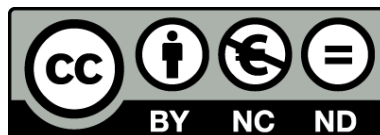




UNIVERSITAT DE
BARCELONA

Quantitative analyses of molecular and cellular features of brain development in the context of recent human evolution

Juan Moriano



Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial – SenseObraDerivada 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial – SinObraDerivada 4.0. España de Creative Commons.**

This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial-NoDerivs 4.0. Spain License.**

Quantitative analyses of molecular and cellular features of brain development in the context of recent human evolution

JUAN MORIANO

Thesis submitted to the
UNIVERSITY OF BARCELONA

in partial fulfillment of the requirements of the degree of

DOCTOR PER LA UB
in Cognitive Science and Language PhD program
Line of Research Lingüística i cognició

Under the supervision and tutorization of:

DR. CEDRIC BOECKX

Facultat de Filologia i Comunicació



September 2023

Abstract

In this thesis, we investigate molecular and cellular substrates of derived traits in our species, and do so under the hypothesis of a *Homo sapiens* species-specific brain ontogenetic trajectory as overarching framework. We exploit advances in the field of paleogenomics (reviewed in chapter 1) to pinpoint mutations that emerged in the *Homo sapiens* lineage after its divergence from the Neanderthal/Denisovan lineage, and advance in the spatiotemporal resolution required to formulate experimentally tractable hypotheses on the evolution of modern human cognitive traits. In chapter 2, we identify evolutionary modifications impacting regulatory regions during human corticogenesis and find regulatory variants associated to candidate genes not previously appreciated, such as those involved in chromatin remodelling, and that significantly extend the focus from the less than one hundred proteins carrying fixed missense mutations derived in our species. In chapter 3, we study developmental mechanisms underlying the expansion and organization of the human neocortex. We perform an integrative analysis of different layers of biological complexity to elucidate transcriptional programs and regulation involved in indirect neurogenesis. We implement a matrix factorization method for single cell transcriptomic analyses that, in combination with gene regulatory network analysis, enable us to resolve a cholesterol metabolic program under the regulation of the zinc-finger transcription factor KLF6, specifically in outer radial glia, a key cell type for neocortical expansion. Furthermore, we computationally infer differential transcription factor binding affinities caused by *Homo* species-specific mutations. We detect instances of positive selection in the *Homo sapiens* lineage in regulatory regions associated to *GLI3*, a key regulator of cortical neurogenesis. The appendix A provides a chronological atlas of derived high-frequency variants in our lineage predicting a recent emergence of *GLI3* regulatory variants. Chapter 3 contributes to the recent evidence on the relevance of metabolic control in the development and evolution of the primate brain. Finally, in chapter 4, we broaden our investigation and undertake a comprehensive transcriptomic analysis of several brain regions across both prenatal and postnatal stages. We particularly focus on genes within special regions of our genome: those under putative positive selection and within deserts of introgression. We identify a dozen of

genes within these special regions and observe distinctive transcriptomic profiles in structures beyond the neocortex. Prominently, we find a Ca^{2+} -dependent activator protein, CADPS2, associated to fixed derived mutations in our lineage, showing a pronounced increase in gene expression in the cerebellum, a critical structure for the derived globular profile of modern human brains.

Resumen

En la presente tesis investigamos los sustratos moleculares y celulares de rasgos derivados en nuestra especie, bajo el marco de la hipótesis de una trayectoria ontogénica del cerebro específica de *Homo sapiens*. Hacemos uso de avances en el campo de la paleogenómica (tratados en el capítulo 1) para identificar mutaciones que emergieron en el linaje de *Homo sapiens* después de la divergencia con el linaje de los Neandertales y Denisovanos, y avanzamos en la resolución espaciotemporal requerida para la formulación de hipótesis testables sobre la evolución de rasgos cognitivos de los humanos modernos. En el capítulo 2, identificamos modificaciones evolutivas en regiones reguladoras activas durante el desarrollo de la corteza cerebral en humanos, y hallamos variantes reguladoras asociadas a genes candidatos que no han sido previamente caracterizados, como genes implicados en la remodelación de la cromatina, extendiendo el foco más allá de aquellas proteínas, no más de cien, que albergan mutaciones de cambio de aminoácido. En el capítulo 3, estudiamos mecanismos del neurodesarrollo implicados en la expansión y la organización de la corteza cerebral humana. Realizamos análisis integradores de diferentes capas de complejidad biológica para elucidar programas transcriptómicos y su regulación durante neurogénesis indirecta. Implementamos un método de factorización de matrices para el análisis de datos transcriptómicos con resolución de una única célula que, combinado con la reconstrucción bioinformática de redes genéticas reguladoras, nos permiten caracterizar un programa metabólico del colesterol bajo la regulación de un factor de transcripción con módulo de zinc, KLF6, específicamente en células de glia radial basal, un tipo celular clave para la expansión de la neocorteza. Además, inferimos computacionalmente cambios de afinidad en la unión de factores de transcripción causados por mutaciones específicas de especies *Homo*. Detectamos casos de selección positiva en *Homo sapiens*, en regiones reguladoras asociadas con el gen *GLI3*, un regulador clave del desarrollo de la corteza cerebral. El apéndice A contiene un atlas cronológico de variantes en alta frecuencia en poblaciones humanas y derivadas en *Homo sapiens*, que predice la aparición reciente de dichas mutaciones asociadas a *GLI3*. El capítulo 3 contribuye a la evidencia creciente de la relevancia del control metabólico en el desarrollo y evolución del cerebro de los primates. Finalmente, en el capítulo 4, ampliamos nuestra investi-

gación para analizar a nivel transcriptómico diversas regiones cerebrales en estadios prenatales y postnatales. Particularmente, nos centramos en genes hallados en regiones especiales en nuestro genoma: regiones bajo selección positiva y en desiertos de introgresión. Identificamos una docena de genes en estas regiones especiales y observamos perfiles transcriptómicos distintivos en estructuras cerebrales más allá de la neocorteza. Prominentemente, encontramos una proteína activadora dependiente de calcio, CADPS2, asociada a mutaciones fijadas en nuestra especie, y que muestra un pronunciado incremento en su expresión génica en el cerebelo, estructura clave para el perfil globular derivado en nuestra especie.

Contents

Abstract	iii
Resumen	v
Contents	vii
List of Figures	x
List of Tables	xiii
Acknowledgments	xv
1 Introduction	1
1.1 Paleogenomics: A window into the genetic basis of derived traits in <i>Homo sapiens</i>	2
1.1.1 The resurrection of ancient <i>Homo</i> genomes	4
A note on low coverage genomes	5
1.1.2 On catalogs and prioritization of variants	6
An hypothesis on the emergence of the modern human cognitive profile	8
1.1.3 Experimental interrogation of the functional relevance of <i>Homo</i> -specific genetic variants	8
A study case: The self-domestication hypothesis	11
1.2 Thesis structure and overview	12
2 Modern human changes in regulatory regions implicated in cortical development	29
2.1 Background	30
2.2 Results	31
2.3 Discussion	34
2.4 Conclusions	35
2.5 Methods	35
2.5.1 Data processing	35

2.5.2	Single-cell RNA-seq analysis	36
2.5.3	Weighted gene co-expression network analysis	36
2.5.4	Enrichment analysis	36
2.6	Supplementary Information	36
3	A multi-layered integrative analysis reveals a cholesterol metabolic program in outer radial glia with implications for human brain evolution	41
3.1	Introduction	42
3.2	Results	44
3.2.1	Inferring neural progenitor states during indirect neurogenesis from the developing human cortex	44
3.2.2	A pseudotime-informed non-negative matrix factorization to identify dynamic gene expression programs	44
	A cholesterol metabolic program activated in the radial glial branch	45
3.2.3	A <i>KLF6</i> centered regulatory network for the activation of a cholesterol metabolism program in human radial glia	45
3.2.4	A paleogenomic interrogation of regulatory regions active during human corticogenesis	46
	Differential transcription factor binding analysis exhibits signals of positive selection in <i>GLI3</i> regulatory islands	47
3.3	Discussion	48
3.4	Methods and Materials	49
3.4.1	Single-cell RNA-seq data processing	49
3.4.2	Gene regulatory network inference and analysis	50
3.4.3	Pseudotime-informed non-negative matrix factorization	50
3.4.4	Paleogenomic analysis	51
3.5	Limitations of this study	52
4	A Brain Region-Specific Expression Profile for Genes Within Large Introgression Deserts and Under Positive Selection in <i>Homo sapiens</i>	67
4.1	Introduction	68
4.2	Results	70
4.2.1	Genes in Large Deserts of Introgression Have Different Expression Levels Relative to the Rest of the Genome	70
4.2.2	The Cerebellar Cortex, the Striatum and the Thalamus Show Divergent Transcriptomic Profiles When Considering Genes Within Large Deserts of Introgression and Under Positive Selection	70

Contents	ix
----------	----

4.2.3	Gene-specific Expression Trajectories of Genes in the Overlapping Desertic and Positively-Selected Regions	73
4.3	Discussion	75
4.4	Methods	76
4.4.1	mRNA-seq analysis	76
4.4.2	Gene-specific expression trajectories	76
	Supplementary Material	79

Appendices

A	Temporal mapping of derived high-frequency gene variants supports the mosaic nature of the evolution of <i>Homo sapiens</i>	95
A.1	Results	97
A.1.1	Variant subset distributions	98
A.1.2	Gene Ontology analysis across temporal windows	100
A.1.3	Gene expression predictions	100
A.2	Discussion	101
A.3	Methods	101
A.3.1	<i>Homo sapiens</i> variant catalog	101
A.3.2	GEVA	102
A.3.3	ExPecto	102
A.3.4	gProfiler2	102
	Supplementary Figures	105

List of publications	119
-----------------------------	------------

List of Figures

2.1	Regulatory regions characterized in this study	31
2.2	Cell-type populations at early stages of cortical development	33
2.3	Progenitor cell-fate decisions shaped by WNT/B-CATENIN signaling	35
3.1	Resolving the tree of neural progenitor cell differentiation during human corticogenesis	53
3.2	Pseudotime-informed non-negative matrix factorization recovers a sequential activation of gene expression programs	54
3.3	Gene regulatory network reconstruction from human neural progenitor single-cell data	55
3.4	Paleogenomic analysis of regulatory variants	56
3.5	Figure 3.1. supplement 1	62
3.6	Figure 3.2 supplement 1	63
3.7	Figure 3.2 supplement 2	64
3.8	Figure 3.3 supplement 1	65
3.9	Figure 3.3 supplement 2	66
4.1	Study outline	70
4.2	Median expression of genes within large deserts	71
4.3	Median expression of genes under putative positive selection within large deserts	72
4.4	Genes within large deserts	73
4.5	Genes under positive selection within large deserts	74
4.6	Gene-specific trajectories	75
4.7	Median expression profile of genes within large deserts, deserts/positively-selected regions and the global dataset, across structures and stages	80
4.8	The cerebellum’s transcriptomic profile significantly diverges at post-natal stages	81
4.9	The transcriptomic profile of the mediodorsal nucleus of the thalamus significantly diverges at fetal stage 1 for genes within deserts under positive selection	81

4.10	The striatum's transcriptomic profile significantly diverges at adolescence for genes within large deserts	82
4.11	Evaluation of transcriptomic divergence considering genes within deserts for chromosome 1	83
4.12	Evaluation of transcriptomic divergence considering genes within deserts for chromosome 3	84
4.13	Evaluation of transcriptomic divergence considering genes within deserts for chromosome 7	85
4.14	Evaluation of transcriptomic divergence considering genes within deserts for chromosome 8	86
4.15	Evaluation of transcriptomic divergence considering genes under positive selection not within large deserts of introgression	87
4.16	Global profile of genes across stages and structures	88
4.17	Visualization of a pairwise post-hoc Tukey test	89
4.18	Decomposition of median expression profile of genes within large deserts per chromosome	90
4.19	Expression profile of genes within large deserts and positively selected overlapping regions	91
A.1	Figure 1	97
A.2	Figure 2	98
A.3	Figure 3	99
A.4	Figure 4	100
A.5	K -means clustering analysis of HF variant temporal distribution . . .	106
A.6	Density distribution of derived Homo sapiens alleles for different subsets used in this study	106
A.7	emporal distribution of introgressed variants linked to phenotypes . .	107
A.8	Temporal distribution of variants shared with each of the extinct human genomes after applying specific population frequency filters . .	108
A.9	Temporal distribution of HF variants in two genes highlighted in early discussions of selective sweeps	109
A.10	Temporal distribution of specific variants	110
A.11	Temporal distribution of variants associated with <i>CADPS2</i>	111
A.12	Temporal distribution of variants in genes found in putative positively-selected genetic windows before early Homo sapiens population divergence	112
A.13	Temporal distribution of high-frequency missense and regulatory variants	113
A.14	GO terms results when thresholding by an adjusted p-value of 0.05 . .	114
A.15	GO terms Reduction of shared terms across time windows	115

A.16	Quantile-quantile plots of predicted expression values of high-frequency variants	116
A.17	Violin plots per time window for 22 brain and brain-related tissues	117

List of Tables

4.1	Genomic coordinates used in this study	69
A.1	Big40 Brain volume GWAS top hits	101

Acknowledgments

I am extremely happy to express here my gratitude to many people that contributed to one the most important stages of my life.

