of CREs overlapping or proximal to protein-coding sequences (**Figure 3.32B**), I focused these analyses on intergenic CREs. For each single cell, I estimated a score based on the mean conservation estimates of all intergenic CREs that were accessible in that cell. I then summarized these metrics across cell types and developmental stages (**Figure 3.32C-E**).

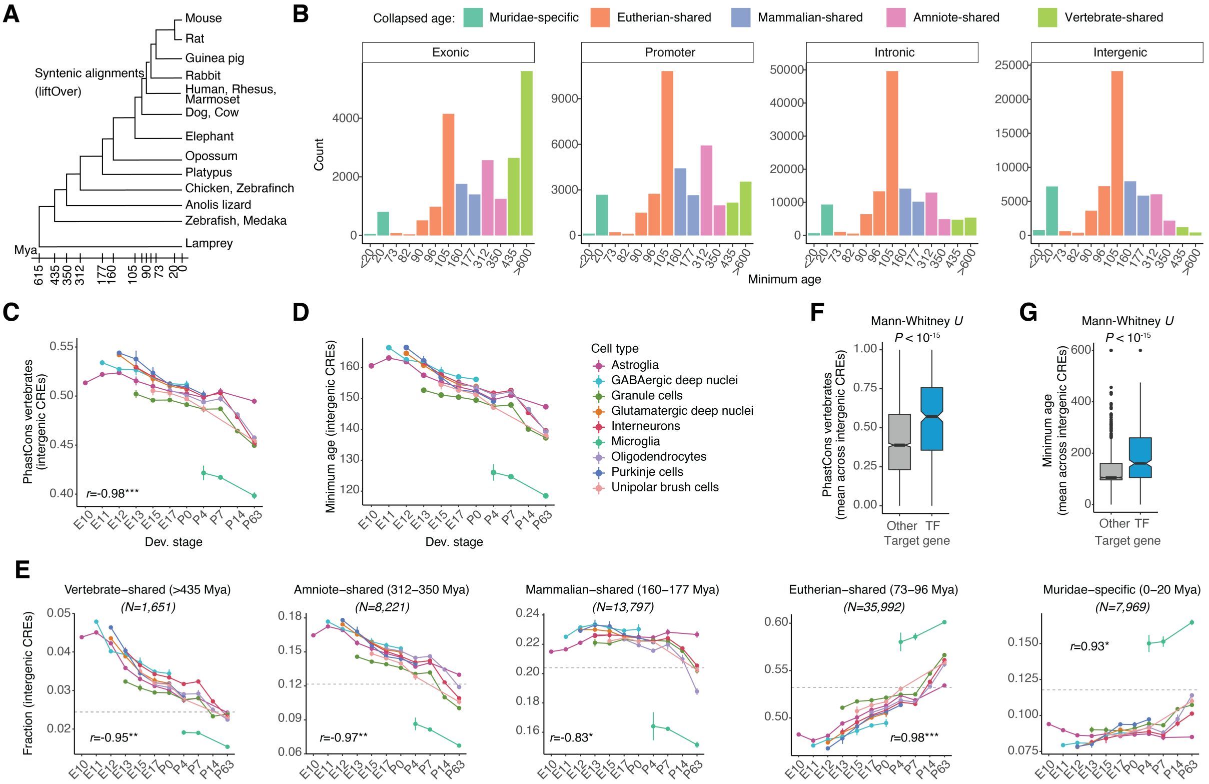


Figure 3.32: Evolutionary dynamics of CREs in developing cerebellar cell types. (A) Species used in the syntenic alignments to infer the minimum age of CREs based on the date of divergence (from *timetree.org*) between mouse and the most distant species in which an alignment was detected. **(B)** Number of CREs across genomic classes and age groups. Colors indicate broader age groups as used in **E**. **(C, D)** Sequence constraint **(C)** and minimum age **(D)** of intergenic CREs accessible per cell, averaged for each cell type and developmental stage. **(E)** Fraction of accessible intergenic CREs assigned to different age groups per cell, averaged for each cell type and developmental stage. Different y-ranges were used across age groups to facilitate comparisons between cell types and stages within each group, as the fraction depends on the number of CREs per group (indicated on top). The fraction of each age group across all intergenic CREs is marked by the dotted horizontal line. For **C-E**, Pearson's *r* correlation coefficients between the estimates and development are shown (median across cell types; *P*<0.05*, *P*<0.01**, *P*<0.001***); vertical bars illustrate difference in average estimates between biological replicates. Only groups with at least 50 cells were considered. **(F, G)** Average sequence constraint **(F)** and minimum age **(G)** of intergenic CREs per target gene for TFs and other genes. *Figure adapted from Sarropoulos et al. 2021* (244).

I observed a significant decrease in CRE sequence constraint during development across all cell types (**Figure 3.32C-D**). Accordingly, the fraction of ancient CREs (older than 300 million years) decreases gradually during development, whereas the fraction of CREs that are younger than 100 million years is significantly higher towards adulthood (**Figure 3.32E**). This suggests that the gene regulatory programs involved in the specification of cell type identity are significantly more conserved than those regulating the functions of the mature cell types. In line with this, CREs associated with TF genes, which are central to defining cell type identity, are older and under more constraint than CREs associated with other genes (**Figure 3.32F-G**). In contrast to these strong temporal patterns, differences in CRE constraint between cell types within the same developmental stage are limited, with the exception of late postnatal stages (see section 3.4.2). Thus, previous observations at the level of whole-organs are largely explained by the decrease in CRE conservation within cell types rather than by changes in the relative abundance of cell types with pronounced differences in evolutionary constraints.

However, even within a cell type, multiple processes might explain the temporal differences in CRE activity and conservation. As discussed in section 3.3.5, even cells from the same cell type and differentiation state show additional differences in chromatin accessibility across developmental stages. These might be related to intrinsic temporal patterning signals (*255, 261*) or to extrinsic factors such as the availability of morphogens and ligands (*140*). To assess the relative contributions of cell type differentiation versus these additional temporal signals, I focused on granule cells, which have a protracted differentiation trajectory (E13-P14), show differences in accessibility between developmental stages (**Figure 3.26, Figure 3.30A**), but are not known to become stratified into distinct temporally specified subtypes (*130*).

I compared the conservation of intergenic CREs across both differentiation (modelled through pseudotime, see section 3.3.5) and developmental stages. Both factors affect CRE constraint, with cells from the earliest developmental stage (E13) and differentiation state (granule cell progenitors) having the most conserved regulatory programs (**Figure 3.33A-C**). To validate this observation in a pseudotime-independent framework, I focused on prenatal granule cell progenitors (E13-P0), which cluster together without any correction across developmental stages (**Figure 3.33D**). Thus, these cells are overall very similar in their chromatin accessibility profiles. However, despite this similarity, I was able to identify 7,527 CREs with decreasing and 11,972 CREs with increasing accessibility profiles during development (**Figure 3.33D**). The sequences of CREs with decreasing accessibility are more constrained than those increasing during development (**Figure 3.33E**), further supporting that the regulatory programs of early granule cell progenitors are more conserved than those of late granule cell progenitors. In support of the biological significance of these differences, CREs with decreasing accessibility are enriched for SOX and RFX motifs, which are shared with other early cell types, whereas those with increasing accessibility are recognized by factors specific to rhombic lip-derived cells, such as ATOH1 and GLI2, or by TFs associated with multiple late-born cell types, such as NFI factors (**Figure 3.33F**). Furthermore, using a published scRNA-seq dataset, I showed that genes adjacent to these temporally dynamic CREs also show concordant temporal differences in their expression in granule cell progenitors (**Figure 3.33G**).

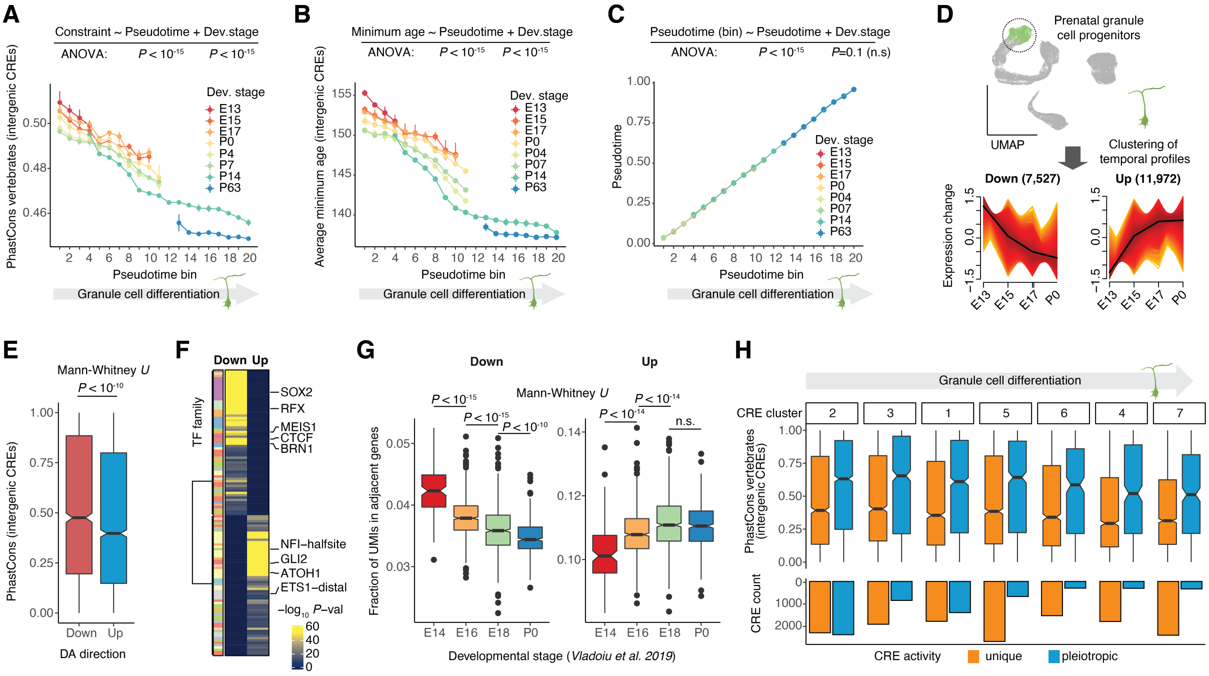


Figure 3.33: Effects of development, differentiation and pleiotropy on CRE conservation. (**A, B, C**) Sequence constraint (**A**) and minimum age (**B**) of intergenic CREs, and pseudotime (**C**) per cell, averaged for each developmental stage (color) and pseudotime interval (step=0.05) of granule cell differentiation. Vertical bars illustrate the difference in average estimates between biological replicates. (**D**) Identification of temporal differences in CRE activity in prenatal granule cell progenitors (top; UMAP of 35,153 granule cells prior to alignment across developmental stages, as in **Figure 3.30A**). Z-score scaled temporal activity of CREs with decreasing or increasing accessibility across development (bottom). Black lines indicate mean values for each cluster. (**E, F**) Sequence constraint (**E**) and TF motifs enrichment (**F**; BH adjusted P < 0.05; hypergeometric test) for intergenic CREs with temporal differences in prenatal granule cell progenitors (clusters from **D**). (**G**) Fraction of UMIs per cell in putative target genes of CREs with decreasing (left) or increasing (right) accessibility during development in prenatal granule cell progenitors. Data and annotations are from *Vladoiu et al. 2019* (*210*). Different y-ranges were used across gene sets to facilitate temporal comparisons within each group as the overall fraction of UMIs depends on the number of genes in each set. (**H**) Sequence constraint (top) and abundance (bottom) of cell type-specific and pleiotropic intergenic CREs active in different stages of granule cell differentiation (from **Figure 3.31A**, ordered from early to late activity during differentiation). *Figure adapted from Sarropoulos et al. 2021* (*244*).

Next, I decided to test whether the decrease in CRE conservation during cell type differentiation (**Figure 3.33A-B**) is associated with the parallel decrease in CRE pleiotropy (**Figure 3.31E-F**), as the latter is thought to impose evolutionary constraints (*172*, *254*). Re-examining the granule cell differentiation trajectory, I found that pleiotropic CREs (i.e., those dynamic in at least two neuron types) are more constrained than those dynamic in a single neuron type (**Figure 3.33H**). These differences are pervasive across all differentiation states and are larger than those observed between CREs of similar pleiotropy but active in different differentiation states. Thus, the gradual decrease in CRE constraint during neuronal differentiation is mostly explained by the decrease in the fraction of pleiotropic CREs, which are more conserved than those active in a single cell type.

3.4.2 Differences in CRE sequence constraint across cerebellar cell types

Although developmental differences in the constraint of intergenic CREs can be observed within all cell types, there are additional differences between cell types, which are most prominent in the adult cerebellum. Microglia, the immune cells of the brain, show the fastest divergence in their CREs (**Figure 3.32C-D**), in agreement with the rapid evolution of their gene expression programs and morphology (*262*). Microglia CREs are also enriched for genomic repeats, especially for TE classes with recent expansions in rodents, such as SINEs B1, B2 and B4, endogenous retrovirus sequences (ERVs) and L1 elements (**Figure 3.34A-B**). By contrast, astrocytes (assigned to the astroglial lineage together with multipotent cerebellar progenitors from earlier stages) have the most constrained CRE sequences in the mature cerebellum (**Figure 3.32C-D**). This increase in overall constraint is associated with a particular increase in the contribution of CREs that originated between 160 and 177 million years ago in common therian or mammalian ancestors (**Figure 3.32E**). Additionally, these mammalian-shared CREs show higher sequence constraint than what would be expected based on their evolutionary age (**Figure 3.34C**). Finally, differences amongst neurons are subtler than those I observed between glial cell types. The most prominent outlier is granule cells, which have the youngest and fastest evolving CRE sequences compared to other neuron types in the same developmental stage, with differences being most pronounced during prenatal development (**Figure 3.32C-D**). This enrichment for younger CRE sequences could be related to evolutionary innovations in the development of granule cells, such as the emergence of the proliferative granule cell layer in amniotes (*213*, *214*).

To test these observations against an independent dataset, I examined a single-cell ATAC-seq atlas of adult mouse organs, which covers a wide range of cell types. Consistently with my results, amongst cerebellar cell types, astrocytes show the highest and microglia the lowest sequence constraint in their chromatin accessibility profiles (**Figure 3.34D**). However, the wider set of cell types assayed by this atlas allowed me to further extend these comparisons. Despite having the fastest evolving CRE sequences in the cerebellum, microglia show the highest conservation when compared to other immune cell types (**Figure 3.34D**), highlighting the overall stronger evolutionary constraints in the brain (*167*, *169*, *172*). This is further supported by the fact that 8 out of the 10 most conserved cell types are found in the brain (**Figure 3.34D**).

More strikingly, astrocytes are marked by the most constrained CRE sequences not only in the mature cerebellum, but across all assayed cell types in the adult mouse (**Figure 3.34D)**. Based on my previous observations regarding development and pleiotropy, this increased conservation of CREs in astrocytes could be related to the latter maintaining some properties of neural progenitors (*263*) or to increased pleiotropic constraints due to their bridging interactions with multiple cell types, including neurons and the vasculature system (*264*). An alternative explanation could involve the requirement for a more rigid sequence grammar for CRE activity in astrocytes compared to other mature cell types. Collectively, with these analyses I revealed common temporal trends, as well as cell type-specific differences in the evolutionary histories of CREs that are active in the developing mouse cerebellum.

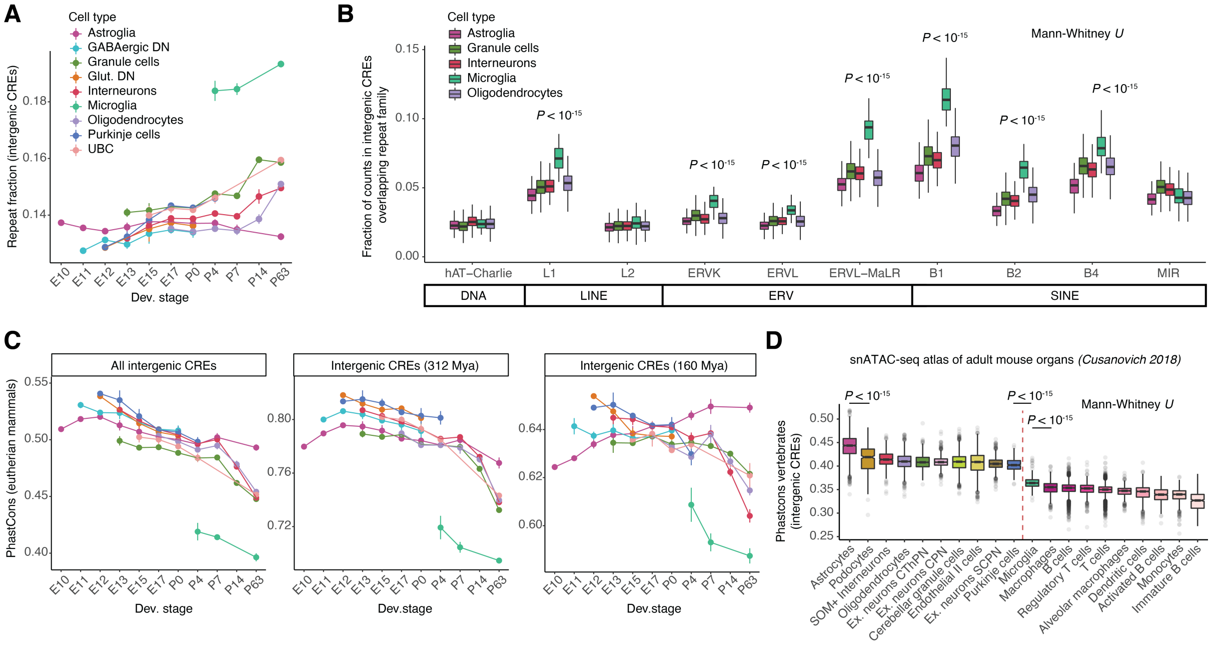


Figure 3.34: Differences in CRE constraint across cell types. **(A)** Fraction of fragments in intergenic CREs accessible per cell overlapping repeats, averaged for each cell type and developmental stage. **(B)** Fraction of fragments in intergenic CREs accessible per cell overlapping transposable elements of different classes, grouped by cell types of the adult mouse cerebellum (P63). **(C)** Sequence constraint across eutherian mammals for all intergenic CREs (left), and subsets that originated 312 (middle) and 160 (right) million years ago (Mya) accessible per cell, averaged for each cell type and developmental stage. Different y-ranges were used across age groups to facilitate comparisons between cell types and stages within each group, as sequence constraint is overall higher for older CREs. In **A** and **C**, vertical bars illustrate the difference in average estimates between biological replicates. **(D)** Sequence constraint of intergenic CREs accessible per cell across cell types in the adult mouse (data from (*62*)). The ten most conserved (left) and all immune (right) cell types are shown. *Figure adapted from Sarropoulos et al. 2021* (*244*).

3.4.3 CRE activity conservation in the mammalian cerebellum

The conclusions of the last two sections are based on assessing the presence and conservation of mouse CRE sequences in the genomes of other vertebrates. However, less than 50% of alignable enhancers between human and mouse show conserved activity in the same organ in both species (*46, 117*). Furthermore, even when sequences retain their capacity to act as CREs in two mammalian species, they often (35%) show repurposed activity across cell types or even organs (*120*). Thus, even though sequence constraint is an important predictor of shared CRE activity, it can often overestimate the true degree of conservation. To assess whether my observations based on sequence constraint extend to the level of regulatory activity conservation, I analyzed a snATAC-seq dataset from the marsupial opossum, which separated from eutherian mammals (including mouse) around 160 million years ago. The dataset was generated by Dr. Mari Sepp and included ~20,000 cerebellar cells across two developmental stages, P21 — which is transcriptionally similar to P4 in mouse (*172*) — and adult (**Figure 3.35A)**.

I analyzed the opossum dataset as described for mouse, identifying a total of 13 cell types and states (**Figure 3.35B-D)**. These match the cellular states identified in the corresponding stages in mouse (P4 and P63), often on the basis of the same marker genes (e.g., *SLC1A3* for astrocytes, *PAX2* for interneurons and *ETV1* for mature granule cells). Relative cell type abundances are also similar for corresponding stages between species, with the majority (79%) of the cells profiled in opossum corresponding to granule cells (**Figure 3.35D)**. These analyses support the overall conservation of the main cellular repertoire of the cerebellum during mammalian evolution.

I then used this dataset to identify a total of 167,341 putative CREs. Of the 72,080 mouse CREs active in the corresponding stages (P4, P63) and having alignable sequences in the opossum genome, only 34,989 (49%) were identified as open chromatin regions in the opossum cerebellum (**Figure 3.35E)**. Regulatory activity conservation is significantly higher for promoters compared to intronic and intergenic CREs (putative enhancers; **Figure 3.35E)**, in agreement with previous observations (*46, 117, 120*).

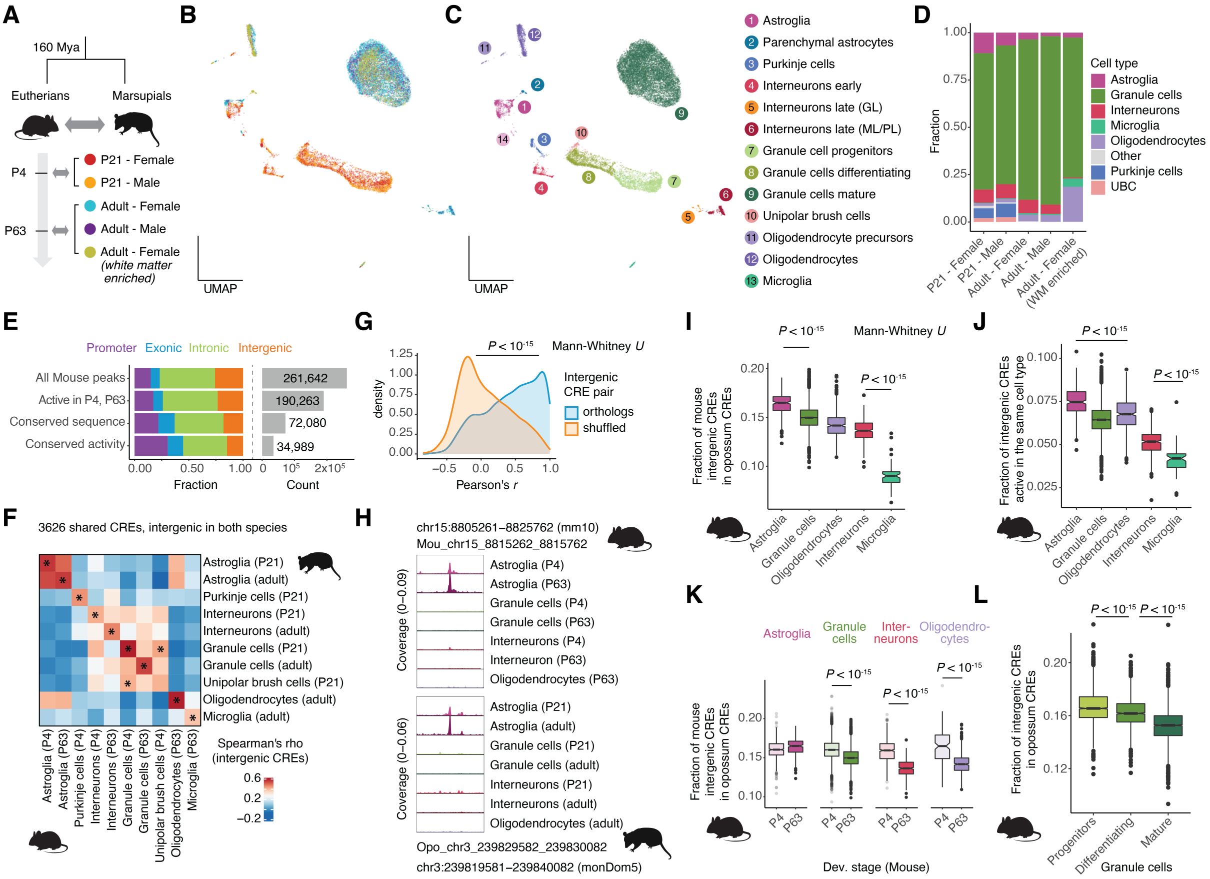


Figure 3.35: CRE activity conservation across therian mammals. **(A)** Overview of opossum snATAC-seq dataset and correspondence to mouse developmental stages based on transcriptomic similarity (*172*). **(B, C)** UMAP projection of 19,204 opossum cells colored by sample **(B)** or cell type and state **(C)**. **(D)** Proportions of broad cell types across samples. **(E)** Distribution of genomic classes (left) and total number (right) for all mouse CREs, CREs active (≥ 5 CPM) in cell types and developmental stages corresponding to those sampled in opossum, and subsets showing conservation of sequence (CRE aligned to genome) and activity (CRE aligned to CRE) in opossum. **(F)** Spearman's correlation of intergenic CRE activity between opossum and mouse corresponding cell types and stages. Stars mark the sample with the highest correlation for each row and column. **(G)** Pearson's correlation coefficients of activity (in CPM) across corresponding cell types and stages for intergenic CRE pairs with true or shuffled orthology relationships. **(H)** Example of a shared intergenic CRE with conserved cell type-specificity between mouse and opossum. Accessibility profiles for each broad cell type and stage were aggregated across cells and scaled by the total number of fragments in each group. **(I, J)** Fraction of mouse intergenic CREs accessible per cell with an ortholog CRE in opossum **(I)** or with an ortholog in the 40,000 most accessible CREs of the corresponding cell type in the adult opossum **(J)**, grouped by cell type in the adult mouse. **(K)** Fraction of mouse intergenic CREs accessible per cell with an ortholog CRE in opossum, grouped by cell type and developmental stage for cell types with at least 100 cells in both stages. **(L)** Fraction of intergenic CREs, accessible per cell in P4 mouse, with an ortholog CRE in opossum, grouped by differentiation state of granule cells. *Figure adapted from Sarropoulos et al. 2021* (*244*).

However, when focusing on intergenic CREs that were shared between the two species, I observed, that all homologous cell types (except for UBCs) show the highest similarity in their accessibility profiles (**Figure 3.35F**). Similarly, orthologous intergenic CREs show significantly higher correlations in their spatiotemporal activity profiles across corresponding cell types and developmental stages compared to shuffled homology relationships (**Figure 3.35G**). Thus, despite the overall rapid turnover of distal CRE activity, most of the CREs that have retained their activity in the cerebellum remain active in the same cell types and developmental stages. A mouse CRE located 100 kb downstream of *Slc1a3* and assigned to that gene based on my correlation analysis (see section 3.3.2) serves as an illustrative example. Both the gene and the CRE have maintained the same astroglia-specific activity in both species suggesting a regulatory interaction that has been conserved across 160 million years of mammalian evolution (**Figure 3.35H**).

Finally, I used the opossum dataset to reexamine my sequence-based comparisons of evolutionary conservation across cell types and developmental stages from the perspective of CRE activity conservation. Focusing on the adult cerebellum, the fraction of mouse CREs with an ortholog showing CRE activity in the opossum cerebellum is the highest for astrocytes and the lowest for microglia (**Figure 3.35I**), in agreement with my previous observations at the level of sequence conservation. The differences between cell types remain significant even when requiring the CRE to be active in the same cell type in both species (**Figure 3.35J**). In support of the developmental decrease in evolutionary conservation, the fraction of mouse CREs conserved in opossum is also higher at P4 compared to P63 for all major cell types except for astroglia (**Figure 3.35K**), which also shows similar levels of sequence constraint between the two stages (**Figure 3.32C-D**). Finally, to assess the effect of differentiation on regulatory activity conservation, I focused on mouse P4 granule cells. In agreement with my sequence-based observations, granule cell progenitors show the highest fraction of CREs conserved in opossum, followed by differentiating and mature granule cells (**Figure 3.35L**). Thus, even though only a subset of conserved CRE sequences also shows conserved activity in opossum, differences in constraint levels between cell types and developmental stages are similar when using either metric.

3.4.4 Comparative multiomic atlases of the mammalian cerebellum

The analyses described in the previous three sections focused on the identification of conserved elements and on comparisons of constraint across cell type and stages. Identifying CREs associated with evolutionary innovation is more challenging due to their rapid turnover, difficulties in associating CREs with genes that have changed their expression, and an incomplete understanding of how the sequences of CREs relate to their function. The next two sections attempt to set up an analytical framework that addresses these challenges, with the final section serving as a proof-of-principle analysis for the identification of CREs that are associated with gene expression changes between species.

To facilitate comparative analyses of gene expression and chromatin accessibility, I integrated the mouse snATAC-seq dataset with a new snATAC-seq dataset for corresponding developmental stages in human, as well as published (*179*) and newly generated snRNA-seq data for both species. All new data described in this section were generated by Dr. Mari Sepp. I used a neighbor-voting procedure followed by manual curation to annotate cell types and states in the new scRNA-seq samples and then transferred annotations to the snATAC-seq dataset in a stage-wise manner (**Figure 3.36A-B**). Collectively, this extended dataset is comprised of 239,627 cells (130,250 RNA and 109,377 ATAC) for human and 222,420 cells (117,114 RNA and 105,306 ATAC) for mouse.

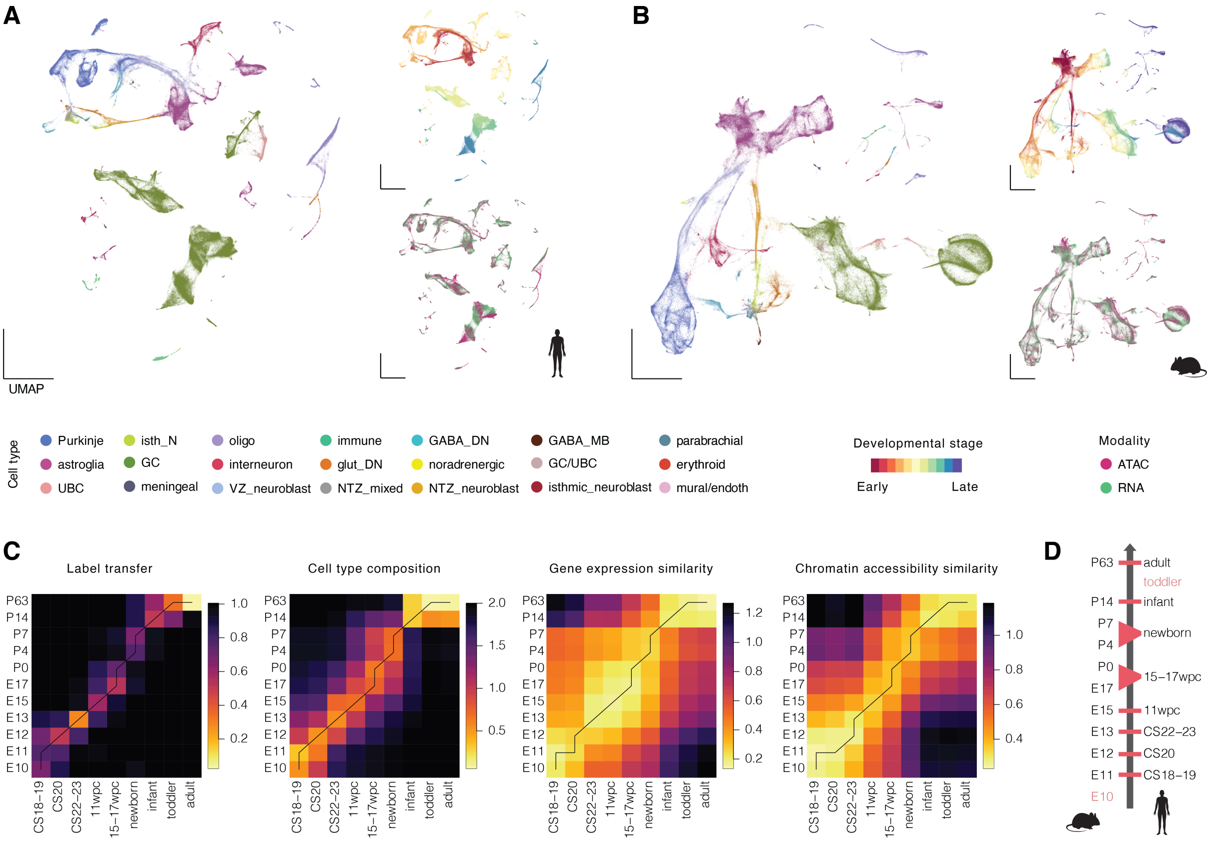


Figure 3.36: Multiomic atlases of cerebellum development in human and mouse. **(A, B)** UMAP projections of 239,627 human **(A)** and 222,420 mouse **(B)** cells colored by broad cell type (left), developmental stage (top-right) and modality (bottom-right). **(C)** Alignment between human and mouse developmental stages based on dynamic time-warping applied on different dissimilarity metrics. **(D)** Consensus correspondence of human and mouse developmental stages based on **C**.

I then used this dataset to identify putative CREs, as described for the mouse dataset (3.3.2), grouping cells at the finest level of cell type annotation (cell states and subtypes). Based on reciprocal syntenic alignments, I detected a total of 199,356 orthologous CRE pairs between the two species.

Next, I reassessed the correspondence in stages of cerebellum development between human and mouse, a necessary step for performing evolutionary comparisons. To this end, I applied a method based on dynamic time-warping, previously used by Dr. Margarida Cardoso-Moreira and later extended by Dr. Mari Sepp, Kevin Leiss and myself. I examined four different metrics that incorporate information about cross-species similarity in gene expression, chromatin accessibility and cellular composition (**Figure 3.36C**). All metrics broadly agree, resulting in a consensus matching of developmental stages between human and mouse (**Figure 3.36D**), which is also in line with previous work from the Kaessmann lab (*172*, *179*).

I then sought to summarize the spatiotemporal activity of CREs from both species across corresponding cell types and developmental stages. To this end, I aggregated CRE activity across cells from the same cell type and developmental stage. Following normalization within and between species (Methods), I applied non-negative matrix factorization (NMF) to identify 18 components summarizing chromatin accessibility patterns in the human and mouse cerebellum (**Figure 3.37A**). These components capture the major factors determining CRE activity, namely cell type and developmental stage, and broadly recapitulate the clusters of CRE activity I previously identified using the mouse dataset (**Figure 3.25A**). However, since pleiotropic CRE activity can be captured by additive contributions in multiple NMF components, rather than requiring a separate cluster, the same patterns can be described using fewer components than clusters. Notably, both human and mouse CREs contribute to all components, demonstrating that the major determinants of chromatin accessibility are shared between the two species (**Figure 3.37A**).

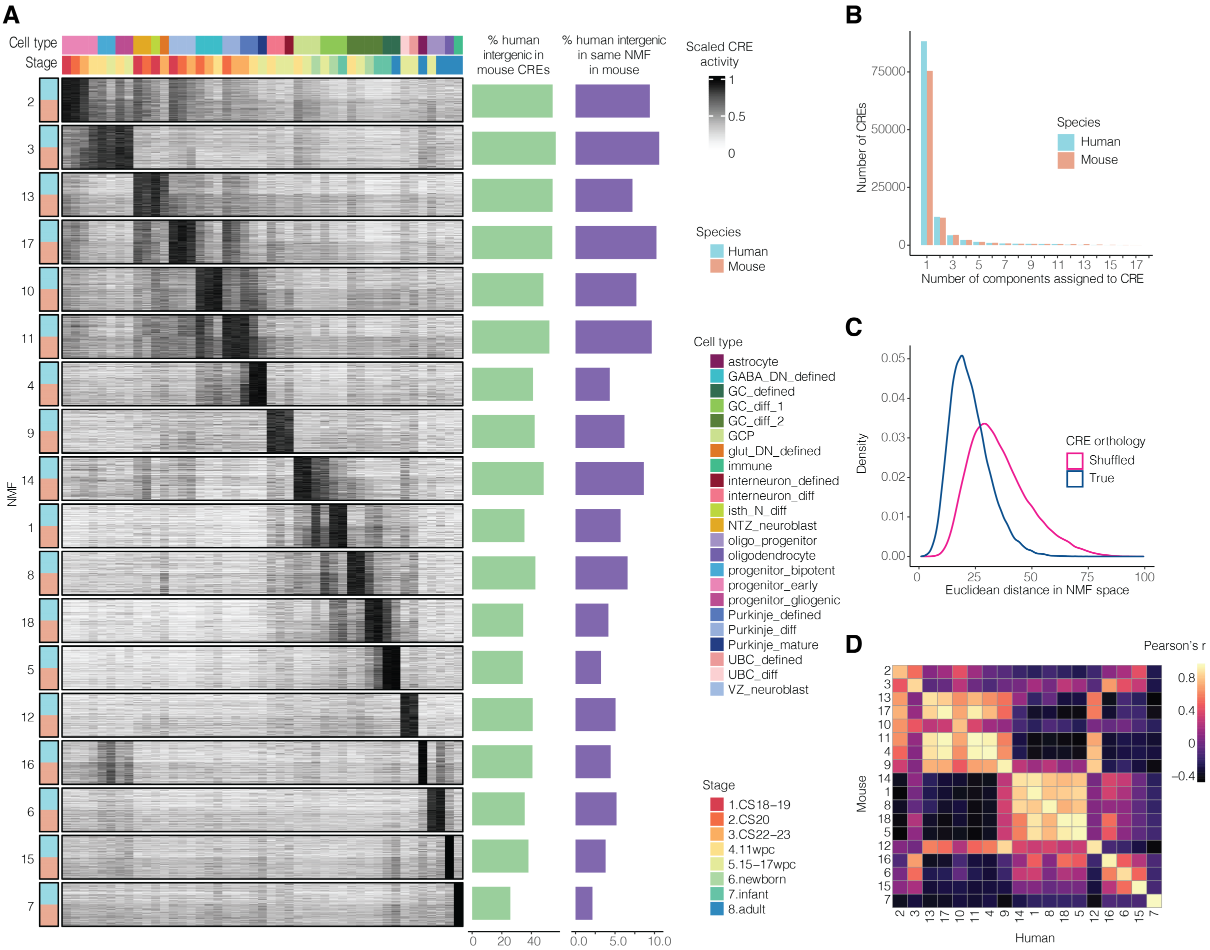


Figure 3.37: Spatiotemporal activity of human and mouse CREs. (A) CRE activity (quantile normalized CPM, scaled by the maximum activity of each CRE) across corresponding human and mouse cell types and developmental stages. 500 CREs with the highest contribution for each NMF component and low contributions to all other components were chosen for each species for visualization. The barplots on the right show, for each NMF component, the fraction of human intergenic CREs with a mouse CRE ortholog (green) and a mouse CRE ortholog assigned to the same component (purple). **(B)** Number of components assigned to each human and mouse CRE. **(C)** Euclidean distance in contributions to NMF components between human and mouse CRE orthologs (blue) or pairs with shuffled orthology relationships (violet). **(D)** Similarity (Pearson's *r*) in TF motif enrichment (normalized enrichment scores obtained from cisTarget) for mouse (rows) and human (columns) CREs assigned to each NMF component.

In line with my previous observations for mouse **(Figure 3.25)**, most CREs (78% in human and 75% in mouse) are associated with a single component, supporting their overall high context-specificity **(Figure 3.37B)**. Orthologous CREs between human and mouse show significantly higher similarity in the NMF space compared to shuffled orthology pairs **(Figure 3.37C)**, recapitulating the notion that a sizeable subset of CREs has retained the same spatiotemporal activity during mammalian evolution. However, overall CRE turnover is fast, with at most 56% of human intergenic CREs also being active in mouse **(Figure 3.37A)**. This percentage decreases further to a maximum of 11% of human intergenic CREs having a mouse ortholog with a high contribution to the same NMF component (**Figure 3.37A**). Thus, the majority of CREs contributing to each component are specific to the human or mouse lineage.

The relative conservation of CRE sequences and activity patterns across NMF components also agrees with my previous observations. Components associated with early development (2, 3, 10, 11, 13 and 17) overall show the highest conservation (**Figure 3.37A**). Mature cell types are typically less conserved than their differentiating precursors (e.g., 11 vs 4 for Purkinje cells, 14 vs 5 for granule cells and 6 vs 15 for oligodendrocytes; **Figure 3.37A**). Finally, amongst components associated with adult cell types (5, 7, 15 and 16), astrocytes (component 16) show the highest conservation (**Figure 3.37A**). By contrast, component 7, associated with immune cells (i.e., microglia) shows the fastest divergence (**Figure 3.37A**). Collectively, this integrated analysis of human and mouse CRE activity across a set of corresponding cell types and developmental stages recapitulated the main findings of my previous work on the mouse dataset.

In contrast to the fast divergence of individual CREs (*46*, *120*), the activity and DNA binding preferences of TFs are generally thought to be conserved across mammals (*124–126*). I thus sought to examine whether CREs with the same spatiotemporal activity (i.e., assigned to the same NMF component) contain motifs recognized by the same TFs in the two species. Although most CREs associated with each component are lineage-specific, I observed high similarity in the TF motifs enriched amongst human and mouse CREs assigned to the same component (**Figure 3.37D**). Thus, the spatiotemporal activity of CREs during the development of the mammalian cerebellum is determined by a set of TFs that are largely conserved between human and mouse.

However, despite the high similarity observed for CREs from the same component between the two species, there are many cases where multiple components have similar motif content (**Figure 3.37D**). For example, components 4, 11, 13 and 17 are associated with early-born cell types and are mostly enriched for Homeobox motifs, whereas components 1, 5, 8, 14 and 18 are associated with various stages of granule cell differentiation and are highly enriched for sequences recognized by bHLH TFs (**Figure 3.37D**). Even though each of these components has a distinct set of CREs bound by a different set of TFs (e.g., ATOH1 in granule cell progenitors, NEUROD1 in differentiating granule cells), the similarity in the motifs recognized by these TFs masks their differences at the level of motif content. The next section describes the use of newly developed computational methods to gain insights into how the activity of CREs is encoded in their sequence, and to investigate whether these more subtle sequence features are also shared between human and mouse.

3.4.5 Cell type-specific regulatory grammar is conserved across mammals

Recent advances in the field of deep learning have enabled the development of models that can predict chromatin accessibility patterns based on DNA sequence (*99–101*). Combined with dedicated interpretation tools (*106, 107*), such models can subsequently highlight the sequence features contributing to their predictions (i.e., the parts of the CRE sequence that the model considers important for determining CRE accessibility), providing insights into the regulatory grammar of different cell types.

Guided by Prof. Dr. Stein Aerts at the University of Leuven and his lab members, Dr. Nikolai Hecker, Ibrahim Taskiran and Carmen Bravo González-Blas, I developed such models to study the regulatory grammar of the developing mammalian cerebellum. First, I explored the mouse data to determine the feasibility of this approach. For each NMF component from **Figure 3.37A**, I extracted the 3,000 mouse CREs with the highest contribution in that component. I then trained a multi-class and multi-label classifier which utilizes convolutional and recurrent neural networks, as described by *Minnoye et al. 2020* (*100*) and *Janssens et al. 2022* (*101*), to predict the NMF component (class) from the DNA sequence (input) of each CRE. To help the model distinguish cell type-specific CREs from sequences with noisy or no CRE activity, I additionally included two more sets of sequences: 3,000 CREs with low contributions across all 18 components (component 19, i.e., noisy peaks) and 3,000 random 500 bp-wide intervals that don't overlap any ATAC peak in the cerebellum dataset. In accord to commonly used approaches, 80% of the data were used for training, 10% for validation and 10% for testing the final model.

To evaluate the performance of the best model trained using only mouse sequences, I estimated the area under the receiver operating characteristic (auROC) and the area under the precision-recall (auPR) curves based on the 10% of CRE sequences reserved for the testing set (**Figure 3.38A**). The median auROC and auPR estimates across all labels for the test set are 0.91 and 0.5 (compared to 0.5 and 0.06 respectively when using shuffled labels) and are comparable to (and even slightly higher than) those achieved for other datasets with a similar model architecture (*100, 101*). Thus, the model was able to learn sequence features that allow it to predict the spatiotemporal activity patterns of previously unseen mouse CRE sequences.

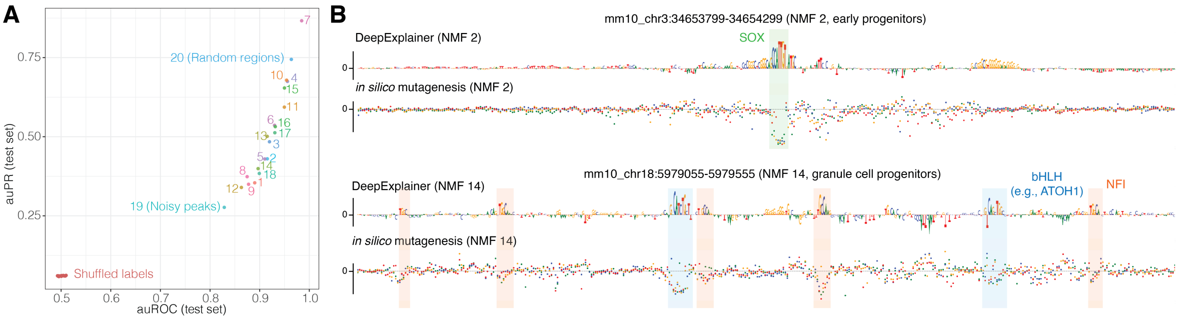


Figure 3.38: Predicting and interpreting the accessibility patterns of mouse CRE sequences. **(A)** Prediction accuracy (x-axis: auROC, y-axis: auPR) for each NMF component class (from **Figure 3.37**), versus random predictions (shuffled labels) for mouse CRE sequences in the testing set. **(B)** DeepExplainer and *in silico* mutagenesis profiles of CRE sequences with high contributions to NMF components 2 (top) and 14 (bottom). Motifs of relevant TFs are highlighted in the boxes.

To further investigate whether the model is able to achieve this task by recognizing sequence features that match our current mechanistic understanding of CRE function (e.g., whether it identifies relevant TF motifs), I used two independent but complementary approaches. The first relies on DeepExplainer (*106*), a method that assesses the relative contribution of different parts of the sequence to the model's prediction. Thus, the DeepExplainer score of each nucleotide reflects its positive or negative contribution to the model's prediction for a particular class. The second approach relies on perturbing each nucleotide of the sequence into all other possible nucleotides and monitoring the effect of the change to the prediction, a process termed "*in silico* mutagenesis". Parts of the sequence that contribute positively to the model's prediction are expected to have negative *in silico* mutagenesis scores, as the prediction drops when they are mutated. As can be seen for two examples of mouse CREs (**Figure 3.38B**), the model's predictions are based on the identification of short sequences that correspond to known motifs (and often their flanking sequences) for major TF regulators of each cell type (e.g., SOX motifs for a CRE active in early progenitors, bHLH and NFI motifs for a CRE active in granule cell progenitors). Thus, the model's high prediction accuracy is based on its ability to identify biologically relevant features of a regulatory sequence, a feature that can be used to study the grammar of CREs.

The high similarity in motif content of CREs assigned to the same component in human and mouse (**Figure 3.37D**) suggests that the sequence features responsible for CRE activity might be to some degree shared between the two species. To test this hypothesis, I trained a similar model based on human CREs and then evaluated the performance of each of the two models in predicting the accessibility of CREs from the same or other species. Surprisingly, both models perform similarly and cross-species predictions are almost indistinguishable in terms of accuracy from within-species predictions (**Figure 3.39**). This suggests that the sequence basis of spatiotemporal CRE activity in the developing cerebellum has remained largely conserved over the last 90 million years of evolution and that both models were able to learn similar aspects of it. Notably, unlike the many-to-many matches observed between related NMF components in terms of TF motif content (**Figure 3.37D**), these sequence-based models are able to distinguish well even between closely related components (**Figure 3.39**). This could possibly be achieved by the model's ability to consider additional sequence features besides the aggregate motif content, such as the distance between adjacent motifs or the composition of their flanking sequences. Detailed mechanistic work has now established that such features are indeed important determinants of CRE activity (*103*, *104*, *265–267*).

The ability of deep learning models to predict spatiotemporal patterns of CRE activity based on DNA sequence, even beyond the species they were trained on, makes them promising tools to study the evolutionary histories of CREs. Specifically, I reasoned that these models would allow me to retrace the evolutionary steps that led to the emergence of new CREs by predicting the accessibility patterns across a range of publicly available mammalian genomes and by inferring the effect of individual sequence changes on the activity of these regions. Even though either the mouse or human model could be used for this purpose, I decided to additionally train models with sequences from both species to minimize the effect of any lineage-specific confounders. Using 3,000 regions per NMF component sampled from both species leads to a slight increase in prediction accuracy (median auPR=0.47 across sequences from both species compared to 0.43-0.44 when using a single-species model). Additionally utilizing the full dataset (3,000 human and 3,000 mouse for a total of 6,000 CREs per component) led to the best performing model with a median auPR of 0.5 across classes and species. This model was used for all analyses described below.

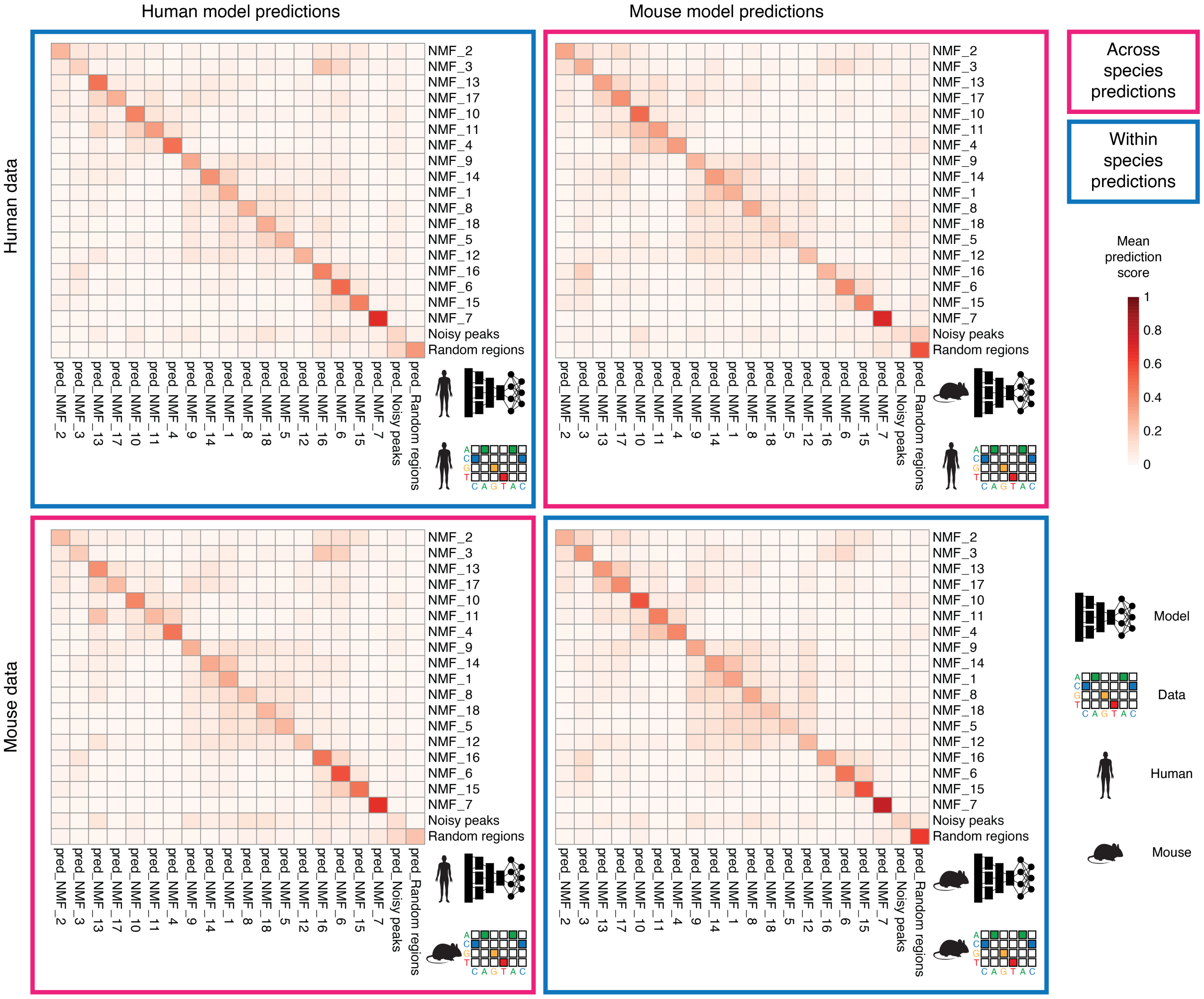


Figure 3.39: The regulatory grammar of cerebellar cell types is conserved between human and mouse. Mean prediction scores of models trained on human (left) and mouse (right) CRE sequences across NMF classes (columns) for human (top) and mouse (bottom) CRE sequences (rows) reserved for the testing set (i.e., not seen by any model during training).