I thank my PhD supervisor Cedric Boeckx for always providing me timely advice and relentless attention. Under his supervision, I have felt privileged to enjoy the freedom to pursue my scientific interests and independence; this was already true when I first contacted him as an undergraduate student. I owe great part of the work of this thesis, and many lessons learned, to him.

During this journey I have been accompanied by great scientists and people. I am in debt to members of the Cognitive Biology of Language group Alejandro Andirkó, Pedro T. Martins, Stefanie Sturm and Thomas O'Rourke. Their support and generosity, talents and interests have inspired me very much, and some of my fondest memories in Barcelona are with them. I thank as well visiting students that enriched us all: Raül Buisán, Mireia Rumbo, Sara Silvente and Lucia Troiani. Also, I was lucky to be part of a Department with supportive colleagues, particularly Faustino Diéguez, Mercé Guisado, María Victoria Novo and Mariona Taulé. I am likewise grateful to our university Library services for all the generous help provided.

Collaborations with exceptional scientists have been a source of great joy during my PhD. I have learned a lot from Susanna Balcells, Daniel Grinberg and Núria Martínez-Gil during my stay at the Department of Genetics of our University. The excitement and deep questions on human evolution transpired every meeting with colleagues from Prof. Giuseppe Testa's group; I am particularly thankful to Davide Aprile, Oliviero Leonardi and Alessandro Vitriolo. I am very grateful as well to Prof. Arnold Kriegstein and his wonderful team, for sharing with me their knowledge and passion since the very first day; they also gifted me unforgettable experiences in San Francisco. Additionally, with great pleasure I have been part of the Institute of Complex Systems student community and the preLights initiative, where I have enjoyed very much being surrounded by very diverse and motivated early career researchers.

It has been truly special to share these years with people whose friendship has nurtured my life in so many different ways. I am forever grateful to the adventurous Carlos and Ángel, for their goodness and their unmatched curiosity, and to the special

kind of magic distilled by Abbi, Valentina, Carlos, Mireia, Simone and Lorena, Ted and Leah, Miguel and Cristian.

Finally, my deepest gratitude to my family, who guide me to the important.

And to my love Adaleen, whose voice and heart transform all that is around.

Chapter 1

Introduction

Homo sapiens are characterized by a gracile skeleton and a rounded cranium. In comparison to our closest relatives, the Neanderthals and Denisovans, anatomically modern humans possess a narrow trunk and pelvis, reduced prognathism, small teeth, and a high and globular neurocranium [1]. The complete suit of anatomical traits characteristic of our species plausibly resulted from the coalescence of heterogeneous populations distributed in Africa rather than from single population emerging at a specific time point and geographic area [2]. Likewise, our species behavioral modernity might not be considered as a monolithic entity but rather the consequence of a multi-factorial gradual process [3]. Recent progress in the research field of human evolution challenges past, more simplistic assumptions about the origin, mode of evolution and nature of *Homo sapiens*, more so with the increasing recognition of the rich and complex behavior of our closet relatives [4].

One of the most remarkable recent breakthroughs in our quest to understand what is quintessential to our species pertain to the ability to reconstruct the genomes of extinct species. The sequencing of the genomes of Neanderthals and Denisovans provides a level of analysis not allowed decades ago, and while the sole sequencing of DNA from extinct *Homo* species does not provide direct evidence for distinctive behaviors, it dramatically empowers the kind of inferences can be made from the examination of the fossil record. This thesis exploits the power of paleogenomics in helping to address the genotype-phenotype mapping problem for the study of the evolution of the human brain, and particularly makes advancements on a hypothesis that links our species-specific brain ontogenetic trajectory to our modern cognitive profile. This hypothesis [5] pertains to the evidence of a perinatal globularization phase, absent in our closest extinct and extant relatives, that results in a neurocranium with steep frontal, expanded parietal and cerebellar areas - and likely acquired at some point after 100 thousand years ago (kya) [6]. In the present chapter we review how paleogenomics is transforming our comprehension of human evolution, with

special attention to the granularity needed to investigate evolutionary modifications that might have impacted the development of the human brain, the substrate of our cognitive abilities. We finally present an overview of the ensuing chapters, where we seek a high spatiotemporal resolution for the investigation of modern human-specific features of brain development and evolution.

1.1 Paleogenomics: A window into the genetic basis of derived traits in *Homo sapiens*

The absence of fossilized brain tissue in the archaeological record significantly hinders inferences about the neurobiological underpinnings of defining cognitive traits of our species, such as language. In a landmark publication in 1997, mitochondrial DNA was retrieved from a 0.4g sample of the right humerus bone of a Neanderthal fossil in the Feldhofer cave (Neander Valley, Germany) [7]. Almost 15 years later, a new hominin species was discovered solely through paleogenomic analysis from extracts of a finger bone found in the Denisova cave in Siberia [8]. What was described once as the emerging field of molecular archeology [9] is now a burgeoning interdisciplinary field dramatically expanding the range of questions researchers can ask to understand the so-called human condition [10]. The retrieval of ancient DNA, defined as extremely short DNA fragments preserved in the fossil record and extracted from material such as bones, teeth, skin, hair or sediments [11], has enabled direct comparisons of the genome of our species to that of our closest extinct relatives, the Neanderthals and Denisovans. Paleogenomics provides a unique entry point to decipher the genetic basis of derived traits in *Homo sapiens*. However, connecting mutations in a DNA sequence to specific molecular, cellular and higher phenotypic levels remains a fundamental problem in biology, and one that requires insights from multiple disciplines, with genetics being only one of them. Additionally, as particularly evident in the field of the neurobiology of language, efforts attempting to resolve the biological underpinnings of traits such as the human faculty of language require the right level of granularity across research domains and explanatory linking hypotheses [12].

A paradigmatic case, highlighting the potential of paleogenomic research, is provided by the study of the forkhead box protein P2 FOXP2. This transcription factor, unequivocally associated to key aspects of (spoken) language, was discovered after the characterization of a three-generation family where around half of the members, with severe linguistic impairments, were found to carry a heterozygous missense mutation in the FOXP2 gene [13]. Experimental research in a variety of model systems have significantly advanced our understanding of the functional roles of FOXP2. The introduction of the arginine-to-histidine substitution present in the affected KE family members in the homolog gene in mice was reported to impair ultrasonic vocalizations

with less complex sequences [14], whereas its knockdown in zebra finches was found to alter dopaminergic signaling and song variability [15]. Additionally, brain imaging studies have revealed that affected members of the KE family present abnormalities in brain structures beyond the neocortex, implicating subcortical regions such as the caudate nucleus or the putamen [16, 17]. In light of its possible key roles for the emergence of language in our species, research studies focused on the evolution of the *FOXP2* gene, and found that despite its high sequence conservation, the human gene harbors two non-synonymous mutations (both in exon 7) that emerged after the last common ancestor with the chimpanzees [18, 19]. Later paleogenomics analyses on ancient DNA from Neanderthal remains found at El Sidrón Cave in Asturias (Spain) revealed that these two amino acid changes were shared with the Neanderthals, concluding that most likely those substitutions were present in the last common ancestor of *Homo sapiens* and the Neanderthals [20]. It is expected that most of the genetic differences distinguishing us from our closest relatives likely emerged by drift and do not exert any differential functional roles, but those that were under selective pressures in our lineage not only affected coding regions but also impacted the regulatory landscape that controls gene expression in a spatiotemporal precise manner. Adding evidence to the evolving story of *FOXP2* [21], a derived variant in the *Homo sapiens* lineage located in the intron 8 of *FOXP2* was shown to cause differential transcription factor binding of POU3F2 [22]. The functional consequences of this change remains unclear, and a recent study found that the intronic region is transcribed and reported no evidence of positive selection in the locus [23]. As revealed by the *FOXP2* case, ancient DNA research can illuminate inquiries into what sets our species apart from our closest extinct relatives going beyond comparisons with non-human primates, and clearly exemplifies the need to bring together diverse analytical and experimental approaches to explore the questions about our species cognitive abilities. But just like there is no gene for language, there is no gene for language evolution, and no single mutation will very likely be able to explain changes to our language-ready brain [24].

Evidence for the functional consequences of derived variants in *Homo sapiens* when compared to our closest relatives implicate diverse systems such as the skin, the immune system, the metabolism or the skeleton. Likewise, some of these variants are expected to have impacted the nervous system, and here we focus on the use of paleogenomics to reconstruct ancient phenotypes and approaches to identify and prioritize candidates for experimental interrogation. The emergence of the modern human condition, and the evolution of the ‘modern’ faculty of language in particular, was plausibly accompanied by the gradual accumulation of genetic mutations rather than the result from a single mutation, and the efforts now propelled by the field of paleogenomics urge a comprehensive characterization of the genetic mutations that were under positive selection in our lineage. Of note, the quantitative revolution in the field of developmental neuroscience [25] closely aligns to the efforts to overcome the

genotype-to-phenotype problem presented above, and a final subsection is dedicated to examples on how concrete hypotheses facing this mapping problem in explaining the evolution of modern human traits can be experimentally tested.

1.1.1 The resurrection of ancient *Homo* genomes

The degradation and chemical modification of ancient DNA material, along with the contamination from microbial and human sources, pose severe challenges for the reconstruction of the genomes of ancient species. For instance, early insights obtained from high-throughput sequencing of nuclear Neanderthal genomes were later shown significantly contaminated by non-endogenous DNA [26, 27]. Advancements in the purification, amplification and high-throughput sequencing of ancient DNA now allows to confidently reconstruct at high-coverage the genomes of our closest relatives with a quality comparable to that obtained from present-day humans. Comparative genomics using ancient *Homo* genomes has been fostered likewise by a wider, more diverse characterization of genomes from present-day human populations across continents, thanks to efforts like the 1000 Genomes Project [28] or the Simons Genome Diversity Project [29], as well as by the genomes of other primate species [30]. At present, four high-coverage archaic genomes have been obtained at exceptional quality from remains belonging to both the Neanderthals and Denisovan species. In 2012, the genome of a female Denisovan individual was assembled utilizing DNA extracts from a finger phalanx from the Denisova Cave [31]. The ancient DNA from this Denisovan individual was obtained at a 30-fold coverage by employing a single-stranded DNA library preparation, which significantly enhanced the final sequencing yield compared to previous efforts [31]. Two years later, the genome of a female Neanderthal was reconstructed at a 52-fold coverage, also from the Denisova cave in the Altai Mountains [32]. This first high-quality genome from a Neanderthal provided estimates of population split of Neanderthal/Denisovans from the *Homo sapiens* lineage between 550k-765kya, and aided in the depiction of the complex evolutionary history of our ancestors, estimating the level of Neanderthal introgressed DNA material in current modern human populations at around 1.5–2.1%, as well as detecting instances of Neanderthal introgression into Denisovans [32]. The first two high-coverage genomes led to the first comprehensive catalog of single nucleotide variants and short insertions and deletions that distinguish our species since our divergence from the Neanderthal/Denisovan lineage. As a result, a small set of proteins (less than one hundred) were identified carrying fixed amino acid substitutions along with a set of regulatory variants derived in our species where the Neanderthals/Denisovans carry the ancestral allele found in non-human primates. Proteins carrying non-synonymous substitutions derived in our lineage were found related to axonal and dendritic growth and synaptic transmission and, notably, an over-

representation of genes relevant for neural progenitor cell behavior was reported [32]. Two additional Neanderthal genomes have been reconstructed over the past decade. A second genome of female Neanderthal was sequenced at 30-fold coverage from remains at the Vindija cave [33] and a third Neanderthal genome was uncovered from remains of Chagyrskaya Cave in the Altai Mountains, sequenced at a 27-fold genomic coverage [34]. Despite the few genomes currently available, the three Neanderthal genomes at high-quality enabled to more confidently elaborate a catalog of derived genetic variants in the Neanderthal lineage, revealing that genes expressed in the striatum show derived features in the Neanderthal lineage, while some of the fixed changes within genes expressed in the striatum might have been under negative selection in anatomically modern humans [34].