Before investigating CREs with lineage-specific activity, I tested the feasibility of my approach on conserved CREs. **Figure 3.40** shows an example of an intergenic CRE with conserved activity in both human and mouse. Although accessible in multiple cell types, this CRE shows the highest accessibility in granule cell progenitors in both species. In both human and mouse, the model correctly predicted a high contribution to component 14 (which is associated with granule cell progenitors). Investigating the sequence features that contribute to this prediction, I identified three bHLH motifs in each ortholog (most likely recognized by ATOH1). Incorporating information from genomic alignments across vertebrate genomes (PhyloP scores), I observed higher sequence constraint in the position of these three motifs compared to most adjacent regions (**Figure 3.40B**), supporting the notion that these three motifs are indeed important for the conserved activity of this CRE. In the next section, I describe the application of a similar analytical framework to study the emergence of evolutionary novel CREs which are associated with changes in gene expression between human and mouse.

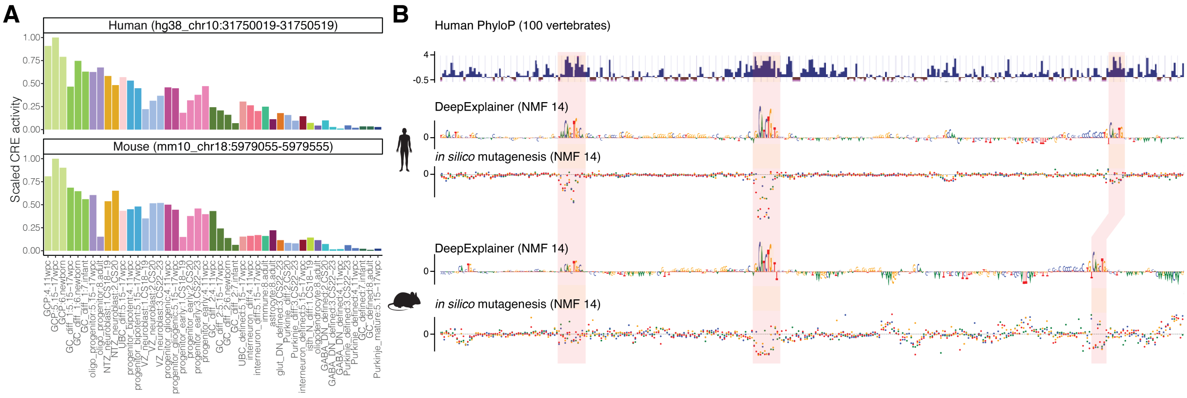


Figure 3.40: Investigating the sequence basis of CRE activity conservation. (A) Spatiotemporal activity (quantile normalized CPM, scaled by the maximum activity for each CRE) for the human (top) and mouse (bottom) ortholog of a conserved CRE. **(B)** DeepExplainer and *in silico* mutagenesis profiles based on the NMF component 14 (granule cell progenitors) for the human (top) and mouse (bottom) sequences of the conserved CRE shown in **A**. The red boxes highlight three bHLH motifs considered important by the model and showing high sequence conservation across vertebrates (PhyloP score from UCSC shown above the human CRE sequence).

3.4.6 CREs associated with changes in gene expression

In the final part of my thesis research, I sought to identify and characterize CREs that are associated with changes in gene expression in at least one context (cell type and developmental window) between human and mouse. First, I developed a new method to conservatively identify changes in gene expression between species, requiring both quantitative and qualitative changes in expression profiles and incorporating information from biological replicates (Methods). To facilitate integration with the patterns of CRE activity described above, I centered this analysis on the NMF components I identified previously (**Figure 3.37A**). For each NMF component, I selected the samples with the highest loadings (e.g., for component 2: progenitor cells from CS18/E11 to CS23/E13 in human/mouse, respectively). Within the samples associated with each component, I then identified genes that show higher expression in human or mouse. In total, I identified 448 genes with higher expression in human and 753 genes with higher expression in mouse, with a median of 69 genes detected as differentially expressed in each component (**Figure 3.41A**). Even though the overall number of detected genes per component is likely also affected by technical differences in the power to identify these changes, components associated with mature cell types consistently show more changes in expression (348 in microglia/NMF7, 236 in oligodendrocytes/NMF15, 134 in mature granule cells/NMF5, 116 in astrocytes/NMF16) compared to progenitors and differentiating cells (e.g., 61 and 14 genes in early/NMF2 and late/NMF3 progenitors respectively, 39 genes in granule cell progenitors/NMF14). **Figure 3.41B** illustrates the expression profiles of four genes identified as having higher expression in humans in at least one NMF component. In the absence of an outgroup species (something I plan to implement in future work), genes with higher expression in humans include those that have gained expression in the human lineage and those that have lost expression in the mouse lineage. However, in the case of these four genes, manual investigation of the expression profiles of their orthologs in the marsupial opossum (*179*) revealed that they have all gained expression in the respective cell types in the human lineage (i.e., they are not expressed in the corresponding cell type in either mouse or opossum).

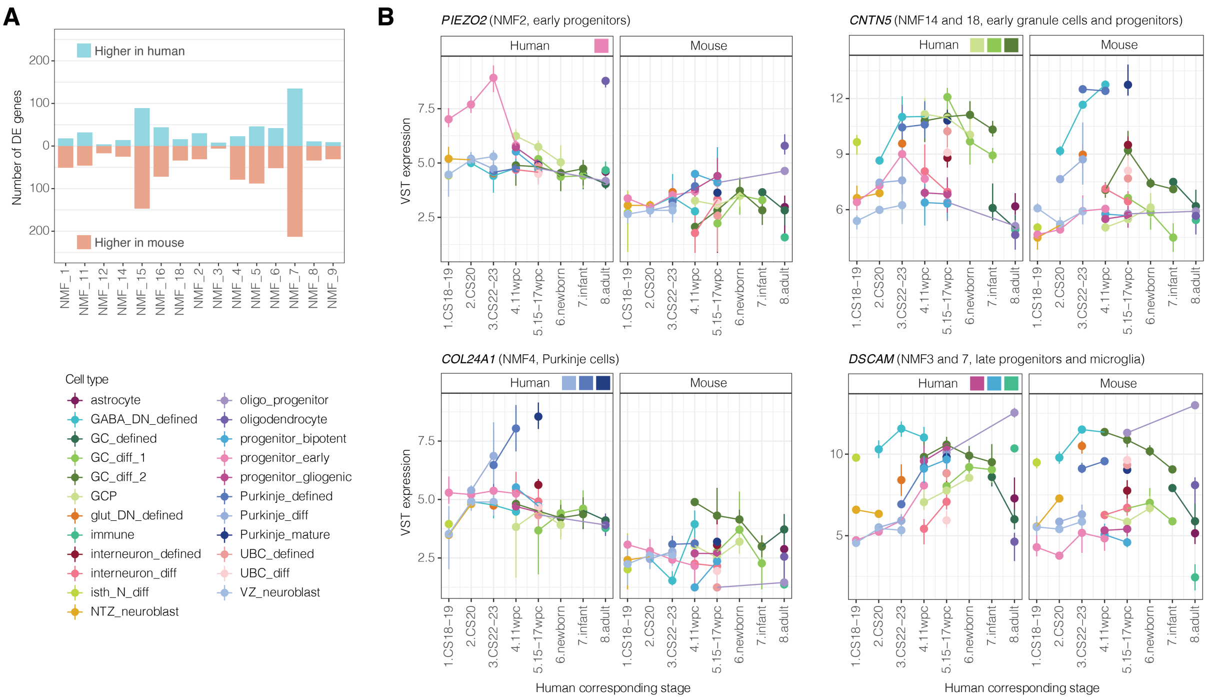


Figure 3.41: Cell type-specific changes in gene expression between human and mouse. **(A)** Number of genes with significantly higher expression in human (blue) or mouse (orange) in samples associated with different NMF components (from **Figure 3.37**). **(B)** Spatiotemporal expression profiles across corresponding cell types and developmental stages for genes that have gained expression (also considering comparisons with opossum) in human cell types. Vertical lines show the maximum and minimum variance stabilized (VST) expression across replicates, dots show the mean. Rectangles indicate cell types with higher expression in human.

I next examined how these evolutionary changes in the expression of genes are related to their chromatin accessibility landscapes. Human CREs adjacent (up to 250 kb from a TSS) to genes with higher expression in human show significantly higher contributions to the corresponding component than those close to genes with higher expression in mouse **(Figure 3.42A)**. Human CREs around genes with higher expression in human are also enriched for CREs that are not shared with mouse (either gained in human or lost in the mouse lineage) compared to those with higher expression in mouse or showing no evidence of change (unresolved) **(Figure 3.42B)**. Finally, even those human CREs that are shared with mouse show higher contributions to NMF components in which the adjacent gene is detected as more highly expressed in humans **(Figure 3.42C)**. Thus, these analyses show that gene expression changes between species are reflected in their broad chromatin accessibility neighborhood and are associated with gains, losses or repurposing of CRE activity in the corresponding cell types.

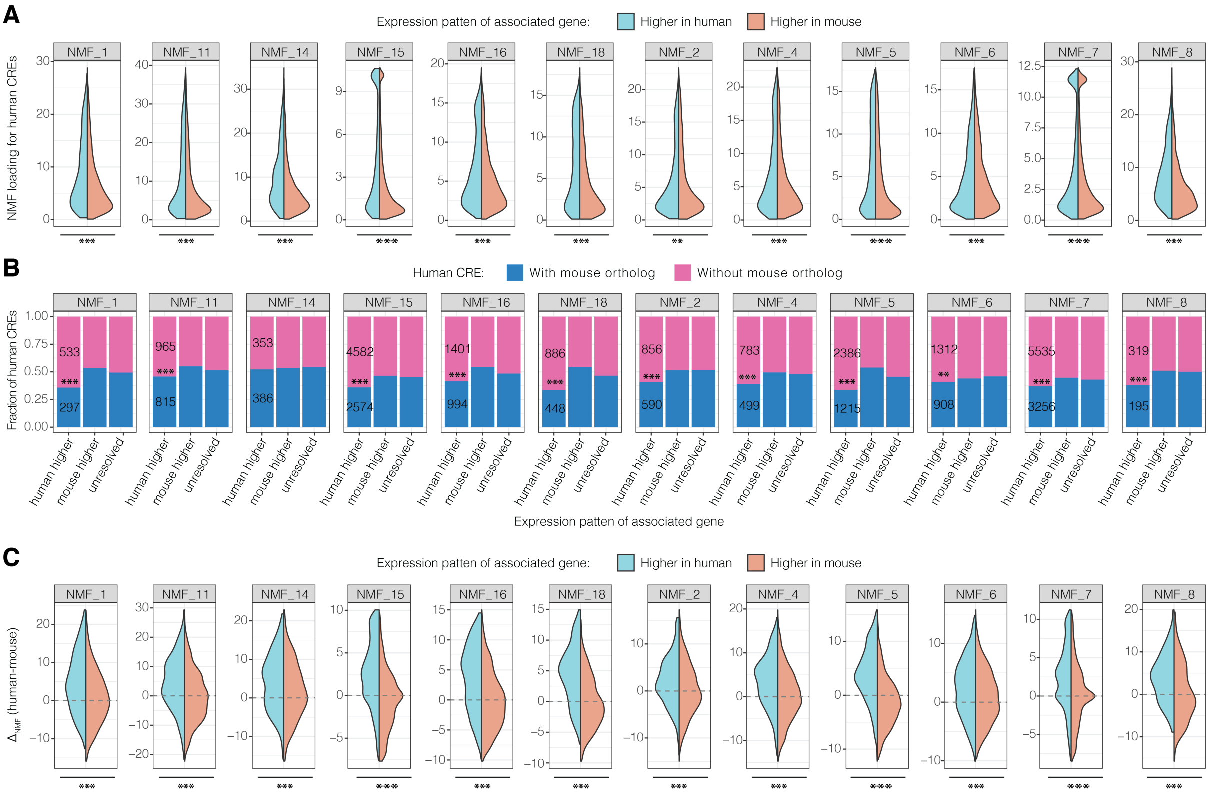


Figure 3.42: Chromatin accessibility landscape of genes with evolutionary changes in expression. **(A)** Distribution of NMF loadings for human CREs within 250 kb of a TSS of a gene with significantly higher expression in human (blue) or mouse (orange) for samples associated with the respective NMF component. Human CREs close to genes with significantly higher expression have higher loadings in the respective component. **(B)** Fraction of human CREs with (blue) or without (pink) a mouse ortholog for sets within 250 kb of a TSS of genes with higher expression in one species or without significant differences (unresolved). Genes with higher expression in humans are enriched for CREs not found in mouse (due to gain of new CREs in human or loss of old CREs in mouse). **(C)** Difference in NMF loadings between the human and mouse CRE orthologs within 250 kb of a TSS of a gene with significantly higher expression in one of the species. Orthologous CRE pairs close to genes with higher expression in human show higher contribution to the corresponding NMF component. *Mann-Whittney U tests (**A, C**) or Fisher's exact test between human-higher and unresolved (**B**) were used for statistical comparisons (P<0.05*, P<0.01**, P<0.001***).*

Despite their signal being detectable in these global comparisons (**Figure 3.42**), the CREs that are associated with a specific gene expression change likely represent only a small fraction amongst all CREs in the 500 kb region flanking that gene. To identify such cases I additionally focused on adjacent CREs that have high contributions in the NMF component where the change was detected, and only in the species where the gene is more highly expressed (either because the CRE is absent from the other species or because it is not active in the respective cell type and developmental window). Focusing on genes with higher expression in humans, I identified a total of 2,603 such CREs (median of 11 CREs per gene, representing 76% of all CREs with high contributions in the respective NMF component around these genes). In comparison, the median number of such CREs next to human genes that are also expressed in mouse is 6 (and only representing 54% of all peaks with high NMF contributions around these genes).

I next used my deep learning model (see section 3.4.5) to further interrogate the predicted accessibility profiles of these CREs in other mammalian species, as well as to investigate the sequence basis of their activity. Automating and scaling up such analyses is still part of my future work plans in the Kaessmann lab. Here, I present a proof-of-principle analysis focused on *PIEZO2* which has gained expression in early progenitor cells (NMF2) in the human lineage. There are 13 human CREs with high contributions in NMF2 within 250 kb of a TSS of *PIEZO2*. Of these, 8 have a high contribution only in human (6 not found in the mouse CRE annotation and 2 annotated as CREs but inactive in early progenitor cells in mouse). Using the deep learning model, I obtained high prediction scores (≥ 0.2) for NMF2 for 3 of these 8 human CREs. While the remaining 5 CREs might still be functionally significant, I did not consider them further for this analysis, as the model cannot be confidently used to predict or interpret their chromatin accessibility profiles based on their sequences.

The following analyses were performed with the help of a first-year PhD student in the Kaessmann lab, Tetsuya Yamada, based on my original project design and under my supervision. Mr. Yamada used a resource of genomic alignments (*268*) to identify corresponding sequences for the remaining 3 CREs and to predict their chromatin accessibility patterns (NMF loadings) in 229 mammalian species **(Figure 3.43A).** Two of these CREs (hg38_chr18:11,199,429-11,199,929 and hg38_chr18:11,258,300-11,258,800) are located 50 and 108 kb upstream from the *PIEZO2* TSS respectively and their sequences could be detected in most of the mammalian species examined (64% and 79% respectively). However, according to the deep learning model, hg38_chr18:11,199,429-11,199,929 is more likely to be accessible in early cerebellar progenitor cells (NMF2) in the old-world monkeys (and two more species from unrelated mammalian lineages; **Figure 3.43A**). The model's prediction for the human sequence mainly relies on one MEIS and three SOX motifs **(Figure 3.43B)**, which overall also show higher sequence conservation across old-world monkeys compared to their flanking regions (**Figure 3.43C**), suggesting that they are indeed important for the conserved activity of this CRE.

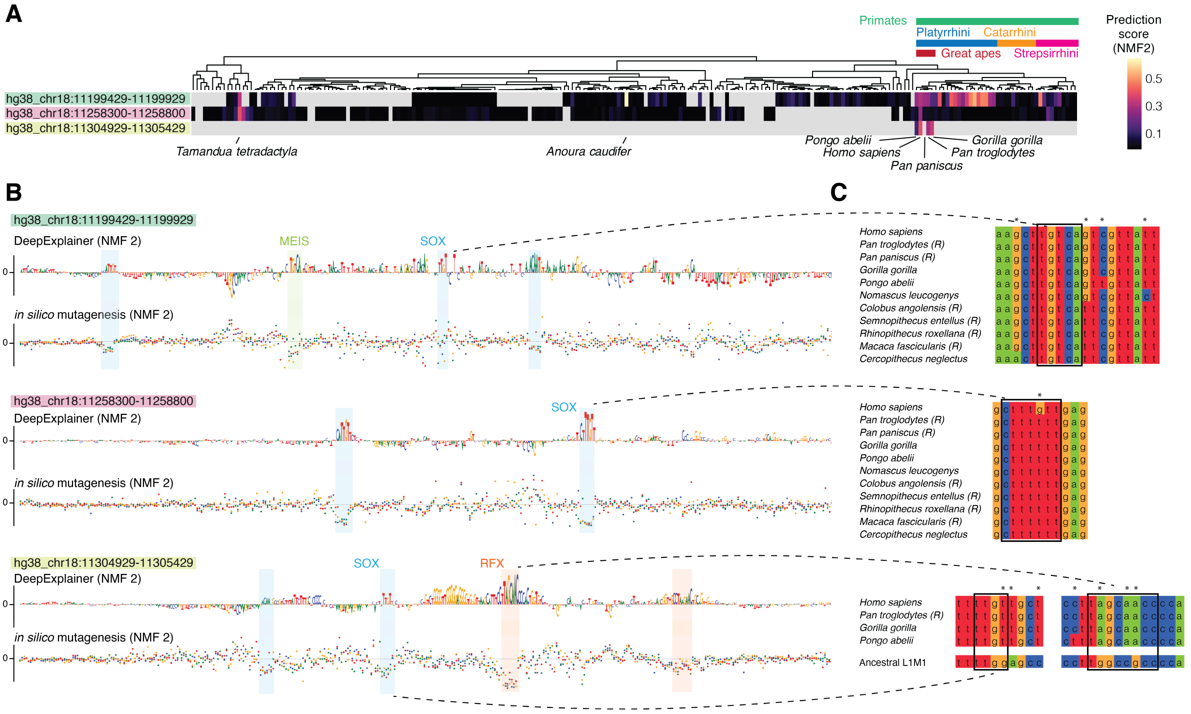


Figure 3.43: The evolution of CRE activity around *PIEZO2* in the human lineage. **(A)** Prediction scores for NMF2 (early progenitors) for sequences orthologous to human CREs around *PIEZO2* across 229 mammalian species. The tree summarizes phylogenetic relationships between the species, different primate taxa are highlighted. Gray rectangles mark species where no orthologous sequence could be identified. **(B)** DeepExplainer and *in silico* mutagenesis profiles for the three human CREs from **A**. Colored rectangles highlight known TF motifs recognized as important by the deep learning model. **(C)** Multiple sequence alignments for regions corresponding to TF motifs recognized by the model across selected primate species. The alignment for hg38_chr18:11,304,929-11,305,429 also includes the ancestral sequence of L1M1 repeats. *The identification of CRE orthologs across mammals and the multiple sequence alignments were performed by Tetsuya Yamada, under my supervision.*

More surprisingly, despite its sequence being found in more than 100 mammalian species, hg38_chr18:11,258,300-11,258,800 is only predicted to be accessible in early cerebellar progenitors in human and two Xenarthra species (**Figure 3.43A**). In this case, the model recognized two strong SOX motifs in the CRE sequence (**Figure 3.43B**), one of which has been created by a single nucleotide substitution (T→G) that occurred after the split between the human and chimp lineage (**Figure 3.43C**). The presence of this CRE sequence in the genome of most placental mammals suggests that it could have been serving a conserved role in a different organ and developmental stage before eventually becoming repurposed to also gain accessibility in human early cerebellar progenitors through this single nucleotide mutation. In support of this hypothesis, this CRE has been previously identified by the ENCODE project

(EH38E1900507, *source: https://screen.encodeproject.org/*) and is highly accessible during human embryonic development in the retina and spinal cord.

Finally, the most distal of the three CREs (hg38_chr18:11,304,929-11,305,429; 155 kb upstream from the *PIEZO2* TSS) has a very different evolutionary history. The sequence emerged from the insertion of a transposable element, L1M1, and can only be detected in four out of five extant great ape species (missing entirely from the bonobo genome; **Figure 3.43A**). Despite the CRE sequence only being present in four species, the model predicts high accessibility in early cerebellar progenitors in all four of them **(Figure 3.43A)**. Notably, sites corresponding to TF motifs that are considered important by the model have diverged considerably from the ancestral L1M1 sequence, but show no additional differences between the four great apes **(Figure 3.43B-C)**. Thus, this CRE seems to have arisen through the exaptation of a TE sequence in the ancestor of all great apes, and has since been selectively preserved during evolution.

Collectively, these proof-of-principle analyses provide a framework for the identification of genes with radical changes in their expression during mammalian evolution, and the subsequent investigation of adjacent CREs that could be associated with these changes. Future directions for improving and expanding these analyses are discussed in section 4.5. More generally, the last three sections of this chapter established that deep learning models can be used to study the sequence basis of CRE activity, that the sequence rules determining cell type-specific CRE activity are largely conserved between mammals, and that deep learning models can serve as powerful tools to study the evolution of non-coding regulatory sequences.

4. Discussion

During my dissertation work I explored the contributions of non-coding genomic elements to the development and evolution of mammalian organs. I focused on two major classes of non-coding elements, lncRNAs and CREs. Even though analyzed separately, based on different methods and at different levels of resolution (e.g., organs versus cell types), I identified many features that are shared between these two sets of non-coding elements. These extend to their context-specificity, developmental dynamics and evolutionary histories.

4.1 Context-specificity of non-coding elements

Both lncRNAs (*47*, *48*) and CREs (*55*) are known to show more specific activity compared to protein-coding genes, something also supported by my findings. Most lncRNAs show dynamic expression in a single organ, whereas my analyses of CREs in the mouse cerebellum revealed that most of them are active in a single cell type and in a restricted developmental window. Although my work did not examine whether the specificity of lncRNAs also extends to the level of cell types, other studies suggest that this is indeed the case (*61*).

This high cell type- and time-specificity adds to the notion that complex gene expression patterns are shaped by modular contributions from multiple context-specific non-coding elements (*36*, *37*). Indeed, my analyses showed that genes with important developmental roles, such as transcription factors and components of signaling pathways, are associated with large numbers of lncRNAs and CREs. Such genes have long been known to be flanked by large intergenic regions, being embedded in “gene deserts” (*269*). Thus, it has been debated whether the large number of non-coding elements around these genes might be a mere consequence of their large intergenic space (*232*). However, my analyses showed that many of the lncRNAs and CREs adjacent to developmental genes show at least partially overlapping spatiotemporal activity patterns, supporting the possibility of *bona fide* regulatory interactions between them. Thus, the large intergenic space surrounding developmental genes is likely an important feature of their complex regulatory landscapes.

The high context specificity of non-coding elements suggests that profiling their spatiotemporal activities can have important implications for the diagnosis and treatment of human disease. Thousands of lncRNAs have been shown to be specifically expressed in the context of disease, most notably in cancer (*270*). Even if these transcripts play no role in disease progression, they can serve as biomarkers allowing earlier diagnosis (*271*). Specific signatures of lncRNA expression have also been used to stratify cancer subtypes and to predict survival in patients with cancer (*272*) or heart failure (*273*). More recently, a population-scale transcriptomic analysis linked 800 lncRNAs to disease-associated polymorphisms that could not be explained by neighboring protein-coding genes (*274*). Such analyses are more likely to reveal lncRNAs with causal associations with disease, paving the way for the development of therapeutic compounds that directly target lncRNA transcripts, such as antisense oligonucleotides and duplex RNAs (*275*). Given the specificity of lncRNA expression, such strategies are expected to have more limited side-effects compared to approaches targeting broadly-expressed protein-coding genes, even without using organ-specific delivery systems.