A note on low coverage genomes

While the sequencing of genomes from extinct *Homo* species at high-quality represents a turning point in the field of human evolution, the importance of low coverage ancient genomes cannot be underestimated. This is perhaps best exemplified by the first drafts of the genomes of both the Neanderthals and the Denisovans at a coverage below 2x [8, 35]. A milestone of ancient DNA in human evolution research was achieved in 2010, when a first draft of a Neanderthal genome from three different individuals at the Vindija cave (Croatia) was presented at a coverage of 1.3x [35]. These first analyses led to the proposal of gene flow from Neanderthals to non-African modern human populations on the basis of Neanderthals sharing more alleles with non-African individuals than African individuals, and it unearthed candidate genes under putative positive selection linked to cognitive aspects of our species [35]. Furthermore, increasing the geographical and temporal resolution of ancient genomes will be pivotal to obtain a more faithful representation of the social structure and history of extinct *Homo* populations. Recently, sodium hypochlorite treatment on bone and tooth powder was used to drastically increase the recovery of ancient DNA from numerous samples with low endogenous DNA content or substantial DNA contamination [36]. Applying this strategy on five different samples of late Neanderthals from fossils distributed across western Eurasia, it was possible to generate single-stranded DNA libraries with an average coverage between 1-fold and 2.7-fold [36]. Paleogenomic analyses on both low and high quality Neanderthal genomes estimated the split times within Neanderthal population structure and found evidence for a population turnover during the last period of Neanderthals presence in western Eurasia [36].

1.1.2 On catalogs and prioritization of variants

Paleogenomic analyses have revealed that only a small set of proteins, less than one hundred, carry fixed non-synonymous substitutions derived in *Homo sapiens*. With increasing efforts to capture the full panoply of genomic diversity in present-day populations [28, 29], the interrogation of nearly fixed substitutions in modern human populations opens up further possibilities. Examining extant variation across present-day human genomes is required to uncover nearly fixed derived variants in *Homo sapiens* associated to phenotypes of clinical relevance, or to detect haplotypes from past interbreeding events that were retained due to its adaptive value. Kuhlwilm and Boeckx [37] stressed the relevance of such an approach revising the list of genes with protein-altering fixed changes, as well as extending the set of proteins with non-synonymous mutations to 571, considering high-frequency substitutions present in at least 90% in present-day humans. In addition, the number of fixed or nearly fixed substitutions falling within non-coding regions largely exceeds protein-altering mutations and therefore might be equally relevant to explain the genetic basis of derived traits in our species, as already suggested decades ago in the comparison with chimpanzees [38]. A significant challenge in the paleogenomic field is to narrow down the list of genetic differences among *Homo* species to those variants that were not the result of drift and represent best candidates for experimental testing.

As highlighted above, the first incursions into the genomes of extinct *Homo* species revealed admixture events between our species and the Neanderthals and Denisovans. Ancient DNA introgression present in modern human populations outside Africa are estimated at around 2% content, while gene flow from Denisovan populations have been detected in present-day individuals from Oceania, Southeast Asia and to a much lesser extent in East and South Asian and Native American individuals [39]. While most of these introgressed variants are predicted to be neutral, some might have been under positive or negative selection in our lineage. The functional dissection of some of these variants have already provided evidence for effects on different aspects of modern human biology, for instance skin pigmentation, immunity or metabolism [40]. A quarter of introgressed variants from the Neanderthals are predicted to impact gene expression regulation [41]; notably, in a comprehensive analysis on more than 50 tissues and 200 individuals, a downregulation of Neanderthal alleles in the brain (more prominently, the cerebellum and the basal ganglia) was found [41]. Additionally, on the basis of two high-coverage Neandertal genomes, it was reported that introgressed alleles were likely under negative selection in promoters and noncoding conserved genomic regions [42].

Complementary, genomic regions that are depleted of Neanderthal ancestry are equally relevant to reveal the biological underpinnings of derived traits in our species. Indeed, the unequal distribution of introgressed alleles in modern human genomes

point to selection against Neanderthal variants, possibly soon after the admixture event [43, 44]. A recent study identified uniquely large genomic regions significantly depleted of introgressed Neanderthal haplotypes [45], the so-called deserts of introgression, distributed across only four chromosomes (partially overlapping previous independent research [46]). Intriguingly, one desert corresponds to a 17Mbp in chromosome 7 where *FOXP2* is found, and likely under strong purifying selection in *Homo sapiens* but not in other populations [47].

Since the split from the Neanderthal/Denisovan lineage, some variants reached high-frequency or even fixation in our lineage and can be found in unusually long genomic regions, suggesting instances of positive selection [48, 49]. A set of 314 genomic regions were identified as putative positively-selected regions in the *Homo sapiens* lineage, and were found enriched in regulatory regions annotated as enhancers and promoters [49]. Also, some of these regions overlap protein-coding genes carrying fixed aminoacid substitutions; this is for example the case of *ADSL*, where evidence for functional consequences has been found (see below).

Another powerful avenue of research to address the genotype-to-phenotype mapping problem builds on biomedical databases that gather large-scale neuroimaging and genomics data, enhancing genome-wide association studies (GWAS) that connect variants to neuroanatomical traits in a statistically significant manner. In a landmark study, neuroimaging genomic data from around 30,000 participants (of European ancestry) from the UK Biobank allowed the identification of polymorphic sites previously annotated as of relevance for *Homo sapiens* brain evolution, with a focus on heritability in specific measures of cortical area size and white matter connectivity [50]. Variants within human-gained enhancers (compared to non-human primates) show enriched heritability in specific cortical areas, prominently Broca's area; additionally, variants within deserts of introgression showed a depletion of heritability in the uncinate fasciculus [50], results that however remain difficult to interpret functionally. In line with the expectation that *Homo* species-specific differences in brain organization affect regions beyond neocortical areas, genes within the aforementioned deserts of introgression were found over-represented in subcortical areas such as the striatum [46]. In another prominent study based on neuroimaging genomic data, a measure of endocranial globularity was established from MRI scans of thousands of present-day individuals (mostly from Europe) and revealed that introgressed Neanderthal variants in two genomic regions were statistically associated to reduced endocranial globularity [51]. Some of these variants were predicted to impact gene expression regulation, and more significantly the expression of *UBR4* in the putamen and *PHLPP1* in the cerebellum, perhaps influencing neurogenesis and myelination, respectively [51].

An hypothesis on the emergence of the modern human cognitive profile

The unprecedented spatiotemporal resolution now afforded by high-throughput sequencing technologies and experimental models recapitulating key aspects of human brain development (see below) make feasible to begin formulating explanatory hypothesis linking *Homo* species-specific genotypes and phenotypic differences. In this context, it is worth mentioning the evidence supporting a *Homo sapiens* species-specific brain ontogenetic trajectory connected to one of the most distinctive anatomical traits of *Homo sapiens*, absent in our closest relatives: a globular neurocranium.

Despite the lack of preservation of brain tissue in the fossil record, three dimensional reconstructions based on computed-tomographic scans of endocasts serve as proxy for the external morphology of the brain. Comparative morphometric analyses on endocasts from *Homo sapiens*, chimpanzees and Neanderthals, encompassing different developmental stages (from birth to adulthood), have revealed a globularization phase in the *Homo sapiens* lineage before birth or during the first postnatal year that results in a steep frontal and markedly expanded cerebellar and parietal regions [52, 53, 54]. Both chimpanzees and Neanderthals were inferred to follow the ancestral pattern of brain growth [52, 53, 54] (but see [55]). Differential expansion of neocortical areas is hypothesized to have impacted the neural substrates required for the implementation of defining cognitive abilities of our species, such as language and the Broca's area functional anterior-posterior specializations [56]. Because the globularization phase concerns a key period for neural connectivity, it has been proposed that this *Homo sapiens*-specific ontogenetic trajectory could have impacted the assembly of neural circuits that ultimately possibilitated the emergence of language in our species, in particular impacting a fronto-parieto-temporal connection involving Broca's and Wernicke's area and functionally related to language-related recursive capacities [5, 57]. This line of research shifts the focus from size (where Neanderthals fall within or above present day human variation) to shape, as well as it encourages to consider other regions beyond the neocortex. For instance, the cerebellum, whose derived status in modern humans acquires special prominence in the context of the evolution of the faculty of language, as the dominant cerebellar hemisphere for language has been found significantly enlarged in *Homo sapiens* when compared to the Neanderthals [58]. Advances in modelling the early stages of brain development *in vitro* are critical to formulate, and test, experimentally tractable hypotheses on the evolution of modern human cognition.

1.1.3 Experimental interrogation of the functional relevance of *Homo*-specific genetic variants

The restricted accessibility to human brain tissue for experimentation stands as a significant obstacle for the elucidation of developmental and evolutionary aspects

of our species brain, specially those that might not be faithfully recapitulated with the use of animal models. In recent years new technologies have been developed to model aspects of nervous system development and function. Prominently among these, brain organoid models stand out as powerful systems to mimic the intricate three-dimensional arrangement and cellular and molecular complexities of dynamic tissues such as the developing human brain, generally with the use of induced pluripotent stem cells (iPSC), including from non-human primate species [59]. Numerous protocols are currently available and being developed to produce unguided neural organoids, or guided organoids to recapitulate developmental processes from specific regions such as the cerebral cortex or the spinal cord [60]. Species-specific *in vitro* models of cell types characterized through single-cell sequencing and functionally interrogated with genetic perturbation tools such as CRISPR–Cas systems or massively parallel reporter assays (MPRAs) are now part of fruitful strategies to reconstruct ancient phenotypes and evaluate their functional implications [61, 62]. The combination of these strategies with machine learning-based predictive tools are likewise key for the formulation of testable hypotheses on the evolutionary consequences of species-specific mutations [63]; in fact, it is now possible to comparatively evaluate sequence divergence informed by paleogenomics and reconstruct past regulatory scenarios such as alternative splicing or chromatin folding [64, 65].

To elucidate relevant molecular and cellular processes that might have contributed to brain ontogenetic divergences before the split of *Homo sapiens* and the Neanderthal and Denisovan lineages, significant efforts have been devoted to identify mutations impacting neurodevelopmental mechanisms of cortical expansion, seeking to explain the three-time fold increase in brain size in humans in comparisons to our closest living relatives. Research using iPSC-derived organoids point to developmental timing as key trait explaining species differences in neural progenitor cell behavior [66, 67, 68]. For instance, a prolonged prometaphase-metaphase in apical progenitors in humans in comparison to other great apes that is associated to an increased proliferative capacity [66]. Another developmental process possibly linked to the dramatic increase in neocortical surface in the *Homo*-lineage pertains to a delayed transition between two progenitor cell types at early neurodevelopmental stages: Comparative brain organoid modelling across primate species revealed a delayed transition of neuroepithelial cells to radial glia cells (two types of progenitor cells found in the germinal zones during brain development) in humans, predicting a 1.9-fold increase in the number of progenitors and neurons in comparison to other apes [69]. A zinc-finger transcription factor, ZEB2, was reported as a plausible critical regulator of this transition, its expression peaking 5 days later in human-derived organoids than in gorilla's [69]. Intriguingly, ZEB2 has been linked to human accelerated regions, raising the question whether human accelerated-regions contribute to its differential regulation finally impacting cerebral cortex expansion [69]. Comparative genomics based

on both brain primary tissue and organoid models across primates species have also provided a catalog of genes that show divergent expression profiles during the early stages of neurodevelopment [70, 71].

How has paleogenomic research contributed to the experimental interrogation of the genetic basis of traits derived in *Homo sapiens*? First, inquiries into the genomes from our closest extinct relatives revealed that a substantial proportion of derived missense variants in our species are associated to genes implicated in the neural progenitor cell division [32]. One of these missense variants corresponds to a lysine-to-arginine substitution that distinguishes *Homo sapiens* and Neanderthals *TKTL1* gene. Using a variety of experimental approaches in human primary tissue, mice, ferret and brain organoids, the *TKTL1* non-synonymous mutation was mechanistically linked to higher phenotypic levels: the *Homo sapiens* derived allele, and not the Neanderthal counterpart, impacts metabolic pathways involved in fatty acid synthesis and selectively affects a specific neural progenitor cell type (outer radial glial cells), amplifying its proliferative capabilities with a concomitant increase in cortical neurogenesis [72]. While the functional effects of the mutation were constrained to the frontal lobe, how to interpret these results in light of the known phenotypic differences in brain organization as inferred from the fossil record remains unclear. Another metabolic pathway impacted in recent human evolution pertains to a single nucleotide variant derived in the *Homo sapiens* lineage causing an alanine-to-valine substitution in the adenylosuccinate lyase enzyme *ADSL* [73]. In comparison to the Neanderthals protein version, the modern human *ADSL* protein is less stable resulting in a decreased synthesis of purines, more prominently in the brain [73].

Another set of genes brought by paleogenomic studies are those related to the spindle apparatus and the kinetochore. In fact, spindle-associated genes, required for the correct placement and segregation of chromosomes, accumulated an excess of missense mutations in our lineage [74]. Contrasting the functional effects of modern human and Neanderthal mutations on three spindle-associated genes in mice and brain organoid models, it was discovered that the modern human-specific allele versions in two genes, *KIF18A* and *KNL1*, cause a longer metaphase with few chromosome segregation errors in neural progenitors when compared to the Neanderthal alleles [75]. Similar experimental manipulations on a third gene, *SPAG5*, revealed no significant functional impact caused by the derived allele in *Homo sapiens*, although it was detected in a genomic region that shows signals of positive selection in modern humans [72, 74]. The genomic divergence observed in spindle genes during recent human evolution also provides more lessons on our shared history with Neanderthals. Substitutions in one gene, the kinetochore scaffold 1 gene *KNL1*, previously thought to be unique to the modern human lineage, were later found to be also present in some Neanderthal individuals, likely as the result of interbreeding with ancestral modern human populations after 265kya [49].

A study case: The self-domestication hypothesis

Other developmental mechanisms and systems have been modified over the course of *Homo sapiens* evolution. The self-domestication hypothesis refers to the idea that certain aspects of the behavioral profile of our species, such as a marked reduction in reactive aggression and a distinctively pro-social phenotype, are linked to phenotypic traits reminiscent to that of domesticated species when compared to their wild counterparts. The traits that encompass the so-called domestication syndrome, to varied degree and asymmetrical distribution across species, include reduced body and brain size, attenuation of sex differences, flattened and smaller faces, or depigmentation [76]. The self-domestication hypothesis states that the evolutionary trajectory leading to our species domestication enabled the emergence of evolved traits characteristic of our species, and perhaps most significantly, of language [77, 78, 79, 80, 81]. Key progress to overcome the genotype-phenotype mapping problem was achieved with the proposition that the domestication syndrome arises from a diminished production and migration of neural crest cells (cell types responsible for the generation of disparate types of tissues, such as the facial bones or the adrenal glands) [76]. This provided a mechanistic explanation whose predictions are experimentally tractable. The self-domestication hypothesis has equally benefited from the increasing number of high coverage genomes from extant and extinct species, which now afford the interrogation of the molecular underpinnings of the hypothesized mild neurocristopathy affecting domesticated animals comparatively across a wide range of species. In recent work, an experimental design was put forward to test whether the transcriptional regulatory networks involved in neural crest cells proliferation and differentiation and affected in human neurocristopathies were under selection in anatomically modern humans [82]. This design involved the use a large cohort of iPSC lines derived from patients clinically diagnosed with Williams-Beuren syndrome and Williams-Beuren region duplication syndrome, as a result of copy number variations at 7q11.23. Overall, the molecular and cellular consequences of *BAZ1B* dosage imbalance, a master regulator of neural crest cells, were dissected and evaluated in light of the phenotypic consequences that bear direct resemblance to key traits in anatomically modern humans: a retracted face and prosocial behavior. The paleogenomic inquiry of the regulatory circuits underlying neural crest cell functioning in these models revealed that the *BAZ1B* regulatory network linked to neurocristopathic facial dysmorphisms underwent evolutionary modifications specific to anatomically modern humans, providing in this way evidence for the molecular correlates behind the craniofacial traits at the core of the self-domestication hypothesis [82]. This approach exemplifies how neurodisease modelling in combination with paleogenomic analyses can be used to experimentally test long standing hypothesis on the evolution of modern humans.

1.2 Thesis structure and overview

Paleogenomic studies have significantly contributed to a better comprehension of the complex history and nature of our species. As part of necessarily interdisciplinary approaches, paleogenomics serves as a powerful entry point where to map genetic changes of evolutionary relevance to molecular and cellular mechanisms implicated in developmental processes responsible for the growth and organization of the human brain.

The generation, proliferation and modes of division of neural progenitor cells are fundamental mechanisms contributing to species-specific brain ontogenetic trajectories. Chapter 2 serves as an initial exploration of the rich catalog developed by [37], that identified *Homo sapiens*-derived alleles where the Neanderthals and Denisovans carry the ancestral version found in chimpanzees. Analyses in chapter 2 extend the previous catalog to consider regulatory variants in enhancers and promoters active during human corticogenesis and lead us to highlight regulators of chromatin dynamics in neural cells as targets of recent selection in the *Homo sapiens* lineage. This study prompted us to investigate in depth regulatory variants that emerged in our lineage in light of hypotheses on mechanisms thought to underlie human brain organization [83, 84], and to exploit recent advances in single-cell profiling of cortical progenitors from the developing human brain. In chapter 3, we first succinctly overview main features of human neocortical development, with a focus on the diversity of neural progenitors and the process of indirect neurogenesis. We then investigate the gene expression programs that unfold during progenitor cell differentiation and inquiry whether positive selection has acted on the regulatory control of such programs. To do so, we take advantage of recent studies characterizing human neocortical development via single-cell high-throughput profiling to capture the intermediate gene expression states that unfold as ventricular radial glia differentiate into basal progenitors. In order to account for the modular nature of gene expression programs during differentiation processes, we implement a linear dimensionality reduction technique known as non-negative matrix factorization, specifically a version tailored to capture continuous signals and originally developed by [85]. This methodological approach, shared as open source code, contributes to the efforts to implement solutions aimed at capturing complex gene expression patterns from transcriptomic data, as one of the core challenges in the field of single cell analysis [86]. The combination of gene regulatory network reconstruction (via [87]) and non-negative matrix factorization enables us to uncover a zinc-finger transcription factor, KLF6, as a putative master regulator of cholesterol biosynthesis specifically in the route leading to outer radial glia. The regulatory roles of metabolic programs in neural progenitor cell behavior are now recognized as fundamental for the development and the evolution of the primate brain [68, 88], and we provide a novel research direction under the light of

paleogenetics. We generate an atlas of open chromatin regions allowing the paleogenomic interrogation of the regulatory regions active during human cortical neurogenesis. Prominently among the identified high-frequency regulatory variants derived in *Homo sapiens*, we detect signals of positive selection around *GLI3* regulatory regions, further informed by a computational inference of transcription factor binding site disruptions.

In chapter 4, we extend our investigations beyond the neocortex and perform a comparative transcriptomic analysis encompassing several brain regions and covering both prenatal and postnatal developmental stages. We exploit the spatiotemporal resolution afforded by transcriptomic datasets available from the PsychENCODE Consortium [89, 90] to profile the expression dynamics of genes within very special regions of the *Homo sapiens* genome: regions within deserts of introgression and under positive selection. We find only a dozen of genes, and characterize their expression patterns across brain regions and developmental time windows. We identify with a marked distinctive expression profile a Ca^{2+} -dependent activator protein, *CADPS2*, which carries several fixed derived mutations in our lineage and shows a pronounced increase in expression around birth and infancy stages, peaking into adulthood, in the cerebellum. Given the prominence of the cerebellum in the derived globular profile of modern human brains, *CADPS2* stands out as a promising candidate for future experimental testing. This chapter contributes to the evidence of possible brain region-specific molecular mechanisms that might have been under differential regulation in *Homo* species, particularly during the globularization phase hypothesized to be unique to our species lineage.

Concluding this thesis, appendix A) is devoted to a chronological atlas of high frequency variants derived in our species. Our atlas sheds light into evolutionary processes shaping our species history such as admixture events or instances of positive selection. Bioinformatic analyses on regulatory variant-gene associations reveal gene ontology functional categories unique to evolutionary-relevant temporal windows, supporting the view of a mosaic-like trajectory of phenotypic traits in our species. This work provides evidence for the recent emergence of the *GLI3* regulatory variants highlighted in chapter 3.

Bibliography

- [1] Chris Stringer. The origin and evolution of *Homo sapiens*. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1698):20150237, July 2016. Publisher: Royal Society.
- [2] Eleanor M. L. Scerri, Mark G. Thomas, Andrea Manica, Philipp Gunz, Jay T. Stock, Chris Stringer, Matt Grove, Huw S. Groucutt, Axel Timmermann, G. Philip Rightmire, Francesco d’Errico, Christian A. Tryon, Nick A. Drake, Alison S.

- Brooks, Robin W. Dennell, Richard Durbin, Brenna M. Henn, Julia Lee-Thorp, Peter deMenocal, Michael D. Petraglia, Jessica C. Thompson, Aylwyn Scally, and Lounès Chikhi. Did Our Species Evolve in Subdivided Populations across Africa, and Why Does It Matter? *Trends in Ecology & Evolution*, 33(8):582–594, August 2018. Publisher: Elsevier.
- [3] Eleanor M. L. Scerri and Manuel Will. The revolution that still isn't: The origins of behavioral complexity in *Homo sapiens*. *Journal of Human Evolution*, 179:103358, June 2023.
- [4] Rebecca Wragg Sykes. *Kindred: 300,000 Years of Neanderthal Life and Afterlife*. Bloomsbury Publishing, 1st edition, 2020.
- [5] Cedric Boeckx. Bilingualism: forays into human cognitive biology. *Journal of anthropological sciences = Rivista di antropologia: JASS*, 91:63–89, 2013.
- [6] Simon Neubauer, Jean-Jacques Hublin, and Philipp Gunz. The evolution of modern human brain shape. *Science Advances*, 4(1):eaao5961, January 2018.
- [7] M. Krings, A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, and S. Pääbo. Neandertal DNA sequences and the origin of modern humans. *Cell*, 90(1):19–30, July 1997.
- [8] David Reich, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, Adrian W. Briggs, Udo Stenzel, Philip L. F. Johnson, Tomislav Maricic, Jeffrey M. Good, Tomas Marques-Bonet, Can Alkan, Qiaomei Fu, Swapan Mallick, Heng Li, Matthias Meyer, Evan E. Eichler, Mark Stoneking, Michael Richards, Sahra Talamo, Michael V. Shunkov, Anatoli P. Derevianko, Jean-Jacques Hublin, Janet Kelso, Montgomery Slatkin, and Svante Pääbo. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468(7327):1053–1060, December 2010. Number: 7327 Publisher: Nature Publishing Group.
- [9] S. Pääbo, R. G. Higuchi, and A. C. Wilson. Ancient DNA and the polymerase chain reaction. The emerging field of molecular archaeology. *The Journal of Biological Chemistry*, 264(17):9709–9712, June 1989.
- [10] Svante Pääbo. The Human Condition—A Molecular Approach. *Cell*, 157(1):216–226, March 2014. Publisher: Elsevier.
- [11] Ludovic Orlando, Robin Allaby, Pontus Skoglund, Clio Der Sarkissian, Philipp W. Stockhammer, María C. Ávila Arcos, Qiaomei Fu, Johannes Krause, Eske Willerslev, Anne C. Stone, and Christina Warinner. Ancient DNA analysis. *Nature Reviews Methods Primers*, 1(1):1–26, February 2021. Number: 1 Publisher: Nature Publishing Group.
- [12] David Poeppel. The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, 29(1-2):34–55, 2012.

- [13] Cecilia S. L. Lai, Simon E. Fisher, Jane A. Hurst, Faraneh Vargha-Khadem, and Anthony P. Monaco. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*, 413(6855):519–523, October 2001. Number: 6855 Publisher: Nature Publishing Group.
- [14] Jonathan Chabout, Abhra Sarkar, Sheel R. Patel, Taylor Radden, David B. Dunson, Simon E. Fisher, and Erich D. Jarvis. A Foxp2 Mutation Implicated in Human Speech Deficits Alters Sequencing of Ultrasonic Vocalizations in Adult Male Mice. *Frontiers in Behavioral Neuroscience*, 10, 2016.
- [15] Malavika Murugan, Stephen Harward, Constance Scharff, and Richard Mooney. Diminished FoxP2 levels affect dopaminergic modulation of corticostriatal signaling important to song variability. *Neuron*, 80(6):1464–1476, December 2013.
- [16] K. E. Watkins, F. Vargha-Khadem, J. Ashburner, R. E. Passingham, A. Connelly, K. J. Friston, R. S. J. Frackowiak, M. Mishkin, and D. G. Gadian. MRI analysis of an inherited speech and language disorder: structural brain abnormalities. *Brain*, 125(3):465–478, March 2002.
- [17] Frédérique Liégeois, Torsten Baldeweg, Alan Connelly, David G. Gadian, Mortimer Mishkin, and Faraneh Vargha-Khadem. Language fMRI abnormalities associated with FOXP2 gene mutation. *Nature Neuroscience*, 6(11):1230–1237, November 2003.
- [18] Wolfgang Enard, Molly Przeworski, Simon E. Fisher, Cecilia S. L. Lai, Victor Wiebe, Takashi Kitano, Anthony P. Monaco, and Svante Pääbo. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, 418(6900):869–872, August 2002.
- [19] Jianzhi Zhang, David M Webb, and Ondrej Podlaha. Accelerated Protein Evolution and Origins of Human-Specific Features: FOXP2 as an Example. *Genetics*, 162(4):1825–1835, December 2002.
- [20] Johannes Krause, Carles Lalueza-Fox, Ludovic Orlando, Wolfgang Enard, Richard E. Green, Hernán A. Burbano, Jean-Jacques Hublin, Catherine Hänni, Javier Fortea, Marco de la Rasilla, Jaume Bertranpetit, Antonio Rosas, and Svante Pääbo. The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals. *Current Biology*, 17(21):1908–1912, November 2007.
- [21] Simon E. Fisher. Human Genetics: The Evolving Story of FOXP2. *Current Biology*, 29(2):R65–R67, January 2019. Publisher: Elsevier.
- [22] Tomislav Maricic, Viola Günther, Oleg Georgiev, Sabine Gehre, Marija Curlin, Christiane Schreiweis, Ronald Naumann, Hernán A. Burbano, Matthias Meyer, Carles Lalueza-Fox, Marco de la Rasilla, Antonio Rosas, Srecko Gajovic, Janet Kelso, Wolfgang Enard, Walter Schaffner, and Svante Pääbo. A recent evolutionary change affects a regulatory element in the human FOXP2 gene. *Molecular Biology and Evolution*, 30(4):844–852, April 2013.