Cell type-specific CREs are also important in the context of human disease. Rare mutations in enhancer sequences (*143*, *276*) or structural variants leading to aberrant gene activation (*277*) have been linked to a series of diseases, collectively referred to as enhanceropathies (*144*). In such cases, targeted genome editing of the affected enhancer can be considered as a treatment option (*144*). For complex diseases, such as metabolic or neurodegenerative disorders, the vast majority of trait-linked genomic variants lies in non-coding regions and is significantly enriched for CRE annotations (*144*). Although the effect size of each individual variant is too small to warrant genome editing-based treatments, jointly considering all variants can help estimate the overall disease risk for an individual based on a personalized polygenic risk score (*144*). Additionally, incorporating information about the spatiotemporal activity of CREs overlapping these variants can offer valuable insights regarding the — often still unknown — molecular and cellular basis of different diseases (e.g., the recently discovered association between microglia and Alzheimer's disease), guiding the development of future treatment strategies (*144*, *278–280*). Finally, the cell type-specific activity of CREs is increasingly being harnessed for the targeted delivery of gene therapy products (*281*). Combining the transgene with context-specific CREs can ensure its targeted expression in the cell type of interest, minimizing off-target effects (*281–283*). Taken together, a precise description of the spatiotemporal activities of non-coding genomic elements is of great value for diagnosing, understanding, and treating human disease.

Although most mammalian lncRNAs and CREs are context-specific, my analyses also identified sizeable sets showing pleiotropic activity (i.e., shared between cell types, organs and developmental stages). This adds to a growing appreciation of the potential of some CREs to act across a diverse set of cell types and developmental stages (*149*, *229*, *254*). Such pleiotropic elements tend to be active in early organ development, when the molecular profiles of different organs and cell types are overall more similar to each other. Furthermore, for both lncRNAs and CREs, pleiotropic activity is associated with stronger evolutionary conservation.

In the context of CREs, there are several non-mutually exclusive hypotheses for this association between pleiotropy and increased sequence conservation (*254*, *284*). If the activity of pleiotropic CREs in different conditions depends on different TFs, these CREs are expected to contain more TF binding sites than context-specific CREs, leading to a higher fraction of their sequence being under constraint (*254*, *284*). In support of this model, a subset of mutations introduced into mouse developmental enhancers with nearly perfect sequence conservation across vertebrates abolished their activity in E14 but not in E11 embryos (*285*), suggesting that activity over multiple developmental stages is encoded in different parts of their sequences. Alternatively, some pleiotropic CREs may reuse the same motifs across contexts, for example because they are identified by different TFs with similar sequence specificities or by the same broadly expressed TF (*254*, *284*). Even in this case, several studies reported that pleiotropic enhancers overall contain more TF motifs than context-specific ones, which would also lead to an increase in the constrained sequence space (*254*, *286*). In support of this hypothesis, the systematic dissection of a developmental fly enhancer revealed that almost every mutation led to some change in the expression of its target gene, often affecting it in multiple ways, suggesting that pleiotropic information can be encoded in the same sequence motif (*287*).

Both of the aforementioned models mainly explain the higher evolutionary conservation of pleiotropic enhancers through a more complex regulatory architecture, which leads to an expansion of the constrained space without increasing the maximum level of constraint observed throughout the sequence. However, even when comparing a pleiotropic and a context-specific CRE with identical TF motifs arranged in the same way, one would expect the pleiotropic CRE to be under stronger constraint. In this case, the fraction of constrained nucleotides would be the same between the two CRE sequences. However, since the pleiotropic CRE is used in multiple contexts, loss of its activity would be associated with a higher fitness cost compared to the more specific CRE. Thus, the degree of constraint at any individual nucleotide affecting regulatory activity is expected to be higher for the pleiotropic CRE. This hypothesis is supported by the association between pleiotropy and evolutionary conservation previously observed for protein-coding genes (*172*). Even though the entire coding region of every protein-coding gene is under some level of constraint (with the exception of synonymous changes), this constraint is significantly higher for pleiotropic genes compared to those with organ-specific expression (*173*). With future studies dissecting the grammar of more regulatory sequences, either by mutational scanning (*285*, *287*, *288*) or through the *in silico* identification of major sequence features (*101*, *289*), it should soon be possible to assess the relative contributions of these different models to the increased evolutionary conservation of pleiotropic CREs.

4.2 Periods of greater developmental change

The analysis of dense temporal time series across organs and cell types also allowed me to assess the degree of change in the activity of non-coding elements during development. Previous analyses of protein-coding gene expression revealed that specific periods of organ development are marked by the extensive rewiring of gene expression programs (*172*). The first period occurs early in embryonic development (e.g., E11-E12 in mouse) with the activation of organ-specific developmental programs (*172*). The second one, occurring around or shortly after birth (e.g., P3-P14 in mouse) is marked by the expression of genes involved in organ specific functions (e.g., synaptic genes in the brain and metabolic enzymes in the liver) (*172*). My analyses of lncRNA expression showed a similar pattern to that observed for protein-coding genes. When analyzing the developmental dynamics of CRE activity within cerebellar cell types, I also identified strong temporal changes in chromatin accessibility that occur in specific periods of development. For example, focusing on granule cell progenitors, I observed that prenatal (E13-P0) populations overall looked very similar to each other but showed significant differences from early postnatal populations (P4-P7). These observations suggest that at least part of the developmental rewiring observed at the whole-organ level is not due to changes in the cellular composition of the organ, but rather because of temporal changes in the molecular profiles of defined cell types.

More intriguingly, some of these temporal changes appear to be shared between different cell types. For example, my analyses of embryonic cerebellar progenitor cells revealed concordant shifts in CRE activity that are shared across all germinal zones. This suggests that the generation of a diverse set of cell types during early cerebellar development is achieved by a combination of spatial and temporal signals, supporting a model of cell fate induction through common temporal cues. Such cues could include extrinsic signals, such as factors secreted from the choroid plexus (*205*), or intrinsic shifts in the activity of transcriptional regulators. The latter model is supported by the recent discovery of a temporal code that is shared across multiple regions of the central nervous system (*255*). NFI transcription factors feature prominently in this temporal patterning system (*255*). Notably, NFI motifs are enriched amongst CREs with increasing accessibility in late development in both multipotent cerebellar progenitors and in fate-committed granule cell progenitors. Another temporal regulator highlighted by my analyses was *Lin28a*, which shows a sharp decrease in early stages of cerebellar development. A recent whole-embryo scRNA-seq atlas of human organogenesis revealed that *LIN28A* is highly expressed in progenitor cells across multiple organs and that its expression is rapidly downregulated around the same developmental period (CS15-16 in human, E11 in mouse) in 96% of the profiled cell types (*290*). Furthermore, the same study showed that this temporal expression pattern has been conserved throughout vertebrate evolution (*290*). These recent studies highlight the importance of previously unappreciated cell type-shared temporal changes during development. With current studies having profiled only a limited subset of all major developmental stages across all organs, there might be more such temporally-regulated genes waiting to be discovered.

4.3 Evolutionary turnover of non-coding elements

Both lncRNA (*47–49*, *77*) and CRE (*46*, *117*, *120*) sequences are known to diverge much faster during evolution compared to protein-coding genes, a notion also supported by my analyses. However, perhaps unintuitively, I also showed that a considerable fraction of those non-coding elements that can be detected in multiple mammalian species retain similar spatiotemporal activity patterns. In the case of lncRNAs, transcripts with dynamic expression in both human and mouse show comparable conservation of their expression profiles to that observed for protein-coding genes. In the context of CREs, corresponding cell types between mouse and opossum show the highest similarity in the chromatin accessibility profiles of orthologous intergenic CREs. Perhaps the rapid evolutionary turnover of the sequences of these non-coding elements also hides the explanation for the comparably slower change in their spatiotemporal activity patterns. Given how easy it is for a new non-coding element to emerge from previously inactive DNA, the need for repurposing of old genomic regions is relatively limited, and likely reserved for pleiotropic elements that contain more dense and complex sequence motifs (see 4.1). Nevertheless, there is also ample evidence that such repurposing does eventually occur during evolution (*120*). This is also consistent with my analyses, in which two of the three CREs I investigated around *PIEZO2* are predicted to be accessible in cerebellar progenitors only within a subset of species in which their sequence can be detected (see 3.4.6). The most likely explanation for this observation is that the ancestral CRE sequence was or still is active in a different cell type and developmental stage, and has been repurposed to (also) be accessible in cerebellar progenitors in the primate/human lineages.

In contrast to the fast turnover of individual CRE sequences, my analyses showed that the sequence rules underlying cell type-specific CRE activity are overall conserved across mammals, at least in the context of the cerebellum. This was highlighted by both the similar TF motif enrichments observed amongst human and mouse CREs with matching patterns of spatiotemporal activity, and by the similar performance of human and mouse deep learning models in predicting CRE accessibility patterns for either human or mouse sequences. Taken together, these findings support a model of CRE evolution in which the main TFs are conserved across species, dictating the same sequence requirements for CRE activity in corresponding cell types. Since these requirements typically involve short motifs, such sequences can easily be created *de novo*, in sharp contrast to how most protein-coding sequences emerge (*15*). The ease in generating new cell type-specific enhancers was recently demonstrated in yeast (*291*) and fly (*289*), where only 5-15 mutations were enough to transform a previously inactive DNA sequence to a highly active regulatory element. The ability to quickly generate new CREs relaxes the constraint on individual elements, allowing for the rapid gain and subsequent loss of CREs with similar sequence properties. Thus, it has been suggested that selection mainly constraints the overall regulatory input around a gene rather than individual TF binding sites and CREs (*117*, *124*). Finally, the rapid incorporation of new CREs into existing gene regulatory networks limits the need for repurposing old CREs, leading to a relatively high conservation of spatiotemporal activities amongst (the few remaining) orthologous regulatory elements. Possible exceptions include CREs that are active in multiple cell types, which have higher sequence complexity and are thus more difficult to create *de novo*. In support of this, pleiotropic CREs are overall more likely to be preserved in mammalian genomes during evolution, as discussed above (4.1).

Despite the overall fast evolution of non-coding sequences, there are differences in constraint between organs, cell types and developmental stages, reflecting potential differences in selective pressures. For lncRNAs, I observed a decrease in evolutionary conservation during the development of all assayed organs. Similarly, focusing on chromatin accessibility of CREs in the mouse cerebellum I showed that intergenic CREs active in earlier developmental stages are more conserved than those active in later development. The single-cell resolution of my chromatin accessibility data further allowed me to identify that this developmental decrease in conservation was shared across all cerebellar cell types, suggesting that the whole-organ patterns are mostly explained by changes within cell types rather than by differences in cell type abundances. In both cases, my analyses showed that this developmental decrease in evolutionary conservation is associated with a parallel decrease in the fraction of pleiotropic CREs, which are more conserved (see section 4.1). Comparing constraint across different organs, nervous tissues show the highest conservation for both lncRNAs and CREs, in line with previous observations for protein-coding genes (*169*, *172*). On the other end, organs and cell types associated with increased levels of evolutionary innovation, such as the liver, the gonads and immune cell types are marked by faster diverging non-coding elements. These observations suggest that differences in constraint that are associated with cell type- and organ-specific functions shape the non-coding genome in similar ways as for protein-coding gene sequences.

4.4 Non-coding elements and evolutionary innovation

Associating individual non-coding elements with evolutionary innovation is considerably more challenging than identifying conserved elements. The rapid evolutionary turnover of CREs, in particular in cases where the gain of a new element compensates for the loss of an old one, suggests that most CRE changes are effectively neutral in regard to gene expression. Even when focusing on genes that have changed their expression during evolution, identifying the associated CREs can be challenging because they might be thousands of base pairs away. Furthermore, our incomplete understanding of how non-coding sequences relate to their functions prohibits the use of methods commonly employed to identify positive selection in protein-coding sequences (*292*). Previous sequence-based approaches to identify non-coding elements associated with evolutionary innovation have focused on the accelerated evolution (*293*) or complete loss (*294*) of evolutionarily conserved elements in the human lineage. Although many of these elements were indeed later shown to regulate gene expression (*295–297*), such approaches are likely to prioritize changes associated with loss of activity in humans and *a priori* dismiss a very large number of evolutionary young non-coding elements. Other approaches aiming to identify promoters with an unusually high number of substitutions (*298*) have been met with skepticism due to known biases in overall mutation rates between genomic regions and classes of regulatory elements (*299*). Collectively, methods relying on DNA sequence alone can only offer limited insights regarding the contributions of non-coding elements to evolutionary innovation.

In this dissertation, I tried to tackle these challenges by incorporating data on the spatiotemporal activity of non-coding elements. In the case of the marsupial-specific lncRNA *FSX*, its broad and female-specific expression and localization on the X chromosome provided a clear hypothesis regarding its putative function. This allowed me to test this hypothesis with targeted analyses and ultimately accumulate substantial evidence to support its involvement in marsupial X chromosome inactivation. For CREs, I devised a strategy to enrich for elements associated with evolutionary innovation by focusing on genes with divergent expression patterns between human and mouse and identifying adjacent CREs with spatiotemporal activities matching those of the gene expression change. I further utilized deep learning models to predict CRE accessibility patterns across hundreds of mammalian species in order to better understand their evolutionary histories. Finally, by considering the sequence features with high importance for the model's predictions I was able to propose specific sequence changes as putatively causal for the observed shifts in CRE activity. Although further refinements will be needed (discussed in detail in the next section), this framework offers a new way for investigating CREs and their contributions to evolutionary innovation.

4.5 Outlook

The final sections of this dissertation described a new approach for the identification and characterization of CREs that are associated with gene expression changes. However, this work was focused on individual cases, mainly intended as proof-of-principle analyses. There are several improvements already planned for these analyses, some of which being implemented by collaborating lab members at the time of writing.

4.5.1 Incorporation of additional species

One major limitation of the method I used to identify differentially expressed genes between human and mouse is that it only considered these two species. Although I was able to confidently detect genes with higher expression in human or mouse for a given cell type and developmental window, I was unable to polarize the changes in the absence of an outgroup species. Therefore, genes with higher expression in human include both those that gained expression in the human lineage and those that lost expression in the mouse lineage. These two gene sets are expected to differ markedly with respect to the evolutionary histories and activities of their adjacent CREs. In the first case, one would expect an enrichment for evolutionarily young CREs in human, whereas in the latter case we expect an increased loss of older CREs that were already present in the ancestor of human and mouse. To distinguish between these two possibilities, I plan to incorporate a previously generated snRNA-seq atlas of the developing cerebellum from the marsupial opossum (*179*), which can provide information about the ancestral gene expression profiles and thus allow me to polarize the gene expression changes.

Additionally, the identified expression differences between human and mouse include all shifts that occurred during the last 90 million years, ranging from changes shared across all primates to those specific to modern humans. To better resolve the timing of these gene expression changes, I further plan to incorporate a new dataset describing cerebellum development in the common marmoset, a New-World monkey that diverged from human about 40 million years ago. This dataset, currently being finalized by Dr. Mari Sepp, will contain paired measurements of gene expression and chromatin accessibility from the same cell across more than 100,000 cells spanning seven stages of cerebellum development. Since the developmental sampling of this dataset is sparser compared to human and mouse, I plan to incorporate it in my analyses in a targeted way. Thus, I will only update the information on human gains of expression when the corresponding cell type and developmental stage was adequately sampled in marmoset.

Since the marmoset dataset includes single-cell measurements of chromatin accessibility, I plan to use this dataset to also evaluate the prediction accuracy of my approach for estimating the evolutionary age of CREs based on their predicted accessibility across the entire mammalian tree. Even though I previously showed that deep learning models trained on a single species can be used to predict accessibility in a different mammalian species, inferring the timing of a gain or loss in CRE activity is a more ambitious task, as it aggregates prediction errors from hundreds of species. Thus, this dataset will allow me to assess what fraction of CREs that are accessible in human but not in mouse, and have high predictions across primate species, are in fact accessible in marmoset. Similarly, CREs with high predictions only within great apes or Old-World monkeys should overall show low accessibility in the marmoset. Furthermore, the lab has been able to obtain a limited number of postnatal cerebellar samples from non-human great apes, which might allow us to generate additional smaller datasets to perform similar assessments for CREs with more recent gains of accessibility.

4.5.2 Phylogenetic models based on predicted CRE accessibility

Besides incorporating new datasets, there are also additional analyses that can be performed based on the existing data. The case studies of CREs that I described in this work were based on the manual investigation of their predicted accessibility across the mammalian phylogeny. The identification of putative TF binding sites and their conservation across primates was also based on the inspection of the model's feature importance and of multiple sequence alignments. However, to globally assess the evolutionary history of thousands of CREs, it is important to scale up such analyses by automatically extracting the relevant features and sequences for each CRE and to incorporate statistical measurements of uncertainty for the reported inferences. To this end, Tetsuya Yamada and I are currently working towards obtaining statistical estimates for the assignment of CREs to different evolutionary groups, based on the detection of non-randomly distributed patterns across the phylogenetic tree. Additionally, inspired by recent studies (*188, 300*), we are exploring the possibility to detect stabilizing and positive selection by comparing the difference in predicted accessibility between two species (one of which could be the inferred ancestor) versus an empirical distribution derived from simulated sequences with the same amount of nucleotide divergence randomly distributed across the CRE. In such an approach, CREs under stabilizing selection are expected to show significantly smaller differences in predicted accessibility than randomly placed mutations because the parts of the sequence that are important for the CRE activity (i.e., TF motifs) have been protected from substitutions during evolution. On the other hand, CREs under positive selection (typically detected based on comparisons to an ancestral sequence) will show significantly larger differences than the empirical distributions because mutations have accumulated in the parts that are important for the high prediction of the sequence in one lineage (e.g., by creating a new TF binding site).

4.5.3 The evolution of gene regulatory networks

Another limitation of the approach I described so far is that it focuses entirely on *cis*-regulatory mechanisms. There is ample evidence to suggest that most evolutionary changes in gene expression are due to changes in CREs rather than in the activity of TFs (*46*, *117*, *120*, *123*, *124*, *167*, *301*) and my analyses of regulatory grammar conservation in cerebellar cell types further support this notion. However, these analyses examine the activity of relatively few "cell type-defining" TFs, which are responsible for the regulation of most CREs accessible in each cell type. There are more than 1,500 TFs encoded in the human genome (*43*), including at least 150 TFs with zinc-finger binding domains that emerged from primate-specific expansions and are thus not present in mouse (*302*). Although representing obvious candidates for driving evolutionary innovation, their study has been hindered by our limited knowledge of their DNA-binding preferences (*302*). Besides lineage-specific TFs, our work on the evolution of gene expression in cerebellar cell types revealed 89 TF genes with major expression shifts between human and mouse (*179*). While the overall contribution of these species-specific or expression-diverged TFs to the global chromatin accessibility landscape of each cell type is likely small, they might explain a disproportionate amount of gene expression changes between species.

To investigate the effect of changes in the *trans*-regulatory environment on the evolution of gene expression, Philipp Schäfer, a master's student in the Kaessmann lab, has been working with me on the inference of cell type-specific gene regulatory networks. Several methods have recently been developed to facilitate the linkage of TFs to their candidate target genes via CREs that are bound by these TFs and are predicted to regulate the target gene (*303*–*306*). While the methods vary in their assumptions and statistical procedures, they all make use of paired measurements of gene expression and chromatin accessibility from single-cells. Even though our human and mouse snRNA-seq and snATAC-seq datasets were acquired separately for each modality, my analyses demonstrated the feasibility of the *in silico* integration of the datasets. Using these integrated datasets, Philipp Schäfer has been working, under my supervision, towards identifying the best performing methods and parameters by evaluating the inferred gene regulatory networks based on independent molecular measurements. Specifically, he has used TF ChIP-seq data (*307*) to evaluate the ability of the different approaches to link TFs to CREs, and chromatin interaction data (*115*) to validate inferred CRE-to-gene links. Additionally, he has been assessing the quality of TF-to-gene connections by using the gene regulatory network to predict changes in gene expression between independent RNA-seq samples that were not used for inferring the network, akin to previous approaches (*304*, *305*). At the moment of writing this dissertation, this benchmarking is nearly complete and Mr. Schäfer will soon finalize his gene regulatory network inference in human and mouse and proceed to evolutionary comparisons between the two species. This will allow us to directly assess the impact of *cis* versus *trans* regulatory changes on gene expression evolution, most likely validating the higher contribution of the former, but also identifying cases where species-specific or expression-diverged TFs lead to downstream changes in gene expression. These cases can be further explored by simulating perturbations in TF expression (e.g., by *in silico* humanizing TF expression patterns in mouse cells) and assessing the impact on downstream gene expression states (e.g., observing a potential shift of mouse cells towards higher expression similarity with the corresponding human cell type).

4.5.4 Gene-centric estimates of regulatory input

In previous sections, I discussed how the rapid evolutionary turnover in CRE sequences, with frequent losses of CREs compensated by gains of new CREs with similar sequence features, complicates the identification of the relatively few regulatory changes that have a sizeable impact on gene expression evolution. It has long been hypothesized that evolutionary shifts in gene expression are ultimately driven by changes in the total regulatory input (approximated by the number of connections per TF) received by each gene. Thus, gains and losses of CREs that lead to the same number of functional TF binding sites around a gene are predicted to be effectively neutral whereas those that lead to the emergence of new connections or to the loss of old ones are more likely to be associated with changes in gene expression. However, such hypotheses have been difficult to test due to the challenges in identifying TF binding sites and linking CREs to their target genes. The new computational tools discussed here — in particular deep learning models of regulatory grammar that allow for a better inference of putatively functional TF binding sites (*101, 103*), and gene regulatory networks that can be used to quantify the TF input received by each gene — offer new paths to tackle this challenge. Specifically, I anticipate that it will soon be feasible to use our datasets to infer the total input of TF connections received for each gene and then test whether genes with significant differences in gene expression between human and mouse also show more differences in their regulatory input. However, such analyses can be complicated by additional factors affecting gene expression, such as the genomic distance between a CRE and its target gene (*308*), quantitative differences in TF expression levels leading to differential binding on CREs (*309*), as well as the presence of non-linear interactions between CREs (*146*), TFs (*267*) and cofactors (*310*).

4.5.5 Experimental validation of computational predictions

The previous sections describe computational methods to investigate the regulatory basis of gene expression changes. However, even the best performing methods are only able to give predictions, which need to be validated experimentally. Furthermore, all my analyses of CREs are based on measurements of chromatin accessibility, serving as a proxy for regulatory activity. Despite being one of the most widely used methods to enrich for CREs, it is important to consider that not all accessible regions correspond to CREs. Moreover, changes in chromatin accessibility might differ in their temporal dynamics compared to changes in CRE activity. Additionally, while the assignment of CREs to their putative target genes based on matching spatiotemporal activities is a reasonable approach to enrich for functionally relevant interactions, it can also lead to false assignments, especially when multiple genes in the same region show similar expression patterns. Finally, while improving on previously available methods, deep learning models can make erroneous predictions regarding the accessibility of CRE sequences and can mistakenly highlight sequence segments that are not actually bound by TFs. By focusing on the classification of CRE accessibility patterns, such models may also fail to capture features that are globally important for CRE activity but don't lead to differences in accessibility between cell types, such as regions bound by constitutively active TFs or by factors modulating quantitative expression levels without altering the context-specificity of the CRE. Given all these limitations, it is important to validate — at least a subset of — these computational predictions based on independent functional assays.

The most widely used method to assess the ability of DNA sequences to activate gene expression is through enhancer reporter assays, in which the examined sequence is placed in the proximity of a reporter gene transcribed from a minimal promoter (*311*). In the original low-throughput implementation of these methods, reporter gene expression is usually monitored based on visual inspection (e.g., by using luciferase, LacZ staining or fluorescent proteins) (*311*). By combining *en masse* cloning with barcodes that are unique to each tested element to monitor the expression of the reporter gene through RNA-sequencing, massively parallel reporter assays (MPRAs) allow the parallel investigation of thousands of putative CREs in a single experiment (*312*). Such methods have been successfully used to study the effect of sequence changes on CRE activity in the context of disease-associated variants (*313*) and during evolution (*296*). Thanks to the high conservation of the regulatory grammar of cerebellar cell types across mammals, it would be feasible to use mice to test a selection of human CREs, together with their ancestral states and their orthologs in other mammalian species. These experiments would allow the testing of my computational predictions regarding the impact of individual sequence changes on CRE activity.

A major limitation of enhancer reporter assays is that they typically test CRE activity outside of their native genomic context (typically in episomal vectors or through lentiviral integration into “safe-harbor” genomic loci) (*312*). Thus, they can only test the ability of CREs to “generally” activate transcription in a cell type of interest, without considering the effect of the local genomic and chromatin context or of potential specificities in enhancer-promoter interactions. An alternative approach that can address these caveats is to directly modify the endogenous DNA through the CRISPR/Cas9 system, for example by “humanizing” mice for a set of CREs (*314*). For CRE sequences that are already present in mouse, this could involve the introduction of mutations to model the orthologous sequences in human, other primates or to reconstruct intermediate ancestral states (*288*). Alternatively, CREs specific to the human lineage can be inserted into their syntenic position in mouse. Profiling of chromatin accessibility and gene expression in these transgenic mice can subsequently assess the effects of these mutations on the accessibility of the CRE and on the expression of its putative target gene. Thus, despite being more limited in throughput, this approach can be more informative by simultaneously testing multiple aspects of my computational predictions, such as the effect of sequence changes on CRE activity and the association of that CRE with the correct target gene.