- [23] Elizabeth Grace Atkinson, Amanda Jane Audesse, Julia Adela Palacios, Dean Michael Bobo, Ashley Elizabeth Webb, Sohini Ramachandran, and Brenna Mariah Henn. No Evidence for Recent Selection at FOXP2 among Diverse Human Populations. *Cell*, 174(6):1424–1435.e15, September 2018. Publisher: Elsevier.
- [24] Bart de Boer, Bill Thompson, Andrea Ravignani, and Cedric Boeckx. Evolutionary Dynamics Do Not Motivate a Single-Mutant Theory of Human Language. *Scientific Reports*, 10(1):451, January 2020.
- [25] James Briscoe and Oscar Marín. Looking at neurodevelopment through a big data lens. *Science (New York, N.Y.)*, 369(6510):eaaz8627, September 2020.
- [26] Richard E. Green, Johannes Krause, Susan E. Ptak, Adrian W. Briggs, Michael T. Ronan, Jan F. Simons, Lei Du, Michael Egholm, Jonathan M. Rothberg, Maja Paunovic, and Svante Pääbo. Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444(7117):330–336, November 2006. Number: 7117 Publisher: Nature Publishing Group.
- [27] Jeffrey D. Wall and Sung K. Kim. Inconsistencies in Neanderthal Genomic DNA Sequences. *PLOS Genetics*, 3(10):e175, October 2007. Publisher: Public Library of Science.
- [28] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green, Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korb, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapani, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Jun Wang, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yingrui Li, Shengmao Liu, Xiao Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Eric S. Lander, David M. Altshuler, Stacey B. Gabriel, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Paul Flicek, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, David R. Bentley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Hans Lehrach, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina,

Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, Stephen T. Sherry, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Gil A. McVean, Richard M. Durbin, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Golub, Teresa Webster, Brant Wong, Yiping Zhan, Adam Auton, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Richard A. Gibbs, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Donna Muzny, Aniko Sabo, Zhuoyi Huang, Jun Wang, Lachlan J. M. Coin, Lin Fang, Xiaosen Guo, Xin Jin, Guoqing Li, Qibin Li, Yingrui Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, David M. Altshuler, Eric Banks, Gaurav Bhatia, Guillermo del Angel, Stacey B. Gabriel, Giulio Genovese, Namrata Gupta, Heng Li, Seva Kashin, Eric S. Lander, Steven A. McCarroll, James C. Nemes, Ryan E. Poplin, Seungtae C. Yoon, Jayon Lihm, Vladimir Makarov, Andrew G. Clark, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Jan O. Korbel, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Paul Flicek, Kathryn Beal, Laura Clarke, Avik Datta, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Richard E. Smith, Daniel Zerbino, Xiangqun Zheng-Bradley, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, David R. Bentley, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Sean Humphray, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Sudbrak, Vyacheslav S. Amstislavskiy, Ralf Herwig, Elaine R. Mardis, Li Ding, Daniel C. Koboldt, David Larson, Kai Ye, Simon Gravel, The 1000 Genomes Project Consortium, Corresponding authors, Steering committee, Production group, Baylor College of Medicine, BGI-Shenzhen, Broad Institute of MIT and Harvard, Coriell Institute for Medical Research, European Bioinformatics Institute European Molecular Biology Laboratory, Illumina, Max Planck Institute for Molecular Genetics, McDonnell Genome Institute at Washington University, US National Institutes of Health, University of Oxford, Wellcome Trust Sanger Institute, Analysis group, Affymetrix, Albert

Einstein College of Medicine, Bilkent University, Boston College, Cold Spring Harbor Laboratory, Cornell University, European Molecular Biology Laboratory, Harvard University, Human Gene Mutation Database, Icahn School of Medicine at Mount Sinai, Louisiana State University, Massachusetts General Hospital, McGill University, and NIH National Eye Institute. A global reference for human genetic variation. *Nature*, 526(7571):68–74, October 2015. Number: 7571 Publisher: Nature Publishing Group.

- [29] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, Pontus Skoglund, Iosif Lazaridis, Sriram Sankararaman, Qiaomei Fu, Nadin Rohland, Gabriel Renaud, Yaniv Erlich, Thomas Willems, Carla Gallo, Jeffrey P. Spence, Yun S. Song, Giovanni Poletti, Francois Balloux, George van Driem, Peter de Knijff, Irene Gallego Romero, Aashish R. Jha, Doron M. Behar, Claudio M. Bravi, Cristian Capelli, Tor Hervig, Andres Moreno-Estrada, Olga L. Posukh, Elena Balanovska, Oleg Balanovsky, Sena Karachanak-Yankova, Hovhannes Sahakyan, Draga Toncheva, Levon Yepiskoposyan, Chris Tyler-Smith, Yali Xue, M. Syafiq Abdullah, Andres Ruiz-Linares, Cynthia M. Beall, Anna Di Rienzo, Choongwon Jeong, Elena B. Starikovskaya, Ene Metspalu, Jüri Parik, Richard Villems, Brenna M. Henn, Ugur Hodoglugil, Robert Mahley, Antti Sajantila, George Stamatoyannopoulos, Joseph T. S. Wee, Rita Khusainova, Elza Khusnutdinova, Sergey Litvinov, George Ayodo, David Comas, Michael F. Hammer, Toomas Kivisild, William Klitz, Cheryl A. Winkler, Damian Labuda, Michael Bamshad, Lynn B. Jorde, Sarah A. Tishkoff, W. Scott Watkins, Mait Metspalu, Stanislav Dryomov, Rem Sukernik, Lalji Singh, Kumarasamy Thangaraj, Svante Pääbo, Janet Kelso, Nick Patterson, and David Reich. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, October 2016. Number: 7624 Publisher: Nature Publishing Group.
- [30] Lukas F. K. Kuderna, Hong Gao, Mareike C. Janiak, Martin Kuhlwilm, Joseph D. Orkin, Thomas Bataillon, Shivakumara Manu, Alejandro Valenzuela, Juraj Bergman, Marjolaine Rousselle, Felipe Ennes Silva, Lidia Agueda, Julie Blanc, Marta Gut, Dorien de Vries, Ian Goodhead, R. Alan Harris, Muthuswamy Raveendran, Axel Jensen, Idrissa S. Chuma, Julie E. Horvath, Christina Hvilsom, David Juan, Peter Frandsen, Joshua G. Schraiber, Fabiano R. de Melo, Fabrício Bertuol, Hazel Byrne, Iracilda Sampaio, Izeni Farias, João Valsecchi, Malu Messias, Maria N. F. da Silva, Mihir Trivedi, Rogerio Rossi, Tomas Hrbek, Nicole Andriaholinirina, Clément J. Rabarivola, Alphonse Zaramody, Clifford J. Jolly, Jane Phillips-Conroy, Gregory Wilkerson, Christian Abee, Joe H. Simmons, Eduardo Fernandez-Duque, Sree Kanthaswamy, Fekadu Shiferaw, Dongdong Wu, Long Zhou, Yong Shao, Guojie Zhang, Julius D. Keyyu, Sascha Knauf, Minh D. Le, Esther Lizano, Stefan Merker, Arcadi Navarro, Tilo Nadler, Chiea Chuen Khor,

- Jessica Lee, Patrick Tan, Weng Khong Lim, Andrew C. Kitchener, Dietmar Zinner, Ivo Gut, Amanda D. Melin, Katerina Guschanski, Mikkel Heide Schierup, Robin M. D. Beck, Govindhaswamy Umapathy, Christian Roos, Jean P. Boubli, Jeffrey Rogers, Kyle Kai-How Farh, and Tomas Marques Bonet. A global catalog of whole-genome diversity from 233 primate species. *Science*, 380(6648):906–913, June 2023. Publisher: American Association for the Advancement of Science.
- [31] Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, Flora Jay, Kay Prüfer, Cesare de Filippo, Peter H. Sudmant, Can Alkan, Qiaomei Fu, Ron Do, Nadin Rohland, Arti Tandon, Michael Siebauer, Richard E. Green, Katarzyna Bryc, Adrian W. Briggs, Udo Stenzel, Jesse Dabney, Jay Shendure, Jacob Kitzman, Michael F. Hammer, Michael V. Shunkov, Anatoli P. Derevianko, Nick Patterson, Aida M. Andrés, Evan E. Eichler, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. A high-coverage genome sequence from an archaic Denisovan individual. *Science (New York, N.Y.)*, 338(6104):222–226, October 2012.
- [32] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare de Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L. F. Johnson, H el ene Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante P a bo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, January 2014. Number: 7481 Publisher: Nature Publishing Group.
- [33] Kay Prüfer, Cesare de Filippo, Steffi Grote, Fabrizio Mafessoni, Petra Korlevi c, Mateja Hajdinjak, Benjamin Vernot, Laurits Skov, Pingsun Hsieh, St ephane Peyr egne, David Reher, Charlotte Hopfe, Sarah Nagel, Tomislav Maricic, Qiaomei Fu, Christoph Theunert, Rebekah Rogers, Pontus Skoglund, Manjusha Chintalapati, Michael Dannemann, Bradley J. Nelson, Felix M. Key, Pavao Rudan,  Zeljko Ku can, Ivan Gu si c, Liubov V. Golovanova, Vladimir B. Doronichev, Nick Patterson, David Reich, Evan E. Eichler, Montgomery Slatkin, Mikkel H. Schierup, Aida M. Andr es, Janet Kelso, Matthias Meyer, and Svante P a bo. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science (New York, N.Y.)*, 358(6363):655–658, November 2017.
- [34] Fabrizio Mafessoni, Steffi Grote, Cesare de Filippo, Viviane Slon, Kseniya A.

- Kolobova, Bence Viola, Sergey V. Markin, Manjusha Chintalapati, Stephane Peyrégne, Laurits Skov, Pontus Skoglund, Andrey I. Krivoschapkin, Anatoly P. Derevianko, Matthias Meyer, Janet Kelso, Benjamin Peter, Kay Prüfer, and Svante Pääbo. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proceedings of the National Academy of Sciences*, 117(26):15132–15136, June 2020. Publisher: Proceedings of the National Academy of Sciences.
- [35] Richard E. Green, Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, Heng Li, Weiwei Zhai, Markus Hsi-Yang Fritz, Nancy F. Hansen, Eric Y. Durand, Anna-Sapfo Malaspinas, Jeffrey D. Jensen, Tomas Marques-Bonet, Can Alkan, Kay Prüfer, Matthias Meyer, Hernán A. Burbano, Jeffrey M. Good, Rigo Schultz, Ayinuer Aximu-Petri, Anne Butthof, Barbara Höber, Barbara Höffner, Madlen Siegemund, Antje Weihmann, Chad Nusbaum, Eric S. Lander, Carsten Russ, Nathaniel Novod, Jason Affourtit, Michael Egholm, Christine Verna, Pavao Rudan, Dejana Brajkovic, Željko Kucan, Ivan Gušić, Vladimir B. Doronichev, Liubov V. Golovanova, Carles Lalueza-Fox, Marco de la Rasilla, Javier Fortea, Antonio Rosas, Ralf W. Schmitz, Philip L. F. Johnson, Evan E. Eichler, Daniel Falush, Ewan Birney, James C. Mullikin, Montgomery Slatkin, Rasmus Nielsen, Janet Kelso, Michael Lachmann, David Reich, and Svante Pääbo. A draft sequence of the Neandertal genome. *Science (New York, N.Y.)*, 328(5979):710–722, May 2010.
- [36] Mateja Hajdinjak, Qiaomei Fu, Alexander Hübner, Martin Petr, Fabrizio Mafessoni, Steffi Grote, Pontus Skoglund, Vagheesh Narasimham, Hélène Rougier, Isabelle Crevecoeur, Patrick Semal, Marie Soressi, Sahra Talamo, Jean-Jacques Hublin, Ivan Gušić, Željko Kućan, Pavao Rudan, Liubov V. Golovanova, Vladimir B. Doronichev, Cosimo Posth, Johannes Krause, Petra Korlević, Sarah Nagel, Birgit Nickel, Montgomery Slatkin, Nick Patterson, David Reich, Kay Prüfer, Matthias Meyer, Svante Pääbo, and Janet Kelso. Reconstructing the genetic history of late Neanderthals. *Nature*, 555(7698):652–656, March 2018.
- [37] Martin Kuhlwilm and Cedric Boeckx. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Scientific Reports*, 9(1):8463, June 2019. Number: 1 Publisher: Nature Publishing Group.
- [38] M. C. King and A. C. Wilson. Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, 188(4184):107–116, April 1975.
- [39] Anders Bergström, Chris Stringer, Mateja Hajdinjak, Eleanor M. L. Scerri, and Pontus Skoglund. Origins of modern human ancestry. *Nature*, 590(7845):229–237, February 2021. Number: 7845 Publisher: Nature Publishing Group.
- [40] Patrick F. Reilly, Audrey Tjahjadi, Samantha L. Miller, Joshua M. Akey, and Serena Tucci. The contribution of Neanderthal introgression to modern human traits. *Current Biology*, 32(18):R970–R983, September 2022. Publisher: Elsevier.

- [41] Rajiv C. McCoy, Jon Wakefield, and Joshua M. Akey. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell*, 168(5):916–927.e12, February 2017. Publisher: Elsevier.
- [42] Martin Petr, Svante Pääbo, Janet Kelso, and Benjamin Vernot. Limits of long-term selection against Neandertal introgression. *Proceedings of the National Academy of Sciences*, 116(5):1639–1644, January 2019. Publisher: Proceedings of the National Academy of Sciences.
- [43] Mateja Hajdinjak, Fabrizio Mafessoni, Laurits Skov, Benjamin Vernot, Alexander Hübner, Qiaomei Fu, Elena Essel, Sarah Nagel, Birgit Nickel, Julia Richter, Oana Teodora Moldovan, Silviu Constantin, Elena Endarova, Nikolay Zahariev, Rosen Spasov, Frido Welker, Geoff M. Smith, Virginie Sinet-Mathiot, Lindsey Paskulin, Helen Fewlass, Sahra Talamo, Zeljko Rezek, Svoboda Sirakova, Nikolay Sirakov, Shannon P. McPherron, Tsenka Tsanova, Jean-Jacques Hublin, Benjamin M. Peter, Matthias Meyer, Pontus Skoglund, Janet Kelso, and Svante Pääbo. Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry. *Nature*, 592(7853):253–257, April 2021. Number: 7853 Publisher: Nature Publishing Group.
- [44] Carl Veller, Nathaniel B Edelman, Pavitra Muralidhar, and Martin A Nowak. Recombination and selection against introgressed DNA. *Evolution*, 77(4):1131–1144, April 2023.
- [45] Lu Chen, Aaron B. Wolf, Wenqing Fu, Liming Li, and Joshua M. Akey. Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell*, 180(4):677–687.e16, February 2020. Publisher: Elsevier.
- [46] Benjamin Vernot, Serena Tucci, Janet Kelso, Joshua G. Schraiber, Aaron B. Wolf, Rachel M. Gittelman, Michael Dannemann, Steffi Grote, Rajiv C. McCoy, Heather Norton, Laura B. Scheinfeldt, David A. Merriwether, George Koki, Jonathan S. Friedlaender, Jon Wakefield, Svante Pääbo, and Joshua M. Akey. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science (New York, N.Y.)*, 352(6282):235–239, April 2016.
- [47] Martin kuhlwilm. The evolution of FOXP2 in the light of admixture. *Current Opinion in Behavioral Sciences*, 21:120–126, June 2018. Publisher: Elsevier.
- [48] Fernando Racimo. Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. *Genetics*, 202(2):733–750, February 2016.
- [49] Stéphane Peyrégne, Michael James Boyle, Michael Dannemann, and Kay Prüfer. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Research*, 27(9):1563–1572, September 2017.
- [50] Gökberk Alagöz, Barbara Molz, Else Eising, Dick Schijven, Clyde Francks, Jason L. Stein, and Simon E. Fisher. Using neuroimaging genomics to investigate the evolution of human brain structure. *Proceedings of the National Academy*

- of Sciences*, 119(40):e2200638119, October 2022. Publisher: Proceedings of the National Academy of Sciences.
- [51] Philipp Gunz, Amanda K. Tilot, Katharina Wittfeld, Alexander Teumer, Chin Yang Shapland, Theo G. M. van Erp, Michael Dannemann, Benjamin Vernot, Simon Neubauer, Tulio Guadalupe, Guillén Fernández, Han G. Brunner, Wolfgang Enard, James Fallon, Norbert Hosten, Uwe Völker, Antonio Profico, Fabio Di Vincenzo, Giorgio Manzi, Janet Kelso, Beate St. Pourcain, Jean-Jacques Hublin, Barbara Franke, Svante Pääbo, Fabio Macciardi, Hans J. Grabe, and Simon E. Fisher. Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity. *Current Biology*, 29(1):120–127.e5, January 2019.
- [52] Simon Neubauer, Philipp Gunz, and Jean-Jacques Hublin. Endocranial shape changes during growth in chimpanzees and humans: a morphometric analysis of unique and shared aspects. *Journal of Human Evolution*, 59(5):555–566, November 2010.
- [53] Philipp Gunz, Simon Neubauer, Bruno Maureille, and Jean-Jacques Hublin. Brain development after birth differs between Neanderthals and modern humans. *Current Biology*, 20(21):R921–R922, November 2010.
- [54] Philipp Gunz, Simon Neubauer, Lubov Golovanova, Vladimir Doronichev, Bruno Maureille, and Jean-Jacques Hublin. A uniquely modern human pattern of endocranial development. Insights from a new cranial reconstruction of the Neandertal newborn from Mezmaiskaya. *Journal of Human Evolution*, 62(2):300–313, February 2012.
- [55] Marcia S. Ponce de León, Thibaut Bienvenu, Takeru Akazawa, and Christoph P. E. Zollikofer. Brain development is similar in Neanderthals and modern humans. *Current biology: CB*, 26(14):R665–666, July 2016.
- [56] Angela D. Friederici. Evolutionary neuroanatomical expansion of Broca’s region serving a human-specific function. *Trends in Neurosciences*, August 2023.
- [57] Cedric Boeckx. The language-ready head: Evolutionary considerations. *Psychonomic Bulletin & Review*, 24(1):194–199, February 2017.
- [58] Takanori Kochiyama, Naomichi Ogihara, Hiroki C. Tanabe, Osamu Kondo, Hideki Amano, Kunihiro Hasegawa, Hiromasa Suzuki, Marcia S. Ponce de León, Christoph P. E. Zollikofer, Markus Bastir, Chris Stringer, Norihiro Sadato, and Takeru Akazawa. Reconstructing the Neandertal brain using computational anatomy. *Scientific Reports*, 8(1):6296, April 2018. Number: 1 Publisher: Nature Publishing Group.
- [59] Silvia Velasco, Bruna Paulsen, and Paola Arlotta. 3D Brain Organoids: Studying Brain Development and Disease Outside the Embryo. *Annual Review of Neuroscience*, 43(1):375–389, 2020. eprint: <https://doi.org/10.1146/annurev-neuro-070918-050154>.

- [60] Sergiu P. Pasca, Paola Arlotta, Helen S. Bateup, J. Gray Camp, Silvia Cappello, Fred H. Gage, Jürgen A. Knoblich, Arnold R. Kriegstein, Madeline A. Lancaster, Guo-Li Ming, Alysson R. Muotri, In-Hyun Park, Orly Reiner, Hongjun Song, Lorenz Studer, Sally Temple, Giuseppe Testa, Barbara Treutlein, and Flora M. Vaccarino. A nomenclature consensus for nervous system organoids and assembloids. *Nature*, 609(7929):907–910, September 2022. Number: 7929 Publisher: Nature Publishing Group.
- [61] Carly V. Weiss, Lana Harshman, Fumitaka Inoue, Hunter B. Fraser, Dmitri A. Petrov, Nadav Ahituv, and David Gokhman. The cis-regulatory effects of modern human-specific variants. *eLife*, 10:e63713, April 2021.
- [62] Alex A. Pollen, Umut Kilik, Craig B. Lowe, and J. Gray Camp. Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. *Nature Reviews Genetics*, pages 1–25, February 2023. Publisher: Nature Publishing Group.
- [63] Sean Whalen, Fumitaka Inoue, Hane Ryu, Tyler Fair, Eirene Markenscoff-Papadimitriou, Kathleen Keough, Martin Kircher, Beth Martin, Beatriz Alvarado, Orry Elor, Dianne Laboy Cintron, Alex Williams, Md. Abul Hassan Samee, Sean Thomas, Robert Krencik, Erik M. Ullian, Arnold Kriegstein, John L. Rubenstein, Jay Shendure, Alex A. Pollen, Nadav Ahituv, and Katherine S. Pollard. Machine learning dissection of human accelerated regions in primate neurodevelopment. *Neuron*, 111(6):857–873.e8, March 2023.
- [64] Evonne McArthur, David C. Rinker, Erin N. Gilbertson, Geoff Fudenberg, Maureen Pittman, Kathleen Keough, Katherine S. Pollard, and John A. Capra. Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence, February 2022. Pages: 2022.02.07.479462 Section: New Results.
- [65] Colin M. Brand, Laura L. Colbran, and John A. Capra. Resurrecting the alternative splicing landscape of archaic hominins using machine learning. *Nature Ecology & Evolution*, 7(6):939–953, June 2023. Number: 6 Publisher: Nature Publishing Group.
- [66] Felipe Mora-Bermúdez, Farhath Badsha, Sabina Kanton, J. Gray Camp, Benjamin Vernot, Kathrin Köhler, Birger Voigt, Keisuke Okita, Tomislav Maricic, Zhisong He, Robert Lachmann, Svante Pääbo, Barbara Treutlein, and Wieland B. Huttner. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *eLife*, 5:e18683, September 2016.
- [67] Tomoki Otani, Maria C. Marchetto, Fred H. Gage, Benjamin D. Simons, and Frederick J. Livesey. 2D and 3D Stem Cell Models of Primate Cortical Development Identify Species-Specific Differences in Progenitor Behavior Contributing to Brain Size. *Cell Stem Cell*, 18(4):467–480, April 2016.

- [68] Ryohei Iwata, Pierre Casimir, Emir Erkol, Leïla Boubakar, Mélanie Planque, Isabel M. Gallego López, Martyna Ditkowska, Vaiva Gaspariunaite, Sofie Beckers, Daan Remans, Katlijn Vints, Anke Vandekeere, Suresh Poovathingal, Matthew Bird, Ine Vlaeminck, Eline Creemers, Keimpe Wierda, Nikky Corthout, Pieter Vermeersch, Sébastien Carpentier, Kristofer Davie, Massimiliano Mazzone, Natalia V. Gounko, Stein Aerts, Bart Ghesquière, Sarah-Maria Fendt, and Pierre Vanderhaeghen. Mitochondria metabolism sets the species-specific tempo of neuronal development. *Science*, 379(6632):eabn4705, January 2023. Publisher: American Association for the Advancement of Science.
- [69] Silvia Benito-Kwiecinski, Stefano L. Giandomenico, Magdalena Sutcliffe, Erlend S. Riis, Paula Freire-Pritchett, Iva Kelava, Stephanie Wunderlich, Ulrich Martin, Gregory A. Wray, Kate McDole, and Madeline A. Lancaster. An early cell shape transition drives evolutionary expansion of the human forebrain. *Cell*, 184(8):2084–2102.e19, April 2021.
- [70] Alex A. Pollen, Aparna Bhaduri, Madeline G. Andrews, Tomasz J. Nowakowski, Olivia S. Meyerson, Mohammed A. Mostajo-Radji, Elizabeth Di Lullo, Beatriz Alvarado, Melanie Bedolli, Max L. Dougherty, Ian T. Fiddes, Zev N. Kronenberg, Joe Shuga, Anne A. Leyrat, Jay A. West, Marina Bershteyn, Craig B. Lowe, Bryan J. Pavlovic, Sofie R. Salama, David Haussler, Evan E. Eichler, and Arnold R. Kriegstein. Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. *Cell*, 176(4):743–756.e17, February 2019.
- [71] Sabina Kanton, Michael James Boyle, Zhisong He, Malgorzata Santel, Anne Weigert, Fátima Sanchís-Calleja, Patricia Guijarro, Leila Sidow, Jonas Simon Fleck, Dingding Han, Zhengzong Qian, Michael Heide, Wieland B. Huttner, Philipp Khaitovich, Svante Pääbo, Barbara Treutlein, and J. Gray Camp. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*, 574(7778):418–422, October 2019. Number: 7778 Publisher: Nature Publishing Group.
- [72] Anneline Pinson, Lei Xing, Takashi Namba, Nereo Kalebic, Jula Peters, Christina Eugster Oegema, Sofia Traikov, Katrin Reppe, Stephan Riesenberger, Tomislav Maricic, Razvan Derihaci, Pauline Wimberger, Svante Pääbo, and Wieland B. Huttner. Human TKTL1 implies greater neurogenesis in frontal neocortex of modern humans than Neanderthals. *Science*, 377(6611):eabl6422, September 2022. Publisher: American Association for the Advancement of Science.
- [73] Vita Stepanova, Kaja Ewa Moczulska, Guido N Vacano, Ilia Kurochkin, Xi-angchun Ju, Stephan Riesenberger, Dominik Macak, Tomislav Maricic, Linda Dombrowski, Maria Schörnig, Konstantinos Anastassiadis, Oliver Baker, Ronald Naumann, Ekaterina Khrameeva, Anna Vanushkina, Elena Stekolshchikova, Alina