Collectively, while the computational analyses presented in the last sections of my dissertation are likely to enrich for CREs that have contributed to the evolution of human-specific biology, it is important to functionally investigate at least a subset of these CREs based on independent methods. This will provide more reliable estimates for the accuracy of my computational predictions and will also reveal individual high-confidence cases that can be investigated further (for example by assessing the effect of validated CRE changes on cerebellar morphology and physiology). Owing to the rapid development of both computational and experimental methods, studies of mammalian evolution that would have been considered unfeasible less than a decade ago, are now within our grasp.

5. Methods

5.1 Analysis of lncRNA expression in mammalian organ development

LncRNA annotation and expression quantification – These steps were performed by Dr. Ray Marin and are described in detail in *Sarropoulos et al. 2019* (*223*). In brief, Dr. Marin used stringtie (1.2.3) (*315*) to assemble transcripts based on RNA-seq data for each sample (species, organ and developmental stage). He then merged transcripts into a single assembly per species using Cufflinks (2.2.1) (*316*). He removed all genes overlapping coding genes or showing evidence for coding potential, which was estimated by three different methods, CPAT (1.2) (*317*), RNAcode (0.3) (*318*) and similarity to known proteins, as determined by blastx (2.4.0)(*319*). Orthologous lncRNA families were determined by identifying signficant sequence similarity between lncRNA annotations of different species based on blastn (2.4.0) (*319*). OrthoMCL (2.0) (*320*) was used to cluster reciprocal best hits into lncRNA families.

To evaluate this approach, I considered the conservation of synteny across species. To this end, for each lncRNA, I identified the closest upstream and downstream protein-coding gene using bedtools (2.25) (*321*) and estimated the fraction of lncRNAs that had at least one conserved neighbor in the same orientation. Additionally, for each age class, I estimated the fraction of species in the dataset in which a lncRNA could be identified and compared this to previous studies (*77*). Gene expression counts were generated based on uniquely mapped reads using HTSeq (0.6.1) (*322*) and were subsequently normalized for different analyses into counts per million (CPM), reads per kilobase of transcript per million mapped reads (RPKM), or variance stabilized counts (VST) as implemented in DESeq2 (1.12.4) (*323*). Tissue and time-specificity indexes were calculated based on the Tau statistic, as previously described (*324*). Developmentally dynamic protein-coding genes and lncRNAs were identified for each organ using masigPro (*226*) on CPM-normalized counts and requiring a goodness-of-fit (R^2) value greater than 0.3.

Genomic classification and comparison to other datasets – To compare lncRNA annotations with other datasets (e.g., transcribed enhancers (*229*), pcRNAs (*233*)), I used bedtools intersect (2.25) (*321*) on exonic coordinates in a strand-specific manner. To classify lncRNAs based on their genomic context, I used FEELnc (1.0) (*325*) to compare them to Ensembl protein-coding gene annotations with a maximum window extension of 100 kb. LncRNAs located more than 100 kb away from the nearest protein-coding gene were labelled as "isolated intergenic". The set of experimentally validated lncRNAs was acquired from lncRNAdb (2.0) (*230*). To integrate with the CRISPRi screen, I intersected the primary TSS from the screen library (extended by 500 bp to each side) with the first exon of each lncRNA in my annotation.

Expression-matched sets of dynamic and non-dynamic lncRNAs – To control for the effect of differences in maximum expression levels between dynamic and non-dynamic lncRNAs, I generated two sets of expression-matched transcripts. For the first set, for each dynamic lncRNA, I identified the non-dynamic lncRNA with the most similar maximum expression level. After eliminating redundancies, I was able to obtain around 3,000 transcripts from each set with nearly identical distributions of maximum expression. To assess the enrichment of lowly expressed dynamic lncRNAs for functionally relevant features, I generated a second set focused on 798 dynamic lncRNAs with maximum expression values in the range of 0.25-0.75 RPRKM. As for the first set, for each of these lncRNAs, I identified the non-dynamic lncRNA with the most similar maximum expression value.

Conservation of lncRNA expression – I estimated expression similarity across species by calculating the Spearman's correlation coefficient between the expression profiles of orthologous lncRNAs across corresponding organs and developmental stages. To assess the effect of evolutionary age on expression constraint, I stratified mouse lncRNAs conserved in rat based on their predicted age and compared the distribution of correlation coefficients between the expression profiles of mouse and rat orthologs. To compare the degree of lncRNA conservation across organs and developmental stages, I estimated the fraction of conserved (older than 80 million years) lncRNAs expressed (RPKM > 1) in each sample. Additionally, for each pair of corresponding samples (organ and developmental stage) between human and mouse, I estimated Spearman's correlations in the expression ranks of all lncRNA orthologs of a given set (e.g., developmentally dynamic in both species).

Regulatory complexity of lncRNAs – I estimated the regulatory complexity of different lncRNA classes based on the number of distinct TFs binding on each lncRNA promoter (defined as -2,000/+1,000 bp from the first exon of the longest isoform). TF binding sites were identified based on publicly available ChIP-seq experiments and were retrieved from GTRD (*227*). I also estimated TF binding in a set of random non-repetitive intergenic regions of matched length (3,000 bp), which I used as a negative control. To assess the relevance of the increased regulatory complexity of dynamic lncRNA promoters in the context of mammalian organ development, for each TF and set of lncRNAs dynamic in each organ, I estimated the fraction of promoters that overlap the binding sites of the TF (termed "TF binding frequency"). I then identified TFs with highly variable binding frequency across organs and compared their binding patterns to their expression profiles.

Developmental expression of lncRNAs – I identified lncRNAs and protein-coding genes with significant expression across adjacent developmental stages using DESeq2 (1.12.4) (*323*) requiring an absolute $\log_2$ fold-change ≥ 0.5 and an adjusted *P*-value < 0.05. To co-cluster lncRNAs and protein-coding genes, I selected dynamic transcripts for each organ and used their VST counts (median across replicates) as input for GPClust, a method to cluster time-series data based on Gaussian processes (*326–328*). I used a noise variance parameter (k2.variance.fix) of 1.0 for mouse and 1.5 for human and otherwise default settings. Gene ontology enrichments for each cluster were identified using WebGestaltR (0.1.1) (*329*).

Co-expression with adjacent coding genes – I used bedtools closest (2.25) (*321*) to assign each lncRNA to its nearest protein-coding gene (lncRNA-mRNA), then used the latter's closest protein-coding gene as a control (mRNA-mRNA). For each lncRNA-mRNA and mRNA-mRNA pair, I estimated Pearson's correlation coefficients between their expression profiles (median VST counts across biological replicates) across all organs and developmental stages, except for sexually mature testis samples. To assess the global extent of lncRNA-mRNA co-expression I compared the distribution of correlation coefficients to those of the mRNA-mRNA controls across the entire dataset and within specific distance ranges. To identify significantly co-expressed lncRNA-mRNA genes, I compared their correlation coefficients to those of the mRNA-mRNA control (applying the Fisher Z-transformation and using the function paired.r() from the R package psych (1.8.4) to obtain *P*-values). Additionally, I required correlation coefficients to be greater than 0.75. This value was determined based on comparisons of paralogous protein-coding gene pairs, which I found to retain significantly higher expression similarity than non-related adjacent mRNA-mRNA pairs.

5.2 X-chromosome inactivation in opossum

Screening for female-specific lncRNAs on the opossum X chromosome – To search for female-specific genes on the opossum X chromosome I examined sex-bias expression scores that were developed by Svetlana Ovchinnikova and Leticia Rodríguez-Montes *(Rodríguez-Montes et al. in prep)*. Briefly, these scores were calculated separately for each organ and estimate the extent of sex-biased expression of a gene within that organ by comparing smoothened developmental expression profiles between male and female samples. This score considers both the mean and maximum difference between sexes along the developmental trajectory, thus prioritizing genes with broader sex-bias during development. Positive and negative scores indicate male- and female-specific expression, respectively. To additionally filter for genes with consistent sex-biased expression across multiple organs, I calculated the mean sex-bias score across all somatic organs in the dataset. I used the R package karyoploteR (*330*) to visualize the genomic positions of all female-specific genes on the opossum chromosome X. Gene densities for the karyotype plot were estimated separately for lncRNAs and protein-coding genes using the function kpPlotDensity() with a window size of 100 kb.

Onset of female-specific expression – To investigate the expression of *FSX* at the onset of marsupial X chromosome inactivation, I used a previously published scRNA-seq dataset, which is based on full-length RNA sequencing (SMART-Seq v.4), with each cell sequenced as a separate library (*241*). Since *FSX* was not included in the genome annotation used in that study, I retrieved raw sequencing data and cell type annotations (sex, developmental stage) from ArrayExpress (E-MTAB-7515) and aligned the RNA-seq data to the opossum genome (monDom5) using STAR (2.7.1a) (*331*). For each cell, I used featureCounts (*332*) to count reads in genes based on my extended annotation of the opossum transcriptome (*223*). Only exonic reads (-t exon) with a minimum mapping quality of 40 (-Q 40) were counted in a strand-specific manner (-s 1). Counts for each cell were combined in a gene-by-cell matrix and normalized for gene length and sequencing depth by calculating RPKM values. I then summarized expression profiles of sex-biased genes in single cells by developmental stage and sex to identify the onset of female-specific expression for *RSX* and *FSX*.

Sequence features of RSX and FSX – Sequence similarity between *RSX* and *FSX*, as well as within their own sequences to facilitate the detection of sequence repetitions, was visualized using EMBOSS Dotmatcher (*333*) with a window of 10 bp and a similarity cutoff of 50. To estimate the fraction of each lncRNA transcript covered by simple repeats, I first collapsed all exons from each gene into a consensus metagene model which covered all exonic regions found across all isoforms. I retrieved the "Simple repeat" track for the opossum genome from the UCSC table browser and used bedtools (2.29) (*321*) to intersect it with the merged exonic regions of opossum lncRNAs. I then divided, for each lncRNA, the length of the overlap with repeats by the total length of the merged exonic regions. To identify overrepresented sequence motifs for *RSX* and *FSX,* I used MEME (Multiple EM for motif elicitation) on the sequence of the merged exonic regions of each lncRNA with a maximum motif width of 50 bp and a minimum of 10 occurrences per motif (*334*).

Evolutionary conservation of RSX and FSX sequence and expression – To assess the conservation of opossum lncRNA sequences in other mammalian genomes, I first extracted the DNA sequences corresponding to the merged exonic regions across all isoforms of each lncRNA. I excluded regions overlapping with protein-coding gene exons (in any orientation) and used blastn (*319*) to search for sequence similarity to other genomes, after applying soft-masking for repeats. Significant alignments were required to be at least 50 bp long, show at least 10% of sequence identity and have an E-value smaller than 10^{-5}. The relative conservation of *RSX* and *FSX* transcript segments across mammalian genomes was visualized using the R package Gviz (*335*).

I also used Gviz (*335*) to visualize the genomic neighborhood around the regions with significant similarity to *RSX* and *FSX* in other marsupial genomes. To filter out smaller spurious alignments, I determined, for each species, the segment with the highest alignment score (bitscore) and only visualized alignments within 500 kb (*RSX*) and 100 kb (*FSX*) of that segment. I used a smaller window for *FSX* due to the higher gene density in that region compared to the *RSX* locus. To aid comparisons across species, I maintained a consistent orientation by reversing the entire genomic region for alignments to the opposite strand compared to the opossum genome.

To assess the female-specific expression of *RSX* and *FSX* in koala, I downloaded raw sequencing data from a previous study (*243*), aligned them to the koala genome (phaCin_unsw_v4.1) and used featureCounts (*332*) to quantify reads in gene exons annotated in the NCBI annotation, and separately in regions with significant sequence similarity to the *RSX* and *FSX* opossum exonic sequences. I then used the counts in the NCBI annotation to determine the sequencing depth of each library, which I used for RPKM normalization of the *RSX*- and *FSX*-aligned regions across samples from different individuals and tissues.

5.3 Chromatin accessibility dynamics in mouse cerebellum development

Data processing, clustering and cell type annotation – Raw sequencing data were converted to tabular fragment files using celllranger atac (1.1.0) (*66*). Quality control, doublet removal and inference of gene scores was performed using ArchR (0.9) (*245*) as described in detail in *Sarropoulos et al. 2021* (*244*). For dimensionality reduction, I used an iterative latent semantic indexing (LSI) procedure with gradually increasing clustering resolution (0.1, 0.2, 0.4, 0.8) to project the data into 100 dimensions. I then performed Louvain clustering on these components (resolution 1.5), identifying a total of 47 clusters. Clusters that could be confidently matched to a single cell type and state were annotated as such, whereas the rest were subjected to subclustering analyses, as described in detail in *Sarropoulos et al. 2021* (*244*). In total, this approach allowed the confident annotation of 97% of cells in the mouse dataset.

Integration with scRNA-seq data – I integrated the mouse snATAC-seq dataset with a scRNA-seq study profiling the development of the mouse hindbrain and cerebellum (*210*) using canonical correlation analysis (CCA) between gene expression (scRNA-seq) and gene scores (snATAC-seq) as implemented in Seurat (3.1) (*249*). To improve the computational efficiency and accuracy of the integration procedure, I reanalyzed the scRNA-seq and snATAC-seq data in a stage-wise manner using Seurat (*249*) and ArchR (*245*), respectively. This allowed me to use the CCA procedure to only integrate cells from corresponding developmental stages. For each stage, the integration was performed based on the 3,000 highly variable genes from the scRNA-seq dataset, which was used as a query to transfer cell type annotations and to impute gene expression estimates to the snATAC-seq dataset, as described in detail in *Sarropoulos et al. 2021* (*244*). After integration, I estimated the concordance in cell type annotations between the two studies based on a Jaccard similarity index for each pair of assigned (ATAC) and predicted (RNA) labels. To assess the utility of gene scores as a proxy for gene expression, I estimated, for each developmental stage and highly variable gene, Pearson's correlations between its gene score and imputed RNA value.

Identification of CREs and assignment to target genes – Open chromatin regions, as a proxy for putative CREs, were identified in a cluster-specific and replicate-aware manner using ArchR (0.9) (*245*), which internally utilizes MACS2 (2.1.2) (*250*), as described in detail in *Sarropoulos et al. 2021* (*244*). ArchR's iterative overlap merging procedure was used to collapse the cluster-specific datasets into a single union peak annotation. To reduce the inclusion of noisy peaks from clusters with large numbers of cells or replicates, I also implemented an additional filtering step, requiring peaks to be accessible in at least 5% of cells in at least one cluster, resulting to the identification of 261,642 putative CREs. Putative CREs were annotated in terms of their genomic context based on their overlap or proximity to genes from the mm10 UCSC annotation, supplemented with the lncRNAs identified in *Sarropoulos et al. 2019* (*223*). I benchmarked this CRE annotation by comparing to other datasets (*62*, *150*, *239*, *252*), estimating the fraction of different element sets recovered in the cerebellum snATAC-seq dataset.

To associate CREs with their putative target genes, I combined two complementary approaches, as described in detail in *Sarropoulos et al. 2021* (*244*). First, I estimated correlations in the accessibility of protein-coding gene promoters and distal (intronic and intergenic) CREs within a 250 kb window upstream and downstream of the promoter. Since the sparsity of single-cell data can hinder correlation-based analyses, I summarized accessibility profiles over 4,083 pseudocells, each aggregated across 50 nearest neighbors in the LSI embedding. The observed correlations were compared to a background distribution generated by correlating each promoter to 10,000 random distal CREs from different chromosomes. I then used a lenient cutoff ($r > 0.15$, FDR < 40%) to select pairs that would be considered in the subsequent analysis. For this second step, I focused on gene scores, which reflect gene expression more accurately than promoter accessibility (*245*). I aggregated gene score and distal CRE accessibility estimates across cells from the same cell type and developmental stage, generating 124 "pseudobulks" comprised by at least 40 cells. After scaling by sequencing depth (CPM) and applying a log-transformation, I computed the Pearson's correlation between gene scores and CRE accessibility for each gene-CRE pair that passed the first filtering step of promoter co-accessibility. Similar to the first step, I generated a background distribution using interchromosomal correlations. Using a strict cutoff ($r > 0.41$, FDR < 5%), and assigning distal CREs that were correlated with multiple genes to the one with the maximum correlation, I identified a total of 32,792 CRE-gene pairs. I evaluated the quality of these assignments by assessing the probability of the CRE and gene to be in the same TAD in neuronal progenitors (*115*) and by estimating the correlation in eRNA and mRNA expression in a series of mouse cerebellar samples (*252*).

CRE activity across cell types and stages – To summarize the major patterns of CRE activity during mouse cerebellum development, I aggregated CRE accessibility profiles across all cells from a given cell type and developmental stage, excluding samples from cell type mixtures (e.g., nuclear transitory zone) and non-neural cell types (e.g., vascular) and only considering groups with at least 50 cells. I scaled the data by sequencing depth (CPM) and then standardized accessibility profiles per CRE by scaling by its maximum CPM value across all cell types and developmental stages. I then used a two-step clustering procedure inspired by *Trevino et al. 2020* (*336*). First, I used k-means clustering to identify 50 clusters of CRE activity (primary clusters). Then, I estimated the average accessibility across cell types and developmental stages for each primary cluster and used it for hierarchical clustering (based on correlation distances). I then used the clustering dendrogram to iteratively merge the two most similar branches into a new cluster and compute a silhouette score for the new clustering. Based on these silhouette scores and the overall structure of the hierarchical clustering dendrogram, I determined the optimal number of final clusters to be 26. A similar approach was used for the CRE clusters in cerebellar progenitors (see below) where the optimal number of clusters was found to be 12. I further assessed the membership of each CRE by calculating the Pearson's correlation between its accessibility and the cluster mean. CREs with $r > 0.5$ were considered "confident cluster members" and used for the identification of enriched gene ontology terms (*337*) and TF motifs (*338*).

Analysis of cerebellar progenitors – The annotation of cerebellar progenitor subtypes was based on iterative subclustering and comparison to public databases (*248*), as described in detail in *Sarropoulos et al. 2021* (*244*). To assess potential lineage relationships between temporally distinct progenitor populations, I identified CREs that are specific to the late progenitor group using ArchR's function getMarkerFeatures(). I then focused on earlier developmental stages and estimated, per cell, the fraction of fragments overlapping the CREs that are specific to the late progenitor population. Progenitor types with a higher accessibility in the CREs that are specific to a given population in a later developmental stage were considered more likely to belong to the same lineage.

To validate the clustering of early progenitor cells by developmental stage, which I first observed in a UMAP projection, I aggregated CRE accessibility profiles across progenitor types and developmental stages, estimated Spearman's correlations across progenitor groups and used the estimates for hierarchical clustering. I assessed the robustness of the clustering pattern by bootstrapping with 1,000 repetitions. To identify the CREs driving the observed clustering pattern I repeated the steps described in "*CRE activity across cell types and stages*" with 30 primary k-means clusters and 12 final clusters based on hierarchical clustering.

To assess the effect of temporally-variant CREs on gene expression, I first sought to identify genes with higher gene score variability across developmental stages compared to across cell types. The precise procedure is described in detail in (*244*). In brief, I aggregated gene score profiles across cells from the same progenitor type, developmental stage and replicate, and then estimated standard deviations after grouping by each of these variables. Temporally-variant genes were identified as those with higher standard deviation across developmental stages compared to either across progenitor types or replicates. I then used a fuzzy c-means clustering algorithm for time-series from the R package Mfuzz (2.4.6) (*339*) to classify temporally-variant genes based on their decreasing or increasing activity during development. Next, I examined the gene expression profiles of these genes in cerebellar progenitors from different developmental stages in a published scRNA-seq dataset (*210*), observing a high concordance between gene scores and RNA-seq measurements.

CRE dynamics during differentiation – For these analyses I focused on the three most abundant cerebellar neuron types (granule cells, Purkinje cells and GABAergic interneurons). I extracted cells assigned to each cell type and projected them into a new low-dimensional embedding using non-iterative LSI, as discussed in detail in *Sarropoulos et al. 2021* (*244*). As these embeddings still captured developmental signals that were independent of differentiation (e.g., pre- and postnatal granule cell progenitors were separated), I applied an additional correction using Harmony (1.0) (*258*). I then used diffusion-based pseudotime (*259*), estimated based on a 20-nearest-neighbors graph constructed from the Harmony-corrected embedding, as a proxy for gradual differentiation and maturation processes. To specify the root of the pseudotime, I selected a random cell belonging to the earliest developmental stage and to a cluster with high gene score for genes known to be highly expressed in early precursors of that cell type (e.g., *Atohl* in granule cells, *Ptf1a* in interneurons).

To identify CREs with dynamic accessibility along pseudotime trajectories, I first divided cells for each neuron type into 50 bins based on their pseudotime ranks. I then aggregated chromatin accessibility profiles across all cells in the same bin and used mutual information between CRE accessibility and pseudotime to identify dynamic CREs, as described in detail in *Sarropoulos et al. 2021* (*244*). After identifying CREs with dynamic accessibility in each neuron type, I used Mfuzz (2.4.6) (*339*) to cluster CREs based on their accessibility patterns. For the PCA, I aggregated CRE accessibility profiles across cells from the same neuron type and pseudotime bin, applied a variance stabilizing transformation as implemented in DESeq2 (1.26) (*323*) and performed PCA using the R package FactoMineR (*340*).

5.4 Evolutionary conservation of CREs

Evolutionary conservation metrics per CRE – To assess sequence constraints within each mouse CRE, I downloaded phastCons scores for vertebrates and eutherian mammals from the UCSC table browser and used a sliding window approach to estimate average phastCons scores over 100 bp intervals within each CRE using the UCSC utility bigWigAverageOverBed (*341*). For each CRE, I kept the average score of the most conserved 100 bp-window. To assign a minimum evolutionary age to each CRE, I assessed its presence or absence in the genomes of 16 other vertebrate species based on syntenic alignments with the mouse. To this end, I used liftover (-minMatch=0.1 -multiple -minSizeQ=50 -minSizeT=50) to identify syntenic regions of mouse CREs in each vertebrate species. I then assigned a minimum age to each CRE based on the estimated time of divergence between mouse and the most distant species in which I could detect an alignment. Finally, I used bedtools intersect (2.29) (*321*) to determine overlaps between mouse CREs and various repeats annotated in the RepeatMasker track for the mm10 genome (*342*).

Cell-wise conservation scores – To compare the evolutionary dynamics of CREs across cell types and developmental stages, I calculated mean statistics (phastCons score, minimum age, repeat fraction) per cell based on all intergenic CREs accessible (i.e., at least one fragment detected) in that cell. I focused this analysis on intergenic CREs to minimize biases from the overlap or proximity to protein-coding sequences which overall show very high sequence constraint. Due to the sparsity of snATAC-seq data, average statistics differ even across cells from the same cell type and developmental stage. To summarize estimates across groups of cells, I calculated the mean per biological replicate, observing overall much smaller variance between replicates than between cell types and developmental stages.

To compare the contributions of different age groups and repeat elements to the chromatin accessibility profiles of different cell types, I estimated the fraction of accessible intergenic CREs per cell belonging to each class (i.e., assigned to a specific age group or overlapping a particular repeat class). I then summarized these estimates by calculating the mean across all cells from the same cell type, developmental stage and biological replicate. To compare the conservation of CREs associated with TFs or other genes, I estimated the average constraint (phastCons, minimum age) across all intergenic CREs assigned to each protein-coding gene (see *Identification of CREs and assignment to target genes)*. I then compared the distributions of these estimates between genes annotated as TFs in AnimalTFDB v3 (*343*) and all other protein-coding genes with at least one intergenic CRE assigned to them using two-sided Mann-Whitney *U* tests.

Effects of development, differentiation and pleiotropy on CRE conservation– I used a linear model to assess the relative contributions of differentiation (as inferred by pseudotime) and absolute developmental time (i.e., age of the mouse embryo) on the temporal decrease in CRE conservation. Both pseudotime and developmental stage were used as predictors of the mean CRE constraint (phastCons, minimum age) observed for each cell. The significance of each predictor term was estimated using ANOVA tests between the full model and an alternative model that excluded that term. To validate the presence of age-related differences in a pseudotime-independent manner, I focused on prenatal granule cell progenitors (E13-P0), which clustered together even prior to any correction with Harmony, and are thus very similar to each other. I grouped cells from the same stage and biological replicate into pseudobulks and aggregated CRE accessibility profiles across them. I then used masigPro (1.58.0) (*226*) with a second-degree polynomial on variance-stabilized counts to identify temporally dynamic CREs (adjusted *P*-value < 0.05 and $R^2 > 0.7$). I classified these CREs into those with decreasing or increasing accessibility during development based on Mfuzz (2.46.0) (*339*) with k=2 and compared the sequence constraint between intergenic CREs from the two groups. Differential TF motif enrichment between the two groups was performed using Homer (*338*).

To assess the potential impact of these temporal changes in CRE accessibility on gene expression, I considered the protein-coding genes that were the closest to each of these CREs and examined their expression in granule cell progenitors from different developmental stages in a previously published scRNA-seq atlas of mouse cerebellum development (*210*). I used proximity instead of my previously described correlation-based assignment strategy to avoid the circularity of examining the concordance between gene expression and chromatin accessibility after filtering for correlations between the two features.

To assess the impact of CRE pleiotropy on the observed differences in evolutionary constraint during differentiation, I stratified CREs with dynamic profiles during granule cell differentiation based on their accessibility patterns (determined via clustering, see *CRE dynamics during differentiation*). For each cluster of CRE activity (i.e., accessibility pattern during differentiation), I compared constraint estimates between CREs only dynamic in granule cells or in at least one more cell type.