- Egorova, Anna Tkachev, Randall Mazzarino, Nathan Duval, Dmitri Zubkov, Patrick Giavalisco, Terry G Wilkinson, David Patterson, Philipp Khaitovich, and Svante Pääbo. Reduced purine biosynthesis in humans after their divergence from Neandertals. *eLife*, 10:e58741, 2021.
- [74] Stéphane Peyrégne, Janet Kelso, Benjamin M Peter, and Svante Pääbo. The evolutionary history of human spindle genes includes back-and-forth gene flow with Neandertals. *eLife*, 11:e75464, July 2022. Publisher: eLife Sciences Publications, Ltd.
- [75] Felipe Mora-Bermúdez, Philipp Kanis, Dominik Macak, Jula Peters, Ronald Naumann, Lei Xing, Mihail Sarov, Sylke Winkler, Christina Eugster Oegema, Christiane Haffner, Pauline Wimberger, Stephan Riesenberger, Tomislav Maricic, Wieland B. Huttner, and Svante Pääbo. Longer metaphase and fewer chromosome segregation errors in modern human than Neanderthal brain development. *Science Advances*, 8(30):eabn7702, July 2022. Publisher: American Association for the Advancement of Science.
- [76] Adam S Wilkins, Richard W Wrangham, and W Tecumseh Fitch. The “Domestication Syndrome” in Mammals: A Unified Explanation Based on Neural Crest Cell Behavior and Genetics. *Genetics*, 197(3):795–808, July 2014.
- [77] Kazuo Okanoya. Sexual communication and domestication may give rise to the signal complexity necessary for the emergence of language: An indication from songbird studies. *Psychonomic Bulletin & Review*, 24(1):106–110, February 2017.
- [78] Richard W. Wrangham. Targeted conspiratorial killing, human self-domestication and the evolution of groupishness. *Evolutionary Human Sciences*, 3:e26, January 2021. Publisher: Cambridge University Press.
- [79] James Thomas and Simon Kirby. Self domestication and the evolution of language. *Biology & Philosophy*, 33(1):9, March 2018.
- [80] Thomas O’Rourke, Pedro Tiago Martins, Rie Asano, Ryosuke O. Tachibana, Kazuo Okanoya, and Cedric Boeckx. Capturing the Effects of Domestication on Vocal Learning Complexity. *Trends in Cognitive Sciences*, 25(6):462–474, June 2021.
- [81] Cedric Boeckx. What made us “hunter-gatherers of words”. *Frontiers in Neuroscience*, 17, 2023.
- [82] Matteo Zanella, Alessandro Vitriolo, Alejandro Andirko, Pedro Tiago Martins, Stefanie Sturm, Thomas O’Rourke, Magdalena Laugsch, Natascia Malerba, Adrianos Skaros, Sebastiano Trattaro, Pierre-Luc Germain, Marija Mihailovic, Giuseppe Merla, Alvaro Rada-Iglesias, Cedric Boeckx, and Giuseppe Testa. Dosage analysis of the 7q11.23 Williams region identifies BAZ1B as a major human gene patterning the modern human face and underlying self-domestication. *Science Advances*, 5(12):eaaw7908, December 2019.

- [83] Pasko Rakic. A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. *Trends in Neurosciences*, 18(9):383–388, September 1995.
- [84] Arnold Kriegstein, Stephen Noctor, and Verónica Martínez-Cerdeño. Patterns of neural stem and progenitor cell division may underlie evolutionary cortical expansion. *Nature Reviews Neuroscience*, 7(11):883–890, November 2006. Number: 11 Publisher: Nature Publishing Group.
- [85] Cécile Hautecoeur and François Glineur. Nonnegative Matrix Factorization over Continuous Signals using Parametrizable Functions. *Neurocomputing*, 416:256–265, November 2020.
- [86] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E. Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Tamar Jessurun Lobo, Emma M. Keizer, Indu Khatri, Szymon M. Kielbasa, Jan O. Korbel, Alexey M. Kozlov, Tzu-Hao Kuo, Boudewijn P.F. Lelieveldt, Ion I. Mandoiu, John C. Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J. Theis, Huan Yang, Alex Zelikovsky, Alice C. McHardy, Benjamin J. Raphael, Sohrab P. Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31, February 2020.
- [87] Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, February 2023. Number: 7949 Publisher: Nature Publishing Group.
- [88] Takashi Namba, Jeannette Nardelli, Pierre Gressens, and Wieland B. Huttner. Metabolic Regulation of Neocortical Expansion in Development and Evolution. *Neuron*, 109(3):408–419, February 2021.
- [89] THE PSYCHENCODE CONSORTIUM. Revealing the brain’s molecular architecture. *Science*, 362(6420):1262–1263, December 2018. Publisher: American Association for the Advancement of Science.
- [90] Mingfeng Li, Gabriel Santpere, Yuka Imamura Kawasawa, Oleg V. Evgrafov, Forrest O. Gulden, Sirisha Pochareddy, Susan M. Sunkin, Zhen Li, Yuray Shin, Ying Zhu, André M. M. Sousa, Donna M. Werling, Robert R. Kitchen, Hyo Jung Kang, Mihovil Pletikos, Jinmyung Choi, Sydney Muchnik, Xuming Xu, Daifeng Wang, Belen Lorente-Galdos, Shuang Liu, Paola Giusti-Rodríguez, Hyejung

Won, Christiaan A. de Leeuw, Antonio F. Pardiñas, BrainSpan Consortium, PsychENCODE Consortium, PsychENCODE Developmental Subgroup, Ming Hu, Fulai Jin, Yun Li, Michael J. Owen, Michael C. O'Donovan, James T. R. Walters, Danielle Posthuma, Mark A. Reimers, Pat Levitt, Daniel R. Weinberger, Thomas M. Hyde, Joel E. Kleinman, Daniel H. Geschwind, Michael J. Hawrylycz, Matthew W. State, Stephan J. Sanders, Patrick F. Sullivan, Mark B. Gerstein, Ed S. Lein, James A. Knowles, and Nenad Sestan. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, 362(6420):eaat7615, December 2018. Publisher: American Association for the Advancement of Science.

Chapter 2

Modern human changes in regulatory regions implicated in cortical development

Published as:

Moriano, J. & Boeckx, C. 2020. Modern human changes in regulatory regions implicated in cortical development. *BMC Genomics*


doi:[10.1186/s12864-020-6706-x](https://doi.org/10.1186/s12864-020-6706-x)

RESEARCH ARTICLE

Open Access

Modern human changes in regulatory regions implicated in cortical development



Juan Moriano^{1,2*}  and Cedric Boeckx^{1,2,3*}

Abstract

Background: Recent paleogenomic studies have highlighted a very small set of proteins carrying modern human-specific missense changes in comparison to our closest extinct relatives. Despite being frequently alluded to as highly relevant, species-specific differences in regulatory regions remain understudied. Here, we integrate data from paleogenomics, chromatin modification and physical interaction, and single-cell gene expression of neural progenitor cells to identify derived regulatory changes in the modern human lineage in comparison to Neanderthals/Denisovans. We report a set of genes whose enhancers and/or promoters harbor modern human single nucleotide changes and are active at early stages of cortical development.

Results: We identified 212 genes controlled by regulatory regions harboring modern human changes where Neanderthals/Denisovans carry the ancestral allele. These regulatory regions significantly overlap with putative modern human positively-selected regions and schizophrenia-related genetic loci. Among the 212 genes, we identified a substantial proportion of genes related to transcriptional regulation and, specifically, an enrichment for the SETD1A histone methyltransferase complex, known to regulate WNT signaling for the generation and proliferation of intermediate progenitor cells.

Conclusions: This study complements previous research focused on protein-coding changes distinguishing our species from Neanderthals/Denisovans and highlights chromatin regulation as a functional category so far overlooked in modern human evolution studies. We present a set of candidates that will help to illuminate the investigation of modern human-specific ontogenetic trajectories.

Keywords: Modern humans, Neanderthals/Denisovans, Paleogenomics, Regulatory regions, Chromatin regulation, SETD1A/histone methyltransferase complex

Background

Progress in the field of paleogenomics has allowed researchers to study the genetic basis of modern human-specific traits in comparison to our closest extinct relatives, the Neanderthals and Denisovans [1]. One such trait concerns the period of growth and maturation of the brain, which is a major factor underlying the characteristic ‘globular’ head shape of modern humans [2]. Comparative genomic analyses using high-quality Neanderthal/Denisovan genomes [3–5] have revealed

missense changes in the modern human lineage affecting proteins involved in the division of neural progenitor cells, key for the proper generation of neurons in an orderly spatiotemporal manner [4, 6]. But the total number of fixed missense changes in the modern human lineage amounts to less than one hundred proteins [1, 6]. This suggests that changes falling outside protein-coding regions may be equally relevant to understand the genetic basis of modern human-specific traits, as proposed more than four decades ago [7]. In this context it is noteworthy that human positively-selected genomic regions were found to be enriched in regulatory regions [8], and that signals of negative selection against Neanderthal

* Correspondence: jmoriano@ub.edu; cedric.boeckx@ub.edu

¹Universitat de Barcelona, Gran Via de les Corts Catalanes, Barcelona, Spain
Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

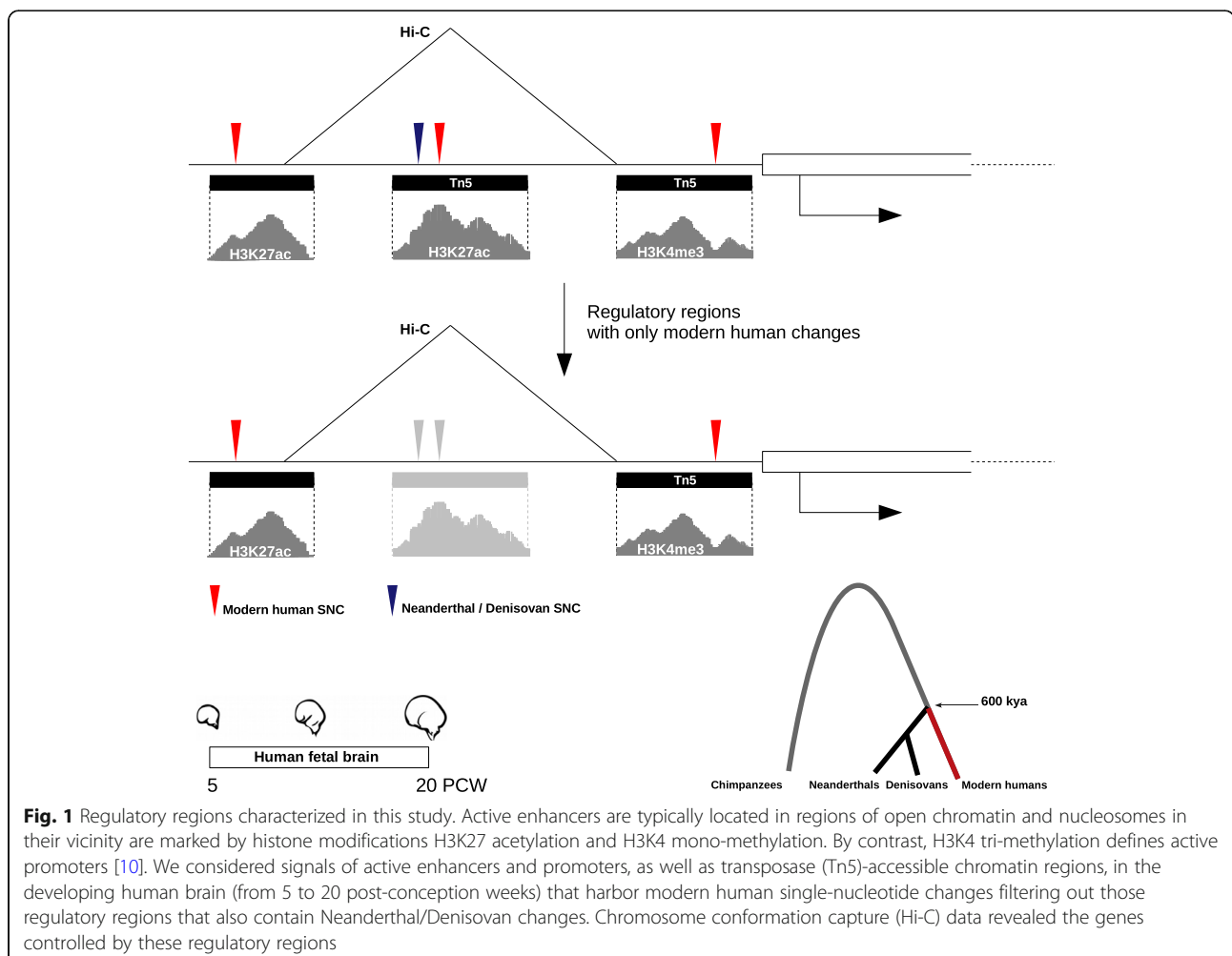
DNA introgression were reported in promoters and conserved genomic regions [9].