Comparisons of CRE accessibility between mouse and opossum – The opossum snATAC-seq was processed as described for mouse, with minor modifications detailed in *Sarropoulos et al. 2021* (*244*). Following CRE identification, I detected orthologous CRE pairs based on reciprocal syntenic alignments between mouse and opossum using liftover (-minMatch=0.1 -multiple -minSizeQ=50 -minSizeT=50). Since I used a fixed width of 500 bp for CRE identification, besides reciprocal 1:1 matches, I also considered 1:2 matches when the two hits were up to 500 bp from each other, retaining the CRE with the highest overlap. Additional one-to-many and many-to-many matches were excluded from downstream analyses. To assess the overall conservation of chromatin accessibility between corresponding cell types and developmental stages, I focused on orthologous CRE pairs that are intergenic in both species. For each species, I aggregated accessibility profiles across all cells from the same cell type and state, scaled by sequencing depth (CPM) and estimated rank-based Spearman's correlations across mouse and opossum cell types. To estimate the degree of conservation of individual CRE pairs, I calculated Pearson's correlations of their accessibility profiles in mouse and opossum across all corresponding cell types and stages. I then compared the distribution of correlation coefficients to the same set of CREs with shuffled homology relationships.

To compare the degree of CRE activity conservation across cell types and developmental stages, I estimated the fraction of mouse CREs accessible per cell that have a CRE ortholog in opossum. To additionally compare cell types based on whether CRE activity was conserved in the same cell type and stage, I estimated the fraction of mouse CREs accessible per cell that have a CRE in the 40,000 most accessible CREs for the corresponding cell type and stage of that cell in opossum.

5.5 Evolutionary innovation in CRE activity

Cross-species multiomics atlases – To generate multiomic atlases of cerebellum development in human and mouse, I applied a uniform processing pipeline to previously published (*179*, *244*) and newly generated data. For snRNA-seq data, I used STARsolo (*331*) to count reads in exons and full-length transcripts, allocating multimapping reads based on an expectation-maximization algorithm. I identified barcodes corresponding to cells based on Gaussian mixture models with two groups on the distribution of full-length UMIs and the fraction of intronic reads, taking the intersection of barcodes in the group with the highest distribution for both metrics. I used scrublet to remove the 10% of cells with the highest doublet score from each sample, as well as cells with more than 2.5 times the median number of full-length UMIs in that sample. For each sample, I used Seurat (4.0) (*249*) to regress out cell cycle scores (only correcting for the difference between S and G2M phases), applied SCTransform and projected the data into 50 principal components, which I utilized for low-resolution Louvain clustering (0.2). I then used these clusters as input to SoupX (1.5.2) (*344*) to correct the expression of transcripts associated with ambient RNA or cellular debris. Contamination estimates appeared much higher when considering exonic counts, suggesting that most contaminating transcripts are already spliced. Thus, I only corrected exonic counts, and subsequently reconstructed full-length expression by adding the corrected exonic counts to the uncorrected intronic values. Additionally, I limited the correction to genes estimated to contribute more than 0.05% to the total contamination to avoid introducing noise in the expression of genes for which the background contamination level could not be reliably estimated. I then used the corrected full-length values to repeat the Seurat analysis described above (cell cycle regression, SCTransform, PCA). Next, I integrated samples from the same developmental stage using the Seurat function IntegrateData() based on the corrected SCT counts in the 3,000 most highly variable features across biological replicates. These stage-wise integrated objects were used for integration with the corresponding snATAC-seq samples, as described below.

To annotate cells from the newly added snRNA-seq samples in a consistent way to our previously published annotation (*179*), I used liger (*345*) to integrate all samples within each species. For this, I used a variance threshold of 0.1 and k=75 for mouse and k=100 for human. I then applied a neighbor-voting procedure, coupled with manual curation of clusters with a large fraction of unannotated cells, to annotate new cells in the dataset. In this process, previously annotated cells overwhelmingly received similar labels to their original ones, establishing the reproducibility of our annotation procedure. In the few cases of disagreement, I kept the old labels to maintain maximum compatibility with our previous studies.

For snATAC-seq data, I identified barcodes corresponding to cells based on a similar procedure as described for snRNA-seq cells, applying Gaussian mixture models on the number of fragments and TSS enrichment score per cell, which I estimated using ArchR (1.0.2) (*245*). Additionally, I required barcodes annotated as cells to have a minimum of 2,500 fragments and a TSS enrichment score of 2.5. I used ArchR (1.0.2) (*245*) to estimate doublet scores and removed the top 10% of cells in each sample, or any barcodes with more than 2.5 times the median number of fragments in that sample. Samples from the same stage were then jointly analyzed using ArchR (1.0.2) (*245*), counting accessibility in 500 bp-wide windows and inferring gene scores. Single-cell chromatin profiles were then projected into 50 latent dimensions based on an iterative LSI with gradually increasing clustering resolution (0.1, 0.2, 0.4, 0.8). These dimensions were then corrected using Harmony (*258*) to facilitate integration between biological replicates.

Integration across modalities was performed separately for each developmental stage using Seurat (4.0) CCA (*249*), as described in section 5.3 for the mouse snATAC-seq dataset. In addition to transferring cell type labels from the snRNA-seq data to the cells profiled with snATAC-seq, I also used the integration to impute coordinates for the merged snRNA-seq embedding (originally estimated based on liger as described above) for all snATAC-seq cells. This allowed me to co-embed the two modalities in the same latent dimensions, which were also used for the UMAP projections shown in **Figure 3.36**. Putative CREs in each species were identified using ArchR (1.0.2) (*245*) and MACS2 (2.1.2) (*250*), grouping cells based on their most precise label, as predicted by the snRNA-seq dataset, and requiring a peak to be identified in at least two replicates. Peak annotations from different cell types were merged as described for mouse in section 5.3. Orthologous CREs between human and mouse were identified based on reciprocal syntenic alignments, as described for mouse and opossum in section 5.4.

Developmental correspondences between human and mouse – Corresponding developmental stages between human and mouse were identified using a dynamic time-warping algorithm, as implemented in the R package dtw (1.22-3) (*346*), applied on four different metrics of dissimilarity. First, I used Seurat (4.0) CCA (*249*) to transfer snRNA-seq annotations from mouse to human based on 2,860 1:1 orthologous genes with highly variable expression in at least two samples in each species. Smoothing across the 30 nearest neighbors, I predicted mouse developmental stage annotations for each human cell in the snRNA-seq dataset. Then, for each human developmental stage, I calculated the mean prediction score for every mouse developmental stage across all cells, which I subtracted from 1 to convert into an estimate of dissimilarity.

As a second estimate, I considered similarities in cell type composition. To this end, I estimated the fraction of cells in each species and developmental stage belonging to every possible cell type (defined at an intermediate level of resolution of the cell type annotation that is consistent between human and mouse, referred to as "developmental state" in *Sepp, Leiss et al. 2021* (*179*)). I then estimated the Manhattan distance based on the fractions of cellular states between all possible combinations of human and mouse developmental stages.

The third estimate of dissimilarity was based on direct comparisons between gene expression profiles. For this, I aggregated gene expression profiles across all cells from the same sample, applied a variance-stabilizing transformation as implemented in DESeq2 (1.32) (*323*), and used the residuals of a linear model to correct expression estimates of each gene for a known bias in 10X Genomics technology (v3 versus v2). I then estimated mean gene expression values across replicates per developmental stage and identified highly variable genes (HVGs) as those with a variance greater or equal to 10% of their mean. Out of 1,412 human HVGs and 1,748 mouse HVGs during development, 734 were shared between the two species and were used for estimating the Spearman's correlation in expression profiles between all possible combinations of human and mouse developmental stages. I then subtracted correlation coefficients from 1 to convert these estimates into dissimilarity measurements.

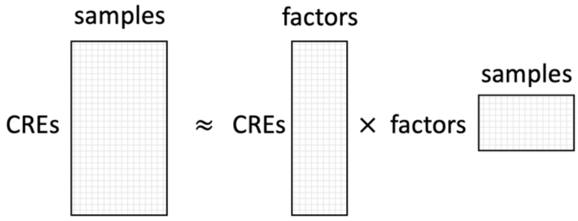
Finally, my fourth metric was based on chromatin accessibility comparisons between the two species. As for gene expression, I aggregated CRE accessibility profiles across all cells from the same sample, scaled for sequencing depth (CPM) and additionally applied a quantile normalization using the Bioconductor package preprocessCore (1.54). I then estimated mean accessibility values across replicates from the same developmental stage and selected highly variable CREs as those with a variance of at least 50% compared to their mean. A total of 37,475

1:1 orthologous CREs were found as highly variable in both species and were used for estimating Spearman's correlation coefficients in accessibility profiles between all possible combinations of human and mouse developmental stages. As with the previous correlation approach, I then subtracted the coefficients from 1 to convert these estimates into dissimilarity measurements. Each of the four dissimilarity matrices was used as input for the package dtw (1.22-3) (*346*). The four alignment paths and the consensus correspondences I inferred from them are shown in **Figure 3.36**.

Characterization of spatiotemporal CRE activity in human and mouse – To jointly characterize CRE accessibility profiles in both species, I first determined a set of groups (corresponding cell types, described at the "developmental state" level, and developmental stages) with at least 50 cells in at least two samples (i.e., biological replicates) for each modality (RNA, ATAC) in both species. I then aggregated gene expression/chromatin accessibility profiles across all cells from the group (cell type and developmental stage) and biological replicate. I then filtered, scaled and transformed these pseudobulks separately for each modality. For gene expression, I considered counts in full-length transcripts (as I found no evidence that these are less comparable between species than considering only exonic UMI counts) and filtered for protein-coding genes reaching at least 20 CPM in at least two replicates from the same group (cell type and developmental stage). I applied a variance stabilizing transformation (*323*) and then used the residuals of a linear model to regress out the effects of 10X technology (v3 versus v2) and of sex. For chromatin accessibility, I scaled for sequencing depth (CPM) and then applied a quantile normalization as implemented in the package preprocessCore (1.54), filtering for putative CREs with at least 5 quantile-corrected CPM in at least two samples (note that the number of CREs is approximately 25 times larger than that of protein-coding genes, leading to a large shift in the distribution of CPM values).

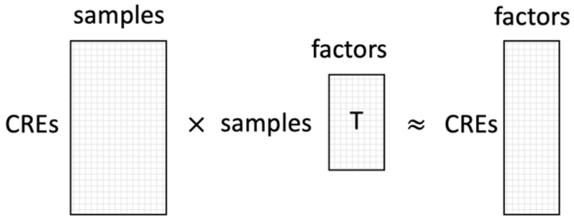
After filtering, I used non-negative matrix factorization (NMF) to summarize the patterns of CRE activity across cell types and developmental stages. To this end, I first calculated mean CRE accessibility values across all replicates from the same cell type and developmental stage. I then determined highly variable CREs across cell types and developmental stages in each species requiring a variance greater or equal to 3 times the mean (note that by separating cells from different cell types, CREs show much greater variance/mean ratios compared to when aggregating per developmental stage, hence the need for a higher cutoff than the one described in the developmental correspondences section). This filtering step resulted to 54,722 and 77,520 highly variable mouse and human CREs, respectively. The accessibility of each CRE was then standardized as a fraction of its maximum value across all groups (i.e., corresponding cell types and developmental stages between human and mouse).

Highly variable CREs from both species were used to learn an NMF representation of the dataset. In NMF, the original matrix [CREs x samples] is approximated by the multiplication of two new matrices that correspond to the loadings of a predetermined number of components on CREs and samples respectively, as illustrated below.



To identify the optimal number of NMF components, I considered multiple values in a range between 2 and 30. After each factorization step, I evaluated the degree of mixing between species by calculating the Euclidean distance between human and mouse CREs in the [CRE x factors] matrix. Additionally, I estimated the reconstruction error between the original [CRE x samples] matrix and the one inferred by the multiplication of the two factor-based matrices. Naturally, the reconstruction error decreases with the addition of more components, albeit with a significantly smaller rate after 18 components. Similarly adding more components, especially above 25, leads to a greater separation of the two species in the factor space. Based on these two metrics, and by considering the biological relevance of the identified associations between samples and factors, I determined the optimal number of components to be 18.

To assess the conservation of CREs associated with different components (i.e., accessible in different cell types and developmental stages), I had to be able to consider all CREs, to ensure that any differences are not due to biases in the cutoffs I used to define highly variable features. To this end, after learning the optimal [factors x samples] matrix by using the highly variable CREs, I multiplied the original [CRE x samples] matrix for all human and mouse CREs with the transpose of the [factors x samples] matrix to get a final matrix with loadings for all CREs on the 18 components.



I then used the loadings of all CREs to identify, in each species, the 10,000 CREs most strongly associated with each component. For each component, I estimated the fraction of human intergenic CREs (amongst the 10,000 assigned to that component) that were present in the mouse CRE annotation, and that had a 1:1 ortholog in the 10,000 mouse CREs assigned to the same component. To visualize CRE accessibility patterns across corresponding cell types and developmental stages, I randomly selected 500 CREs per species and component. To directly assess the level of conservation of spatiotemporal CRE activity between human and mouse, I considered 1:1 orthologs which were assigned to at least one component in at least one species. I then estimated the Euclidean distance between the component loadings of the human and mouse orthologs and compared the distribution of these distances to those observed when shuffling the orthology relationships. Obtaining loadings for all CREs and assigning a fixed number of CREs to each component were essential steps to facilitate unbiased comparisons of evolutionary conservation levels across components. All other analyses described below rely on the highly variable CREs, for which assignment to components is the most confident.

For the TF motif enrichment analysis, I first identified the highly variable CREs with the highest loadings for each component in each species. To account for differences in the distribution of loadings between components, I used elbow plots to determine the optimal loading cutoff for each component, which resulted to 7,000-15,000 putative CREs assigned to each component in each species. I then used pycistarget (1.0.1) to determine enrichments of each CRE set for a collection of more than 49,504 motifs collapsed into more than 8,000 clusters for human and mouse (*305*). I used a lenient normalized enrichment score (NES) cutoff of 0.1 to obtain enrichments for most motif clusters, then compared the similarity between human and mouse by calculating the Pearson's correlation in NES scores between CRE sets for motif clusters with a NES of at least 3 in either species.

Deep learning models of regulatory grammar – In this analysis I used deep learning models to learn the sequence features associated with the assignment of each CRE to different NMF components. This was framed as a classification task, in which each CRE is assigned to zero, one or more NMF components based on its spatiotemporal activity and the model tries to predict its assignment probability for each component based on its DNA sequence. For each component and species, I selected the 3,000 CREs with the highest loadings in that component. Additionally, for each species, I selected the 3,000 CREs with the lowest maximum loading across all components, which I used as a proxy for “noisy CREs” (Component 19). Furthermore, to help the model learn features associated with general CRE activity, I also selected 3,000 random 500 bp-long genomic regions that don’t overlap any CREs in my annotation (Component 20), additionally ensuring that they don’t overlap gaps in the genome assembly or contain undetermined nucleotides. After determining the coordinates of these regions in each species, I extracted their sequences using bedtools getfasta (2.29) (*321*). These sequences were used for training and evaluating different deep learning models, as described below.

The main training strategy was the same across all the models I developed during this work, and closely follows what was recently described by *Janssens et al. 2022* (*101*). In brief, DNA sequences were converted to a one-hot encoding, a matrix with four rows (one for each possible nucleotide) and 500 columns (one for each nucleotide position in the CRE sequence). At each column, the observed nucleotide is marked with 1 and the remaining three options are marked with 0. The total number of DNA sequences was split into 80% used for training the model, 10% for evaluating the performance across different training epochs (validation set), and 10% for testing the performance of the best-performing model (test set). To increase the number of sequences used for training and to allow the model to focus on relevant sequence features, augmentation was performed by extending the training sequences by 100 bp towards each side, then using a sliding window of 500 bp with a stride of 50bp to generate partially overlapping sequences with the same label.

The models use the architecture described by *Janssens et al. 2022* (*101*). Briefly, one-hot encoded sequences are used as input for a convolutional layer with 512 kernels of size 24, followed by a max-pooling layer with size and stride of 16. This is followed by a time-distributed dense layer together with a bidirectional long short-term memory (LSTM) layer with 256 neurons. Finally, the output of the LSTM layer is passed on to a flattened and then to a dense layer, which in turn uses a sigmoid activation function to estimate prediction probabilities for each of the 20 possible classes (18 NMF components, noisy peaks and random regions). To prevent overfitting, dropout layers are introduced after the max-pooling, LSTM and dense layers with dropout rates of 0.5, 0.2 and 0.5, respectively. Model performance was evaluated using the auROC and auPR metrics, estimated based on the testing dataset using the average_precision_score and roc_auc_score functions from the scikit-learn package. All models used the same train-validation-test split for human and mouse sequences, allowing me to compare performance based on sequences not seen by any model during training.

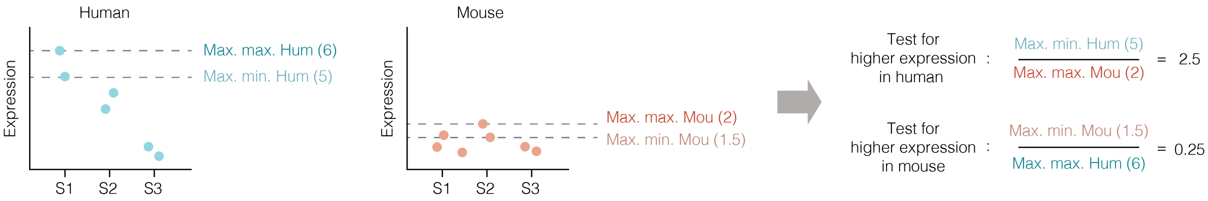
DeepExplainer importance was estimated as described in *Lundberg and Lee 2017* (*347*), using 500 random regions from the non-augmented datasets for initialization. Importance scores for a particular class and sequence were multiplied with the one-hot encoded matrix and visualized based on the viz_sequence function of the package DeepLift (*106*), with custom modifications by Ibrahim Taskiran (*100*). For the *in silico* mutagenesis, all positions of a CRE sequence were sequentially mutated into all other three possible nucleotides, as described in *Minnoye et al. 2020* (*100*). Thus, each new sequence differs from the original sequence by a single nucleotide. The model was then used to generate prediction scores for each of the simulated sequences. The difference in the prediction score between the original and each mutated sequence was estimated for the class of interest and visualized based on the position (x-axis) and nucleotide substitution (color) of each mutation.

Cell type-specific changes in gene expression – To identify changes in the expression of 1:1 orthologous protein-coding genes between human and mouse, and to relate these to my previous analyses of CRE accessibility, I used the pseudobulks generated for the same samples (corresponding cell types and developmental stages) that were considered for the NMF analysis (see *Characterization of spatiotemporal CRE activity in human and mouse*). Briefly, these are aggregated gene expression profiles across all cells from the same cell type, developmental stage and biological replicate, only considering groups with enough cells in both species and modalities. As described above, gene expression measurements were transformed using a variance-stabilizing transformation (VST), and corrected for differences in 10X technology and sex. To further minimize biases in VST estimates between species, I applied an additional median-scaling normalization.

To relate gene expression changes to different NMF components, I first identified the samples (cell types and developmental stages) with the highest loadings for each component. These were selected as all samples with at least 40% of the maximum loading for that component. For example, for component 1, these included granule cell progenitors from newborn humans and P4-P7 mice, and differentiating granule cells (subtype 1) from newborn and infant humans and P4-P14 mice. The goal of the analysis was to identify genes that are highly expressed in at least some of the samples associated with this component in one species but not in the other, while accounting for known technical biases and limitations of cross-species comparisons of gene expression levels.

Contrasting absolute expression levels (e.g., VST, CPM) between species is especially prone to differences in gene annotations (e.g., loss of reads in one species due to incomplete UTR annotations). Additionally, full-length transcripts are biased by the number of polyA stretches in gene introns (accounting for a significant fraction of total UMIs), which can differ markedly between species. Such confounders are expected to affect expression estimates in the same way across conditions (i.e., cell types and developmental stages) and can thus be addressed by standardizing expression values within each species (e.g., through estimating Z-scores). These relative gene expression values can then be compared between species, as a metric for changes in cell type- or time-specificity of expression (e.g., identifying a gene that is expressed in cell type A in mouse and in both cell types A and B in human). However, comparing relative expression values can also lead to the identification of genes that in practice only show minor quantitative differences between species, a problem especially prominent in lowly or broadly expressed genes. To address the technical issues associated with both approaches, I decided to consider both the difference in VST counts (as a quantitative measure) and the difference in Z-scores (as a relative measure).

Additionally, I aimed to identify changes that were reproducible across biological replicates. For this, I performed comparisons between human and mouse in an asymmetric way depending on whether I was testing for higher expression in human or mouse. For each sample (cell type and developmental stage) associated with a component, and for each species, I estimated the maximum and minimum expression value (VST or Z-score) across biological replicates. Since the goal of the analysis was to identify genes with high expression in at least some of the samples in one species but significantly lower expression in all samples in the other species, I then calculated the maximum values of these two estimates across all samples for that component. Then, to test for higher expression of a gene in human, I compared the minimum value across replicates in the sample with the maximum expression in human (Max. min. Hum.) to the maximum value across replicates in the sample with the maximum expression in mouse (Max. max. Mou.). Similarly, to test for higher expression of a gene in mouse, I compared the minimum value across replicates in the sample with the maximum expression in mouse (Max. min. Mou.) to the maximum value across replicates in the sample with the maximum expression in human (Max. max. Hum.). Thus, the identified differences in VST counts or Z-scores represent conservative estimates of the minimum difference observed between the highest expression of that gene across corresponding samples in the two species, as shown in the diagram below.



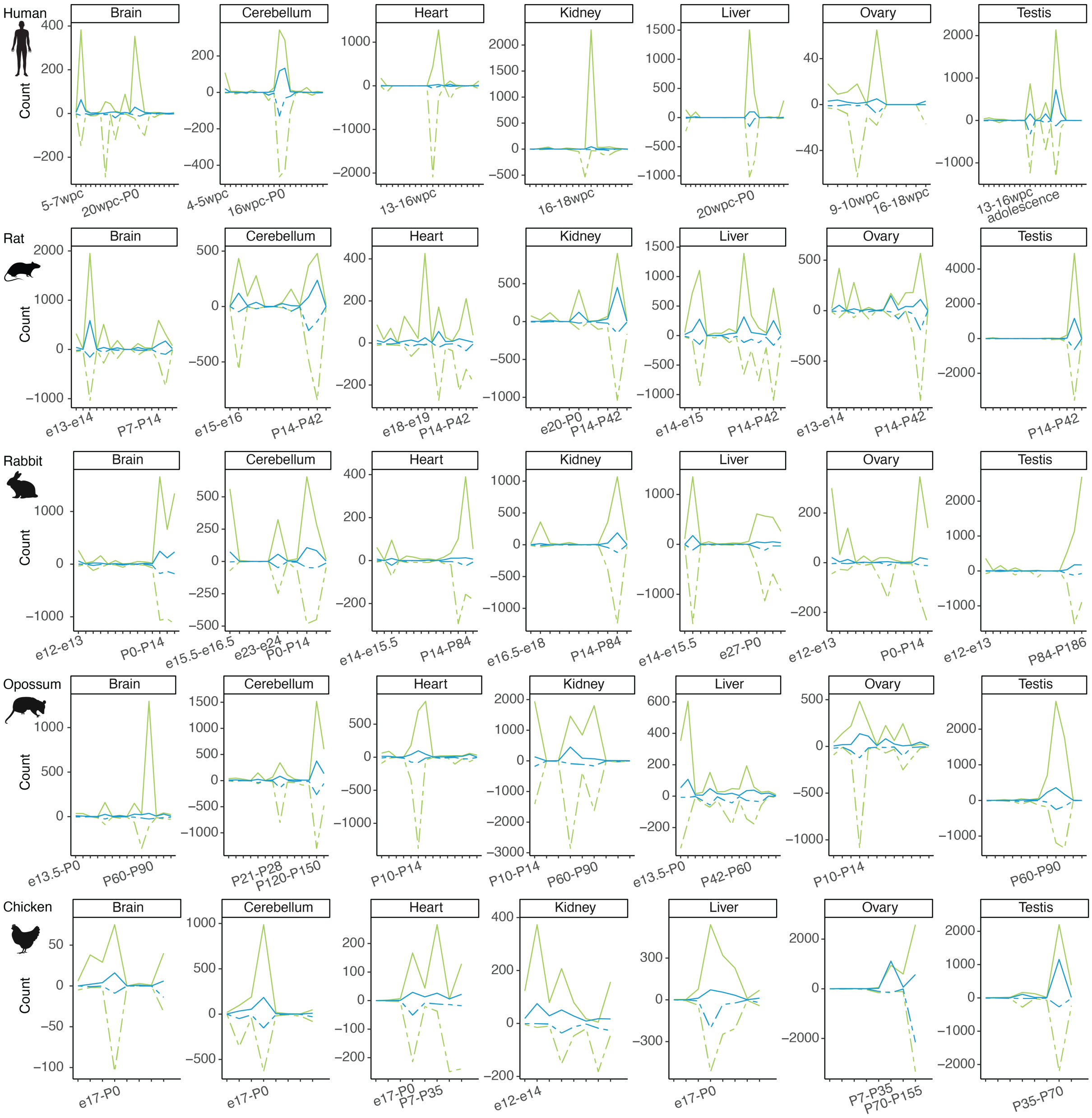
Collectively, for each component and for each gene, I estimated two metrics (based on VST counts and Z-scores, respectively) to test for higher expression in human and two metrics to test for higher expression in mouse. I then identified genes with higher expression in one species as those with ΔVST ≥ 2 and ΔZ ≥ 0.5, additionally requiring the gene to reach at least 7.5 VST counts and a Z-score of 0 in at least one replicate across all samples associated with that component for the species in which it was called as highly expressed. These cutoffs were optimized based on manual inspection of the gene expression profiles of the identified genes, but are admittedly arbitrary. They serve the main purpose of this analysis, which is to identify the most striking and confident changes in gene expression as a first step towards exploring the regulatory basis of these changes. However, in the future, I plan to explore the possibility to obtain more unbiased estimates for these cutoffs and to infer the statistical significance of each identified gene expression change.

Regulatory basis of gene expression changes – For the global analysis of the chromatin accessibility landscape around genes with changes in their expression, I considered all CREs within 250 kb upstream or downstream of a gene's TSS. I compared the distribution of the loadings of human CREs adjacent to genes with higher expression in human or mouse in the samples associated with the corresponding component. I then compared the fraction of human CREs that were conserved in the mouse CRE annotation (based on reciprocal syntenic alignments) between CREs adjacent to genes with higher expression in human compared to mouse and those adjacent to genes with high expression in humans (VST ≥ 7.5) but no significant increase compared to mouse. Finally, for 1:1 orthologous CREs that are adjacent to genes with higher expression in human or mouse, I estimated the difference in the loadings for the corresponding NMF component between the human and the mouse CRE ortholog.

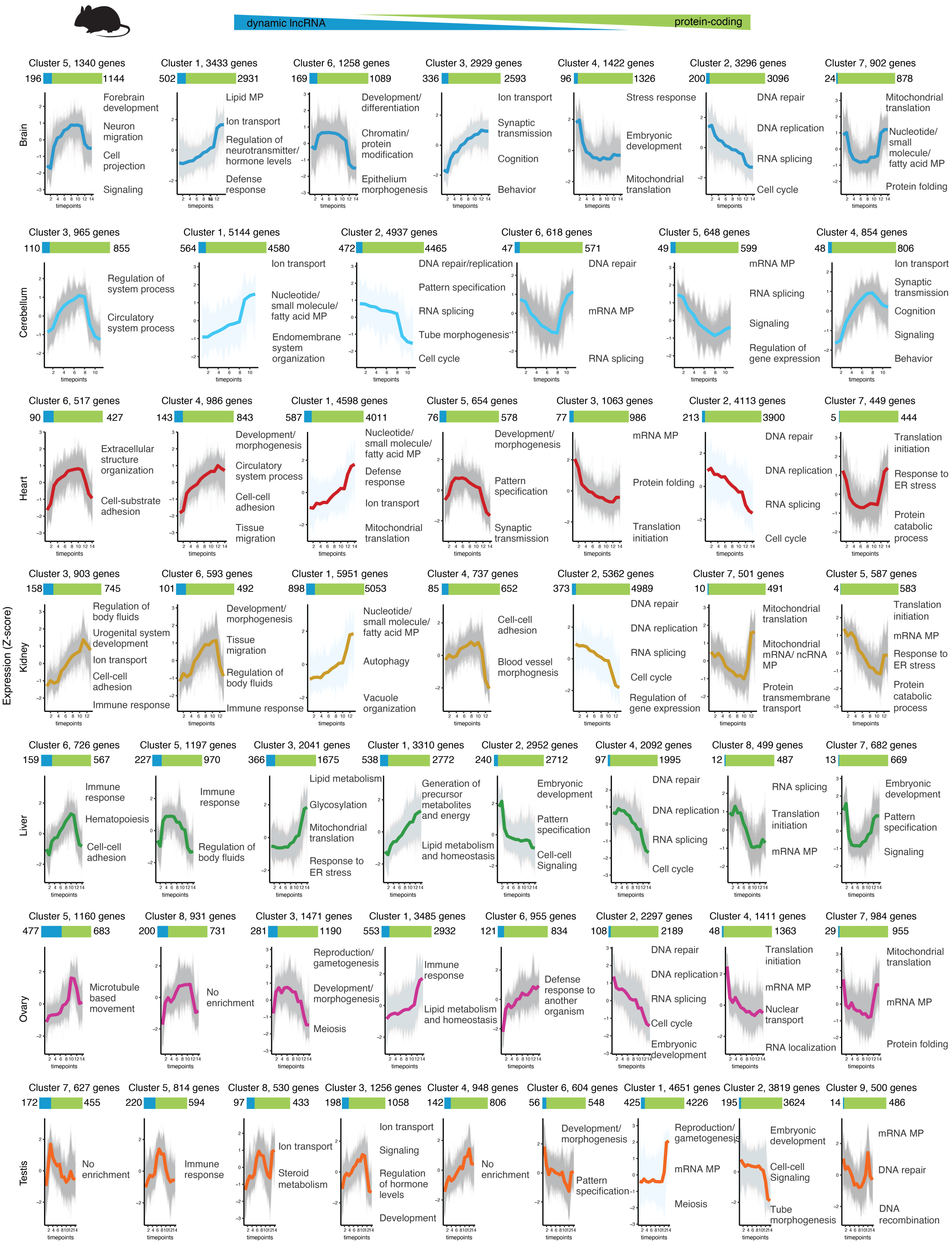
To associate specific CREs with the expression change in *PIEZO2*, I identified all human CREs within 250 kb upstream or downstream of its TSS. I then filtered for CREs with high loadings in the corresponding component (NMF2) based on the elbow plot procedure described above. I considered CREs that were human-specific (no mouse CRE ortholog) and those for which the mouse CRE ortholog was not included in the CREs with high loadings for that component, identified based on the same elbow procedure (i.e., repurposed CREs). I then used the deep learning model trained with the full set of human and mouse CRE sequences to predict accessibility patterns (i.e., assignment to different components for the remaining human CRE sequences) and only considered those with a prediction score of at least 0.2.

The predictions of CRE accessibility across mammalian species were performed by Mr. Tetsuya Yamada, based on my original design and supervision. Briefly, Mr. Yamada downloaded genomic alignments across mammalian species from the Zoonomia consortium (*268*) in .hal format. He used halLiftover (*348*) to extract orthologous sequences for these CREs in each mammalian species, and then applied HALPER (*349*) to identify the putative summit of the orthologous region and to extend it by 250 bp to each direction. After converting the coordinates to fasta sequences and eventually to one-hot encoded matrices, as described above, he used the deep learning model I trained based on the full set of human and mouse CRE sequences to predict accessibility patterns for each mammalian species in which an alignment could be detected. He then used MAFFT (*350*) to generate multiple sequence alignments for each CRE. I used these data to manually inspect the predicted evolutionary histories of human CREs and to explore multiple sequence alignments — visualized using Jalview (*351*) — in regions of the human CREs highlighted as important by my DeepExplainer and *in silico* mutagenesis analyses.

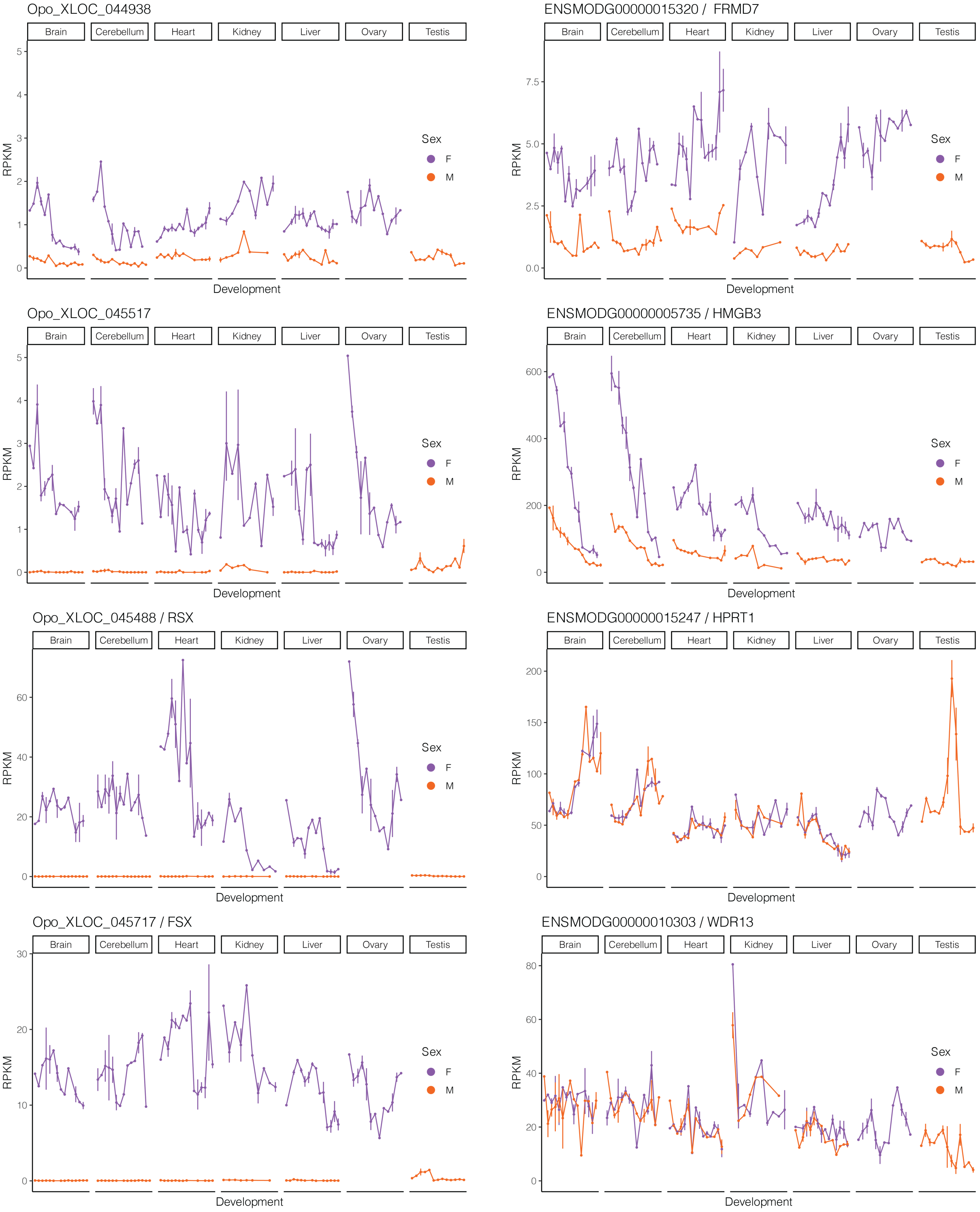
6. Supplementary Figures



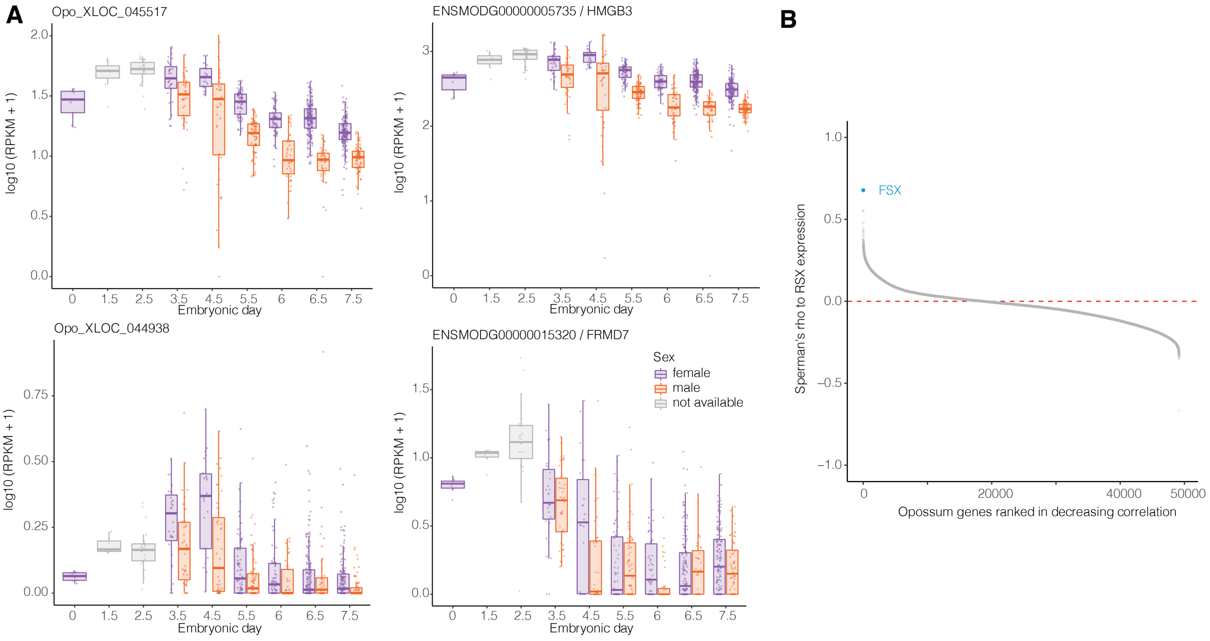
Supplementary Figure 1: Gene expression changes during mammalian organ development. Number of differentially expressed protein-coding genes (green) and dynamic lncRNAs (blue) between adjacent stages of organ development in human, rat, rabbit, opossum and chicken. *Figure adapted from Sarropoulos et al. 2019* (223).



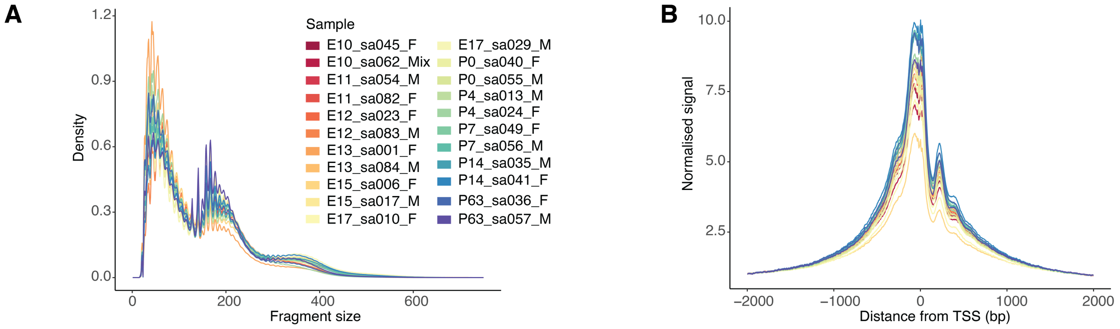
Supplementary Figure 2: Clustering of dynamic lncRNAs based on developmental trajectories. Clusters of developmentally dynamic lncRNAs and protein-coding genes across mouse organs. Grey lines represent individual gene trajectories and solid lines posterior mean trajectories for each cluster. Clusters are arranged by decreasing fraction of lncRNAs. Enriched representative biological processes (Benjamini–Hochberg adjusted $P < 0.05$, hypergeometric test) are shown for each cluster. *Figure reproduced from Sarropoulos et al. 2019* (223).



Supplementary Figure 3: Expression profiles of female-specific genes and their neighbors. Gene expression profiles (RPKM) across developing organs and sexes for four female-specific lncRNAs on the opossum X chromosome (left) and their closest protein-coding genes (right).



Supplementary Figure 4: Expression profiles of other genes at the onset of marsupial X-chromosome inactivation. **(A)** Expression profiles (RPKM) of the remaining four originally identified female-specific genes on the opossum X in single-cells from early opossum embryos. **(B)** Spearman's correlation coefficients between the expression profiles of all opossum genes and *RSX*, sorted in decreasing order. *FSX* is highlighted in blue.



Supplementary Figure 5: Quality control of mouse snATAC-seq libraries. **(A)** Fragment size distribution of the snATAC-seq libraries, labeled by developmental stage, library code and sex (F: female; M: male; Mix: pooled embryos from both sexes). **(B)** TSS enrichment scores of the snATAC-seq libraries. *Figure adapted from Sarropoulos et al. 2021* (*244*).

7. References

1. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

2. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol* **15**, e2003243 (2017).

3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

4. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature 2003 420:6915* **420**, 520–562 (2002).

5. Breschi, A., Gingeras, T. R. & Guigó, R. Comparative transcriptomics in human and mouse. *Nat Rev Genet* **18**, 425–440 (2017).

6. Ponting, C. P. & Hardison, R. C. What fraction of the human genome is functional? *Genome Res* **21**, 1769–1776 (2011).

7. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

8. Doolittle, W. F. Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences* **110**, 5294–5300 (2013).

9. Graur, D. *et al.* On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE. *Genome Biol Evol* **5**, 578–590 (2013).

10. Eddy, S. R. The ENCODE project: Missteps overshadowing a success. *Current Biology* **23**, R259–R261 (2013).

11. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**, 6131–6138 (2014).

12. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat Rev Genet* **13**, 36–46 (2012).

13. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: From conflicts to benefits. *Nat Rev Genet* **18**, 71–86 (2017).

14. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* **9**, 397–405 (2008).

15. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Research* **20**, 1313–1326 (2010).

16. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: Recent insights and future perspectives. *Nat Rev Genet* **17**, 207–223 (2016).

17. Gil, N. & Ulitsky, I. Production of Spliced Long Noncoding RNAs Specifies Regions with Increased Enhancer Activity. *Cell Syst* **7**, 537-547.e3 (2018).

18. Tan, J. Y., Biasini, A., Young, R. S. & Marques, A. C. Splicing of enhancer-associated lincRNAs contributes to enhancer activity. *Life Sci Alliance* **3**, e202000663 (2020).

19. Ulitsky, I. & Bartel, D. P. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154**, 26–46 (2013).

20. Kopp, F. & Mendell, J. T. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* **172**, 393–407 (2018).

21. Blobel, G. & Potter, V. R. Studies on free and membrane-bound ribosomes in rat liver: I. Distribution as related to total cellular RNA. *J Mol Biol* **26**, 279–292 (1967).

22. Esteller, M. Non-coding RNAs in human disease. *Nat Rev Genet* **12**, 861–874 (2011).

23. Schuller, A. P. & Green, R. Roadblocks and resolutions in eukaryotic translation. *Nat Rev Mol Cell Biol* **19**, 526–541 (2018).

24. Maniatis, T. & Reed, R. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature* **325**, 673–678 (1987).

25. Kiss, T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J* **20**, 3617–3622 (2001).

26. Okamura, K. & Lai, E. C. Endogenous small interfering RNAs in animals. *Nat Rev Mol Cell Biol* **9**, 673–678 (2008).

27. Bartel, D. P. Metazoan MicroRNAs. *Cell* **173**, 20–51 (2018).

28. Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K. & Hannon, G. J. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**, 744–747 (2007).

29. Ponting, C. P., Oliver, P. L. & Reik, W. Evolution and Functions of Long Noncoding RNAs. *Cell* **136**, 629–641 (2009).

30. Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164**, 69–80 (2016).

31. Munschauer, M. *et al.* The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* **561**, 132–136 (2018).

32. Williamson, L. *et al.* UV Irradiation Induces a Non-coding RNA that Functionally Opposes the Protein Encoded by the Same Gene. *Cell* **168**, 843-855.e13 (2017).

33. Ponting, C. P. & Haerty, W. Genome-Wide Analysis of Human Long Noncoding RNAs: A Provocative Review. *Annu Rev Genomics Hum Genet* **23**, 153-172 (2022).

34. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7**, 29–59 (2006).

35. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* **21**, 71–87 (2020).

36. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: From properties to genome-wide predictions. *Nat Rev Genet* **15,** 272–286 (2014).

37. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**, 1170–1187 (2016).

38. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat Rev Genet* **17**, 661–678 (2016).

39. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).

40. Ørom, U. A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46–58 (2010).

41. Gisselbrecht, S. S. *et al.* Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. *Mol Cell* **77**, 324-337.e8 (2020).

42. Huang, D. & Ovcharenko, I. Enhancer-silencer transitions in the human genome. *Genome Res* **32**, 437–448 (2022).

43. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

44. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13**, 329–342 (2012).

45. Ulitsky, I. Evolution to the rescue: Using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet* **17**, 601–614 (2016).

46. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).

47. Cabili, M. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25**, 1915–1927 (2011).

48. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635–640 (2014).

49. Washietl, S., Kellis, M. & Garber, M. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* **24**, 616–628 (2014).

50. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).

51. Dominic Mills, J., Kawahara, Y. & Janitz, M. Strand-Specific RNA-Seq Provides Greater Resolution of Transcriptome Profiling. *Curr Genomics* **14**, 173–181 (2013).

52. Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199–204 (2017).

53. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat Genet* **49**, 1731–1740 (2017).

54. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* **19**, 535–548 (2018).

55. Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521–1531 (2013).

56. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* **6**, e21856 (2017).

57. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207–220 (2019).

58. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).

59. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13**, 599–604 (2018).

60. Cabili, M. N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol* **16**, 20 (2015).

61. Liu, S. J. *et al.* Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol* **17**, 67 (2016).

62. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).

63. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).

64. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145 (2015).

65. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).

66. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925–936 (2019).

67. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* **37**, 916–924 (2019).

68. Argelaguet, R. *et al.* Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).

69. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).

70. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22**, 1775–1789 (2012).

71. Melé, M. *et al.* Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res* **27**, 27–37 (2017).

72. Unfried, J. P. & Ulitsky, I. Substoichiometric action of long noncoding RNAs. *Nat Cell Biol* **24**, 608–615 (2022).

73. Soumillon, M. *et al.* Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Rep* **3**, 2179–2190 (2013).

74. Murat, F. *et al.* The molecular evolution of spermatogenesis across mammals. *bioRxiv* 2021.11.08.467712 (2021).

75. Hong, S. H., Han, G., Lee, S. J., Cocquet, J. & Cho, C. Testicular germ cell–specific lncRNA, Teshl, is required for complete expression of Y chromosome genes and a normal offspring sex ratio. *Sci Adv* **7**, eabg5177 (2021).

76. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: Past, present, and future. *Genetics* **193**, 651–669 (2013).

77. Hezroni, H. *et al.* Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species. *Cell Rep* **11**, 1110–1122 (2015).

78. Kapusta, A. *et al.* Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genet* **9**, e1003470 (2013).

79. Chen, J. *et al.* Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol* **17**, 19 (2016).

80. Wu, X. & Sharp, P. A. Divergent Transcription: A Driving Force for New Gene Origination? *Cell* **155**, 990–996 (2013).

81. Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M. & Kaessmann, H. Repurposing of promoters and enhancers during mammalian evolution. *Nat Commun* **9**, 4066 (2018).

82. Hezroni, H. *et al.* A subset of conserved mammalian long non-coding RNAs are fossils of ancestral protein-coding genes. *Genome Biol* **18**, 162 (2017).

83. Duret, L., Chureau, C., Samain, S., Weissanbach, J. & Avner, P. The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* **312**, 1653–1655 (2006).

84. Yotova, I. Y. *et al.* Identification of the human homolog of the imprinted mouse Air non-coding RNA. *Genomics* **92**, 464–473 (2008).

85. Kirk, J. M. *et al.* Functional classification of long non-coding RNAs by k-mer content. *Nat Genet* **50**, 1474–1482 (2018).

86. Young, A. I., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**, 1396–1400 (2019).

87. Clemson, C. M., McNeil, J. A., Willard, H. F. & Lawrence, J. B. XIST RNA paints the inactive X chromosome at interphase: Evidence for a novel RNA involved in nuclear/chromosome structure. *Journal of Cell Biology* **132**, 259–275 (1996).

88. Brown, C. J. *et al.* The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**, 527–542 (1992).

89. Loda, A. *et al.* Genetic and epigenetic features direct differential efficiency of Xist-mediated silencing at X-chromosomal and autosomal locations. *Nat Commun* **8**, 690 (2017).

90. Chu, C. *et al.* Systematic Discovery of Xist RNA Binding Proteins. *Cell* **161**, 404–416 (2015).

91. Rinn, J. L. *et al.* Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* **129**, 1311–1323 (2007).

92. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. *Mol Cell* **44**, 667–678 (2011).

93. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, aah7111 (2017).

94. Bester, A. C. *et al.* An Integrated Genome-wide CRISPRa Approach to Functionalize lncRNAs in Drug Resistance. *Cell* **173**, 649-664.e20 (2018).

95. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50,** D165–D173 (2021).

96. Thanos, D. & Maniatis, T. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).

97. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **5**, 276–287 (2004).

98. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the ACM* **60,** 84–90 (2017).

99. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**, 990–999 (2016).

100. Minnoye, L. *et al.* Cross-species analysis of enhancer logic using deep learning. *Genome Res* **31**, 1815–1834 (2020).

101. Janssens, J. *et al.* Decoding gene regulation in the fly brain. *Nature* **601**, 630–636 (2022).

102. Trevino, A. E. *et al.* Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053-5069.e23 (2021).

103. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat Genet* **53**, 354–366 (2021).

104. de Almeida, B. P., Reiter, F., Pagani, M. & Stark, A. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet* **54**, 613–624 (2022).

105. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).

106. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv* 1704.02685 (2017).

107. Shrikumar, A. *et al.* Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. *arXiv* 1811.00416 (2018).

108. Nair, S., Shrikumar, A., Schreiber, J. & Kundaje, A. fastISM: performant in silico saturation mutagenesis for convolutional neural networks. *Bioinformatics* **38,** 2397–2403 (2022).

109. Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. *Science* **361**, 1341–1345 (2018).

110. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet* **20**, 437–455 (2019).

111. Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* **569**, 345–354 (2019).

112. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

113. Berthelot, C., Muffato, M., Abecassis, J. & Roest Crollius, H. The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Rep* **10**, 1913–1924 (2015).

114. Eres, I. E. & Gilad, Y. A TAD Skeptic: Is 3D Genome Topology Conserved? *Trends in Genetics* **37**, 216–223 (2021).

115. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e24 (2017).

116. Ahanger, S. H. *et al.* Distinct nuclear compartment-associated genome architecture in the developing mammalian brain. *Nat Neurosci* **24**, 1235–1242 (2021).

117. Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* **2**, 152–163 (2018).

118. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* **351**, 1083–1087 (2016).

119. Roller, M. *et al.* LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions. *Genome Biol* **22**, 62 (2021).

120. Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012 (2014).

121. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040 (2010).

122. Stefflova, K. *et al.* Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. *Cell* **154**, 530–540 (2013).

123. Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans-mechanisms and functional implications. *Nat Rev Genet* **15**, 221–233 (2014).

124. Stergachis, A. B. *et al.* Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**, 365–370 (2014).

125. Arendt, D. *et al.* The origin and evolution of cell types. *Nat Rev Genet* **17**, 744–757 (2016).

126. Cheng, Y. *et al.* Principles of regulatory information conservation between mouse and human. *Nature* **515**, 371–375 (2014).

127. Carter, B. & Zhao, K. The epigenetic basis of cellular heterogeneity. *Nat Rev Genet* **22**, 235–250 (2021).

128. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**, 25–36 (2008).

129. Ingham, P. W. & McMahon, A. P. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev* **15**, 3059–3087 (2001).

130. Leto, K. *et al.* Consensus Paper: Cerebellar Development. *Cerebellum* **15**, 789–828 (2016).

131. Bermingham, N. A. *et al.* Math1: an essential gene for the generation of inner ear hair cells. *Science* **284**, 1837–1841 (1999).

132. Ben-Arie, N. *et al.* Functional conservation of atonal and Math1 in the CNS and PNS. *Development* **127**, 1039–1048 (2000).

133. Yang, Q., Bermingham, N. A., Finegold, M. J. & Zoghbi, H. Y. Requirement of Math1 for secretory cell lineage commitment in the mouse intestine. *Science* **294**, 2155–2158 (2001).

134. Kawaguchi, Y. *et al.* The role of the transcriptional regulator Ptf1a in converting intestinal to pancreatic progenitors. *Nat Genet* **32**, 128–134 (2002).

135. Fujitani, Y. *et al.* Ptf1a determines horizontal and amacrine cell fates during mouse retinal development. *Development* **133**, 4439–4450 (2006).

136. Leung, C. Y. & Zernicka-Goetz, M. Mapping the journey from totipotency to lineage specification in the mouse embryo. *Curr Opin Genet Dev* **34**, 71–76 (2015).

137. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).

138. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

139. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol* **16**, 144–154 (2015).

140. Nord, A. S. & West, A. E. Neurobiological functions of transcriptional enhancers. *Nat Neurosci* **23**, 5–14 (2020).

141. Dickinson, M. E. *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).

142. Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**, 797–803 (2005).

143. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* **46**, 61–64 (2013).

144. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol Med* **27**, 1060–1073 (2021).

145. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).

146. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).

147. Kvon, E. Z., Waymack, R., Gad, M. & Wunderlich, Z. Enhancer redundancy in development and disease. *Nat Rev Genet* **22**, 324–336 (2021).

148. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).

149. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).

150. Gorkin, D. U. *et al.* An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).

151. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* **21**, 432–439 (2018).

152. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).

153. Pijuan-Sala, B. *et al.* Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis. *Nat Cell Biol* **22**, 487–497 (2020).

154. Li, Y. E. *et al.* An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**, 129–136 (2021).

155. Stergachis, A. B. *et al.* Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. *Cell* **154**, 888–903 (2013).

156. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184,** 5985-6001 (2021).

157. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).

158. Mayr, E. *Animal Species and Evolution*. Harvard University Press, 1963.

159. Lewis, E. B. A gene complex controlling segmentation in Drosophila. *Nature* **276**, 565–570 (1978).

160. Pearson, J. C., Lemons, D. & McGinnis, W. Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* **6**, 893–904 (2005).

161. Heide, M. *et al.* Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. *Science* **369**, 546–550 (2020).

162. Namba, T. *et al.* Human-Specific ARHGAP11B Acts in Mitochondria to Expand Neocortical Progenitors by Glutaminolysis. *Neuron* **105**, 867-881.e9 (2020).

163. Fiddes, I. T. *et al.* Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* **173**, 1356-1369.e22 (2018).

164. Enard, W. *et al.* Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**, 869–872 (2002).

165. Schreiweis, C. *et al.* Humanized Foxp2 accelerates learning by enhancing transitions from declarative to procedural performance. *Proc Natl Acad Sci U S A* **111**, 14253–14258 (2014).

166. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).

167. Necsulea, A. & Kaessmann, H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet* **15**, 734–748 (2014).

168. Kvon, E. Z. *et al.* Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633-642.e11 (2016).

169. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011).

170. Romero, I. G., Ruvinsky, I. & Gilad, Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet* **13**, 505–516 (2012).

171. Perry, G. H. *et al.* Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res* **22**, 602–610 (2012).

172. Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019).

173. Cardoso-Moreira, M. *et al.* Developmental Gene Expression Differences between Humans and Mammalian Models. *Cell Rep* **33**, 108308 (2020).

174. Bakken, T. E. *et al.* A comprehensive transcriptional map of primate brain development. *Nature* **535**, 367–375 (2016).

175. Sousa, A. M. M. *et al.* Molecular and cellular reorganization of neural circuits in the human lineage. *Science* **358**, 1027–1032 (2017).

176. la Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580.e19 (2016).

177. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).

178. Krienen, F. M. *et al.* Innovations present in the primate interneuron repertoire. *Nature* **586**, 262–269 (2020).

179. Sepp, M. *et al.* Cellular development and evolution of the mammalian cerebellum. *bioRxiv* 2021.12.20.473443 (2021).

180. Cotney, J. *et al.* The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb. *Cell* **154**, 185–196 (2013).

181. Reilly, S. K. *et al.* Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155–1159 (2015).

182. Prescott, S. L. *et al.* Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. *Cell* **163**, 68–83 (2015).

183. Liu, X. *et al.* Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. *Genome Res* **22**, 611–622 (2012).

184. Kanton, S. *et al.* Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature* **574,** 418–422 (2019).

185. Abzhanov, A. Von Baer's law for the ages: Lost and found principles of developmental evolution. *Trends in Genetics* **29,** 712–722 (2013).

186. Irie, N. & Kuratani, S. The developmental hourglass model: a predictor of the basic body plan? *Development* **141**, 4649–4655 (2014).

187. Kalinka, A. T. *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811–816 (2010).

188. Liu, J. *et al.* The hourglass model of evolutionary conservation during embryogenesis extends to developmental enhancers with signatures of positive selection. *Genome Res* **31**, 1573–1581 (2021).

189. Potrzebowski, L. *et al.* Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol* **6**, 709–716 (2008).

190. Lahn, B. T. & Page, D. C. Four evolutionary strata on the human X chromosome. *Science* **286**, 964–967 (1999).

191. Wilson, M. A. & Makova, K. D. Genomic analyses of sex chromosome evolution. *Annu Rev Genomics Hum Genet* **10**, 333–354 (2009).

192. Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P. & Lovell-Badge, R. Male development of chromosomally female mice transgenic for Sry. *Nature* **351**, 117–121 (1991).

193. Kashimada, K. & Koopman, P. Sry: The master switch in mammalian sex determination. *Development* **137**, 3921–3930 (2010).

194. Charlesworth, B. & Charlesworth, D. The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society B: Biological Sciences* **355**, 1563–1572 (2000).

195. Bachtrog, D. Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* **14**, 113–124 (2013).

196. Cortez, D. *et al.* Origins and functional evolution of y chromosomes across mammals. *Nature* **508**, 488–493 (2014).

197. Julien, P. *et al.* Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol* **10**, e1001328 (2012).

198. Galupa, R. & Heard, E. X-chromosome inactivation: A crossroads between chromosome architecture and gene regulation. *Annu Rev Genet* **52**, 535–566 (2018).

199. Graves, J. A. M. Evolution of vertebrate sex chromosomes and dosage compensation. *Nat Rev Genet* **17**, 33–46 (2016).

200. McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232–236 (2015).

201. Grant, J. *et al.* Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* **487**, 254–258 (2012).

202. Chaumeil, J. *et al.* Evolution from XIST-Independent to XIST-Controlled X-Chromosome Inactivation: Epigenetic Modifications in Distantly Related Mammals. *PLoS One* **6**, e19040 (2011).

203. Sathyanesan, A. *et al.* Emerging connections between cerebellar development, behaviour and complex brain disorders. *Nat Rev Neurosci* **20**, 298–313 (2019).

204. Guerreiro Stucklin, A. S. & Grotzer, M. A. Cerebellar tumors. *Handb Clin Neurol* **155**, 289–299 (2018).

205. Butts, T., Green, M. J. & Wingate, R. J. T. Development of the cerebellum: Simple steps to make a 'little brain'. *Development* **141**, 4031–4041 (2014).

206. White, J. J. & Sillitoe, R. v. Development of the cerebellum: From gene expression patterns to circuit maps. *Wiley Interdiscip Rev Dev Biol* **2**, 149–164 (2013).

207. Cerrato, V. *et al.* Multiple origins and modularity in the spatiotemporal emergence of cerebellar astrocyte heterogeneity. *PLoS Biol* **16**, e2005513 (2018).

208. Parmigiani, E. *et al.* Heterogeneity and bipotency of astroglial-like cerebellar progenitors along the interneuron and glial lineages. *Journal of Neuroscience* **35**, 7388–7402 (2015).

209. Carter, R. A. *et al.* A Single-Cell Transcriptional Atlas of the Developing Murine Cerebellum. *Current Biology* **28**, 2910-2920.e2 (2018).

210. Vladoiu, M. C. *et al.* Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature* **572**, 67–73 (2019).

211. Wizeman, J. W., Guo, Q., Wilion, E. M. & Li, J. Y. H. Specification of diverse cell types during early neurogenesis of the mouse cerebellum. *Elife* **8**, e42388 (2019).

212. Aldinger, K. A. *et al.* Spatial and cell type transcriptional landscape of human cerebellar development. *Nat Neurosci* **24**, 1163–1175 (2021).

213. Butts, T., Modrell, M. S., Baker, C. V. H. & Wingate, R. J. T. The evolution of the vertebrate cerebellum: Absence of a proliferative external granule layer in a non-teleost ray-finned fish. *Evol Dev* **16**, 92–100 (2014).

214. Iulianella, A., Wingate, R. J., Moens, C. B. & Capaldo, E. The generation of granule cells during the development and evolution of the cerebellum. *Developmental Dynamics* **248**, 506–513 (2019).

215. Kebschul, J. M. *et al.* Cerebellar nuclei evolved by repeatedly duplicating a conserved cell-type set. *Science* **370**, eabd5059 (2020).

216. Libé-philippot, B. & Vanderhaeghen, P. Cellular and Molecular Mechanisms Linking Human Cortical Development and Evolution. *Annu Rev Genet* **55**, 555-581 (2021).

217. Pinson, A. & Huttner, W. B. Neocortex expansion in development and evolution—from genes to progenitor cell biology. *Curr Opin Cell Biol* **73**, 9–18 (2021).

218. Herculano-Houzel, S., Catania, K., Manger, P. R. & Kaas, J. H. Mammalian Brains Are Made of These: A Dataset of the Numbers and Densities of Neuronal and Nonneuronal Cells in the Brain of Glires, Primates, Scandentia, Eulipotyphlans, Afrotherians and Artiodactyls, and Their Relationship with Body Mass. *Brain Behav Evol* **86**, 145–163 (2015).

219. Barton, R. A. Embodied cognitive evolution and the cerebellum. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 2097–2107 (2012).

220. Barton, R. A. & Venditti, C. Rapid evolution of the cerebellum in humans and other great apes. *Current Biology* **24**, 2440–2444 (2014).

221. Neubauer, S., Hublin, J. J. & Gunz, P. The evolution of modern human brain shape. *Sci Adv* **4**, eaao5961 (2018).

222. Haldipur, P. *et al.* Spatiotemporal expansion of primary progenitor zones in the developing human cerebellum. *Science* **366**, 454–460 (2019).

223. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510–514 (2019).

224. Sarropoulos, I. Functional roles and evolutionary dynamics of mammalian developmentally dynamic lncRNAs. MSc thesis, Heidelberg University, 2017.

225. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* **46**, D754–D761 (2018).

226. Conesa, A., Nueda, M. J., Ferrer, A. & Talón, M. maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* **22**, 1096–1102 (2006).

227. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: A database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res* **45**, D61–D67 (2017).

228. Olson, E. N. Gene regulatory networks in the evolution and development of the heart. *Science* **313,** 1922–1927 (2006).

229. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

230. Quek, X. C. *et al.* lncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res* **43**, D168–D173 (2015).

231. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).

232. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147**, 1537–1550 (2011).

233. Amaral, P. P. *et al.* Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci. *Genome Biol* **19**, 32 (2018).

234. Jiang, W., Liu, Y., Liu, R., Zhang, K. & Zhang, Y. The lncRNA DEANR1 facilitates human endoderm differentiation by activating FOXA2 expression. *Cell Rep* **11**, 137–148 (2015).

235. Jian, X. & Felsenfeld, G. Insulin promoter in human pancreatic β cells contacts diabetes susceptibility loci and regulates genes affecting insulin metabolism. *Proc Natl Acad Sci U S A* **115**, E4633–E4641 (2018).

236. Spigoni, G., Gedressi, C. & Mallamaci, A. Regulation of Emx2 expression by antisense transcripts in murine cortico-cerebral precursors. *PLoS One* **5**, e8658 (2010).

237. Sauvageau, M. *et al.* Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2,** e01749 (2013).

238. Luo, S. *et al.* Divergent lncRNAs regulate gene expression and lineage differentiation in pluripotent cells. *Cell Stem Cell* **18**, 637–652 (2016).

239. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res* **35**, D88–D92 (2007).

240. Studer, T. The developmental sex-biased expression of genes escaping X chromosome inactivation across mammals. PhD thesis, Heidelberg University, 2018. doi:10.11588/HEIDOK.00025740.

241. Mahadevaiah, S. K., Sangrithi, M. N., Hirota, T. & Turner, J. M. A. A single-cell transcriptome atlas of marsupial embryogenesis and X inactivation. *Nature* **586**, 612–617 (2020).

242. Sprague, D. *et al.* Nonlinear sequence similarity between the Xist and Rsx long noncoding RNAs suggests shared functions of tandem repeat domains. *RNA* **25**, 1004–1019 (2019).

243. Hobbs, M. *et al.* A transcriptome resource for the koala (Phascolarctos cinereus): Insights into koala retrovirus transcription and sequence diversity. *BMC Genomics* **15**, 786 (2014).

244. Sarropoulos, I. *et al.* Developmental and evolutionary dynamics of cis-regulatory elements in mouse cerebellar cells. *Science* **373**, eabg4696 (2021).

245. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* **53**, 403–411 (2021).

246. Millen, K. J., Steshina, E. Y., Iskusnykh, I. Y. & Chizhikov, V. v. Transformation of the cerebellum into more ventral brainstem fates causes cerebellar agenesis in the absence of Ptf1a function. *Proc Natl Acad Sci U S A* **111**, E1777–E1786 (2014).

247. Prekop, H. T. *et al.* Sox14 is required for a specific subset of cerebello–olivary projections. *Journal of Neuroscience* **38**, 9539–9550 (2018).

248. Allen Institute for Brain Science, Developing Mouse Brain Atlas. http://developingmouse.brain-map.org/ (2008).

249. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

250. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

251. Birnbaum, R. Y. *et al.* Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res* **22**, 1059–1068 (2012).

252. Dalby, M., Rennie, S. & Andersson, R. FANTOM5 transcribed enhancers in mm10 [Data set]. http://doi.org/10.5281/zenodo.1411211 (2018).

253. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-871.e8 (2018).

254. Sabarís, G., Laiker, I., Preger-Ben Noon, E. & Frankel, N. Actors with Multiple Roles: Pleiotropic Enhancers and the Paradigm of Enhancer Modularity. *Trends in Genetics* **35**, 423–433 (2019).

255. Sagner, A. *et al. A shared transcriptional code orchestrates temporal patterning of the central nervous system*. *PLoS Biology* **19**, e3001450 (2021).

256. Zhang, T. *et al.* Generation of excitatory and inhibitory neurons from common progenitors via Notch signaling in the cerebellum. *Cell Rep* **35**, 109208 (2021).

257. Seto, Y. *et al.* Temporal identity transition from Purkinje cell progenitors to GABAergic interneuron progenitors in the cerebellum. *Nat Commun* **5**, 3337 (2014).

258. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).

259. Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods* **13**, 845–848 (2016).

260. la Manno, G. *et al.* Molecular architecture of the developing mouse brain. *Nature* **596**, 92–96 (2021).

261. Telley, L. *et al.* Temporal patterning of apical progenitors and their daughter neurons in the developing neocortex. *Science* **364**, eaav2522 (2019).

262. Geirsdottir, L. *et al.* Cross-Species Single-Cell Analysis Reveals Divergence of the Primate Microglia Program. *Cell* **179**, 1609-1622.e16 (2019).

263. Götz, M., Sirko, S., Beckers, J. & Irmler, M. Reactive astrocytes as neural stem or progenitor cells: In vivo lineage, In vitro potential, and Genome-wide expression analysis. *Glia* **63**, 1452–1468 (2015).

264. Volterra, A. & Meldolesi, J. Astrocytes, from brain glue to communication elements: The revolution continues. *Nat Rev Neurosci* **6**, 626–640 (2005).

265. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).

266. Zhu, F. *et al.* The interaction landscape between transcription factors and the nucleosome. *Nature* **562**, 76–81 (2018).

267. Ibarra, I. L. *et al.* Mechanistic insights into transcription factor cooperativity and its impact on protein-phenotype interactions. *Nat Commun* **11**, 124 (2020).

268. Genereux, D. P. *et al.* A comparative genomics multitool for scientific discovery and conservation. *Nature* **587**, 240–245 (2020).

269. Nobrega, M. A., Ovcharenko, I., Afzal, V. & Rubin, E. M. Scanning Human Gene Deserts for Long-Range Enhancers. *Science* **302**, 413 (2003).

270. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**, 199–208 (2015).

271. Yan, X. *et al.* Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers. *Cancer Cell* **28**, 529–540 (2015).

272. Du, Z. *et al.* Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol* **20**, 908–913 (2013).

273. Kumarswamy, R. *et al.* Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure. *Circ Res* **114**, 1569–1575 (2014).

274. de Goede, O. M. *et al.* Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **184**, 2633-2648.e19 (2021).

275. Matsui, M. & Corey, D. R. Non-coding RNAs as drug targets. *Nat Rev Drug Discov* **16**, 167–179 (2017).

276. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).

277. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).

278. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).

279. Calderon, D. *et al.* Landscape of stimulation-responsive chromatin across diverse human immune cells. *Nat Genet* **51**, 1494–1505 (2019).

280. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).

281. Matharu, N. & Ahituv, N. Modulating gene regulation to treat genetic disorders. *Nat Rev Drug Discov* **19**, 757–775 (2020).

282. Mich, J. K. *et al.* Functional enhancer elements drive subclass-selective expression from mouse to primate neocortex. *Cell Rep* **34**, 108754 (2021).

283. Graybuck, L. T. *et al.* Enhancer viruses for combinatorial cell-subclass-specific labeling. *Neuron* **109**, 1449-1464.e13 (2021).

284. Snetkova, V., Pennacchio, L. A., Visel, A. & Dickel, D. E. Perfect and imperfect views of ultraconserved sequences. *Nat Rev Genet* **23,** 182–194 (2021).

285. Snetkova, V. *et al.* Ultraconserved enhancer function does not require perfect sequence conservation. *Nat Genet* **53**, 521–528 (2021).

286. Fish, A., Chen, L. & Capra, J. A. Gene regulatory enhancers with evolutionarily conserved activity aremore pleiotropic than those with species-specific activity. *Genome Biol Evol* **9**, 2615–2625 (2017).

287. Fuqua, T. *et al.* Dense and pleiotropic regulatory information in a developmental enhancer. *Nature* **587**, 235–239 (2020).

288. Kvon, E. Z. *et al.* Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. *Cell* **180**, 1262-1271.e15 (2020).

289. Taskiran, I. I., Spanier, K. I., Christiaens, V. & Mauduit, D. Cell type directed design of synthetic enhancers. *bioRxiv* 2022.07.26.501466 (2022).

290. Xu, Y. *et al.* A single-cell transcriptome atlas of human early embryogenesis. *bioRxiv* 2021.11.30.470583 (2021).

291. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory DNA. *Nature* **603**, 455–463 (2022).

292. Bustamante, C. D. *et al.* Natural selection on protein-coding genes in the human genome. *Nature* **437**, 1153–1157 (2005).

293. Pollard, K. S. *et al.* Forces shaping the fastest evolving regions in the human genome. *PLoS Genet* **2**, 1599–1611 (2006).

294. McLean, C. Y. *et al.* Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216–219 (2011).

295. Vermunt, M. W. *et al.* Epigenomic annotation of gene regulatory alterations during evolution of the primate brain. *Nat Neurosci* **19**, 494–503 (2016).

296. Uebbing, S. *et al.* Massively parallel discovery of human-specific substitutions that alter enhancer activity. *Proc Natl Acad Sci U S A* **118**, e2007049118 (2021).

297. Keough, K. *et al.* Machine-learning dissection of Human Accelerated Regions in primate neurodevelopment. *bioRxiv* 256313 (2022).

298. Haygood, R., Fedrigo, O., Hanson, B., Yokoyama, K. D. & Wray, G. A. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. *Nat Genet* **39**, 1140–1144 (2007).

299. Taylor, M. S. *et al.* Rapidly evolving human promoter regions. *Nat Genet* **40**, 1262–1263 (2008).

300. Liu, J. & Robinson-Rechavi, M. Robust inference of positive selection on regulatory sequences in the human brain. *Sci Adv* **6**, eabc9863 (2020).

301. Wilson, M. D. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434–438 (2008).

302. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: Function, expression and evolution. *Nat Rev Genet* **10**, 252–263 (2009).

303. Fleck, J. S. *et al.* Inferring and perturbing cell fate regulomes in human cerebral organoids. *bioRxiv* 2021.08.24.457460 (2021).

304. Kamal, A. *et al.* GRaNIE and GRaNPA: Inference and evaluation of enhancer-mediated gene regulatory networks applied to study macrophages. *bioRxiv* 2021.12.18.473290 (2021).

305. González-Blas, C. B. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *bioRxiv* 2022.08.19.504505 (2022).

306. Argelaguet, R., Lohoff, T., Li, J. G., Nakhuda, A. & Drage, D. Decoding gene regulation in the mouse embryo using single-cell multi-omics. *bioRxiv* 2022.06.15.496239 (2022).

307. Frank, C. L. *et al.* Regulation of chromatin accessibility and Zic binding at enhancers in the developing cerebellum. *Nat Neurosci* **18**, 647–656 (2015).

308. Zuin, J. *et al.* Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).

309. Blassberg, R. *et al.* Sox2 levels regulate the chromatin occupancy of WNT mediators in epiblast progenitors responsible for vertebrate body formation. *Nat Cell Biol* **24**, 633–644 (2022).

310. Neumayr, C. *et al.* Differential cofactor dependencies define distinct types of human enhancers. *Nature* **606**, 406–413 (2022).

311. Kvon, E. Z. Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics* **106**, 185–192 (2015).

312. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

313. Cooper, Y. A. *et al.* Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science* **377**, eabi8654 (2022).

314. Dutrow, E. V. *et al.* Modeling uniquely human gene regulatory function via targeted humanization of the mouse genome. *Nat Commun* **13**, 304 (2022).

315. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295 (2015).

316. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).

317. Wang, L. *et al.* CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**, e74 (2013).

318. Washietl, S. *et al.* RNAcode: Robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* **17**, 578–594 (2011).

319. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

320. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res* **13**, 2178–2189 (2003).

321. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

322. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

323. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

324. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

325. Wucher, V. *et al.* FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res* **45**, e57 (2017).

326. Hensman, J., Rattray, M. & Lawrence, N. D. Fast nonparametric clustering of structured time-series. *IEEE Trans Pattern Anal Mach Intell* **37**, 383–393 (2015).

327. Hensman, J., Rattray, M. & Lawrence, N. D. Fast variational inference in the conjugate exponential family. in *Advances in Neural Information Processing Systems* **4**, 2888–2896 (2012).

328. Hensman, J., Lawrence, N. D. & Rattray, M. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics* **14**, 252 (2013).

329. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: A more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* **45**, W130–W137 (2017).

330. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).

331. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

332. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

333. Rice, P., Longden, L. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**, 276–277 (2000).

334. Bailey, T. L. Discovering Novel Sequence Motifs with MEME. *Curr Protoc Bioinformatics* doi: 10.1002/0471250953.bi0204s00 (2003).

335. Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and Bioconductor. *Methods in Molecular Biology* **1418**, 335–351 (2016).

336. Trevino, A. E. *et al.* Chromatin accessibility dynamics in a model of human forebrain development. *Science* **367**, eaay1645 (2020).

337. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495–501 (2010).

338. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**, 576–589 (2010).

339. Kumar, L. & Futschik, M. E. Mfuzz: A software package for soft clustering of microarray data. *Bioinformation* **2**, 5–7 (2007).

340. Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J Stat Softw* **25**, 1–18 (2008).

341. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: Enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207 (2010).

342. Karolchik, D. *et al.* The UCSC table browser data retrieval tool. *Nucleic Acids Res* **32**, D493-D496 (2004).

343. Hu, H. *et al.* AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res* **47**, D33–D38 (2019).

344. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, giaa151 (2020).

345. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177**, 1873-1887.e17 (2019).

346. Giorgino, T. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *J Stat Softw* **31,** doi:10.18637/jss.v031.i07 (2009).

347. Lundberg, S. M. & Lee, S. I. A Unified Approach to Interpreting Model Predictions. *arXiv* 1705.07874 (2017).

348. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).

349. Zhang, X., Kaplow, I. M., Wirthlin, M., Park, T. Y. & Pfenning, A. R. HALPER facilitates the identification of regulatory element orthologs across species. *Bioinformatics* **36**, 4339–4340 (2020).

350. Katoh, K., Misawa, K., Kuma, K. I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066 (2002).

351. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).