Here, we report a set of genes under the control of regulatory regions that harbor modern human-lineage genetic changes and are active at early stages of cortical development (Fig. 1). We integrated data on chromatin immunoprecipitation and open chromatin regions identifying enhancers and promoters active during human cortical development, and the genes regulated by them as revealed by chromatin physical interaction data, together with paleogenomic data of single-nucleotide changes (SNC) distinguishing modern humans and Neanderthal/Denisovan lineages. This allowed us to uncover those enhancer and promoters that harbor modern human SNC (thereafter, mSNC) at fixed or nearly fixed frequency (as defined by [6]) in present-day human populations and where the Neanderthals/Denisovans carry the ancestral allele (Methods section). Next, we analysed single-cell gene expression data and performed co-expression network analysis to identify the genes plausibly under human-specific regulation within genetic

networks in neural progenitor cells (Methods section). Many of the genes controlled by regulatory regions satisfying the aforementioned criteria are involved in chromatin regulation, and prominently among these, the SETD1A histone methyltransferase (SETD1A/HMT) complex. This complex, which has not figured prominently in the modern human evolution literature until now, appears to have been targeted in modern human evolution and specifically regulates the indirect mode of neurogenesis through the control of WNT/ β -CATENIN signaling.

Results

Two hundred and twelve genes were found associated to regulatory regions active in the developing human cortex (from 5 to 20 post-conception weeks) that harbor mSNCs and do not contain Neanderthal/Denisovan changes (Suppl. Mat. Tables S1 & S2). Among these, some well-studied disease-relevant genes are found: *HTT* (Huntington disease) [11], *FOXP2* (language impairment) [12], *CHD8* and *CPEB4* (autism spectrum



disorder) [13, 14], *TCF4* (Pitt-Hopkins syndrome and schizophrenia) [15, 16], *GLI3* (macrocephaly and Greig cephalopolysyndactyly syndrome) [17], *PHC1* (primary, autosomal recessive, microcephaly-11) [18], *RCAN1* (Down syndrome) [19], and *DYNC1H1* (cortical malformations and microcephaly) [20].

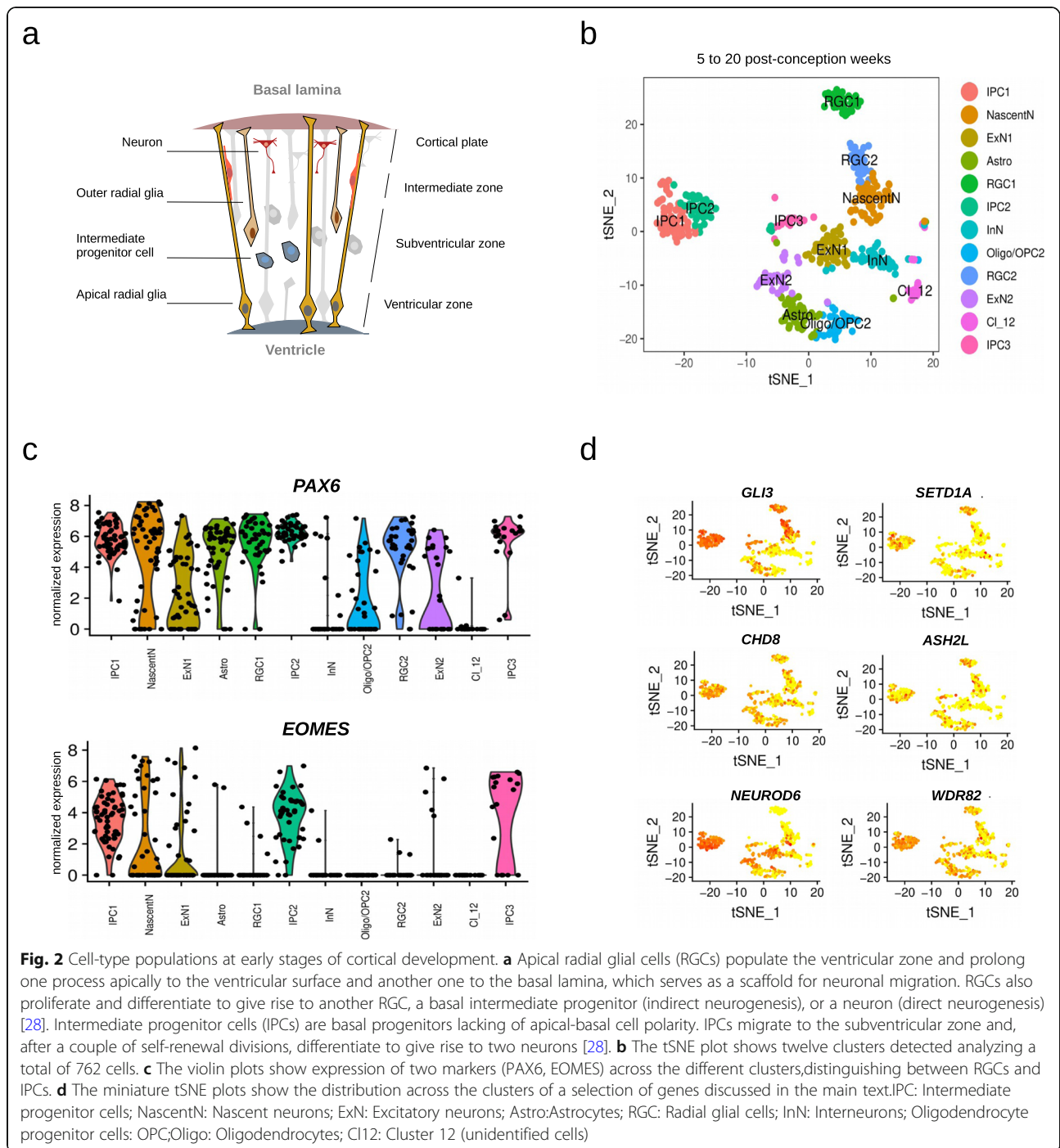
Twelve out of the 212 genes contain fixed mSNCs in enhancers (*NEUROD6*, *GRIN2B*, *LRRC23*, *RNF44*, *KCNA3*, *TCF25*, *TMLHE*, *GLI4*, *DDX12P*, *PLP2*, *TFE3*, *SPG7*), with *LRRC23* having three such changes, and *GRIN2B*, *DDX12P* and *TFE3*, two each. Fourteen genes have fixed mSNCs in their promoters (*LRRC23*, *SETD1A*, *FOXJ2*, *LIMCH1*, *ZFAT*, *SPOP*, *DLGAP4*, *HS6ST2*, *UBE2A*, *FKBP1A*, *RPL6*, *LINC01159*, *RBM4B*, *NFIB*). Only one gene, *LRRC23*, exhibits fixed changes in both its enhancer and promoter regions. To identify putatively mSNC-enriched regions, we ranked regulatory regions by mutation density (Methods section). Top candidate enhancers (top 5% in hits-per-region length distribution) were associated with potassium channel *KCNQ5*, actin-binding protein *FSCN1*, and neuronal marker *NEUROD6*. Top candidate promoters were linked to cytoplasmic dynein *DYNC1H1*, nuclear factor *NFIB*, PHD and RING finger domains-containing *PHRF1*, and kinesin light *KLC1* (Suppl. Mat. Table S3 & S4). Interestingly, most of these are known to be involved in later stages of neurogenesis (differentiation and migration steps).

Previous work has shown an enrichment of enhancer and promoter regions within modern human putative positively-selected regions [8]. For those regulatory regions containing mSNC, a significant over-representation was found for enhancers (permutation test; p -value 0.01) and promoters (permutation test; p -value 10^{-4}) overlapping with modern human candidate sweep regions [8] (Suppl. Fig. 1; Suppl. Mat. Table S5). In addition, we found a significant enrichment for enhancers (permutation test; p -value 0.04; while for promoter regions p -value 0.08) overlapping with genetic loci associated to schizophrenia [21]. By contrast, no significant overlap was found for enhancers/promoters and autism spectrum disorder risk variants ([22], retrieved from [23]) (Suppl. Fig. 1). Single-nucleotide variants can have an impact on epigenetic signals and transcription factor binding affinity in regulatory regions, and thus can alter gene expression levels [24–26]. We also performed motif enrichment analysis for our enhancer/promoter region datasets (Methods section). We found a motif enrichment in enhancer regions for transcriptional regulators *IRF8*, *PU.1*, *CTCF* (Benjamini q -value 0.01) and *OCT4* (Benjamini q -value 0.02); while for promoter regions a motif enrichment was detected for the zinc finger-containing (and *WNT* signaling regulator) *ZBTB33* (Benjamini q -value 0.03).

Next, we evaluated relevant gene ontology and biological categories in our 212 gene list (Methods section). We identified a substantial proportion of genes related to β -catenin binding (GO:0008013; hypergeometric test (h.t.): adj p -value 0.11) and transcriptional regulation (GO:0044212; h.t.: adj p -value 0.17), and detected a significant enrichment from the CORUM protein complexes database for the SETD1A/HMT complex (CORUM:2731; h.t.: adj p -value 0.01). Indeed, three members of the SETD1A/HMT complex are present in our 212 gene list: SETD1A (fixed mSNC in promoter), ASH2L (mSNC in enhancer) and WDR82 (mSNC in enhancer). SETD1A associates to the core of an H3K4 methyltransferase complex (ASH2L, WDR5, RBBP5, DPY30) and to WDR82, which recruits RNA polymerase II, to promote transcription of target genes through histone modification H3K4me3 [27]. Furthermore, the *SETD1A* promoter and the *WDR82* enhancer containing the relevant changes fall within putative positively-selected regions in the modern human lineage [8] (Suppl. Mat. Table S5).

The abundance of transcriptional regulators and the specific enrichment for the SETD1A/HMT complex led us to examine the gene expression programs likely under their influence in neural progenitor cells. From 5 to 20 post-conception weeks, different types of cells populate the germinal zones of the developing cortex (Fig. 2). We re-analyzed gene expression data at single-cell resolution from a total of 762 cells from the developing human cortex, controlling for cell-cycle heterogeneity as a confounding factor in the analysis of progenitor populations (Methods section). We focused on two progenitor cell-types—radial glial and intermediate progenitor cells (RGCs and IPCs, respectively)—two of the main types of progenitor cells that give rise, in an orderly manner, to the neurons present in the adult brain (Fig. 2). Two clusters of RGCs were identified (*PAX6+* and *EOMES-* cells), and three clusters of IPCs were detected (*EOMES*-expressing cells, with cells retaining *PAX6* expression and some expressing differentiation marker *TUJ1*), largely replicating what has been reported in the original publication for this dataset (Suppl. Fig. 2). We next identified genetic networks (based on highly-correlated gene expression levels) in the different cluster of progenitor cells (except for IPC cluster 3, which was excluded due to the low number of cells) (Methods section; Suppl. Figs. 3 & 4). This allowed us to identify genes present in the 212 gene list within modules of co-expressed genes whose biological relevance was assessed through a functional enrichment analysis using *g:Profiler2* R package [29] (hypergeometric test; see Methods).

An over-representation of genes related to the human phenotype ontology term ‘Neurodevelopmental abnormality’ was detected in the RGC-cluster 2 turquoise



module (HP:0012759; h.t.: adj *p*-value 0.03, [Suppl. Mat. Table S6](#)). Indeed, a considerable amount of genes were found to be associated to phenotype terms ‘Neurodevelopmental delay’ and ‘Skull size’ (HP:0012758 and HP:0000240, respectively; h.t.: adj *p*-value 0.07 and 0.13, respectively; [Suppl. Mat. Table S7](#)). These terms have appeared prominently in the human evolution literature in the context of neoteny and delay brain maturation, brain growth and the craniofacial phenotype in between

species comparisons [6, 30–32]. Two chromatin regulators with mSNC in regulatory regions are present in these two ontology terms and are associated to neurodevelopmental disorders: KDM6A (mSNC in promoter), which associates to the H3K4 methyltransferase complex [27], and is mutated in patients with Kabuki syndrome [33]; and PHC1 (mSNC in promoter), a component of the repressive complex PRC1 [27], found in patients with primary microcephaly-11 [18]. Among the total

genes related to the ‘Skull size’ term ($n = 109$), we found an over-representation of genes (*CDON*, *GLI3*, *KIF7*, *GAS1*) related to the hedgehog signaling pathway (KEGG:04340; h.t.: adj p -value 0.05). Of these, *GLI3* (mSNC in promoter) is perhaps the most salient member, highlighted in previous work as a gene harbouring an excess of mSNC [6]. *GLI3* is a gene linked to macrocephaly and the craniofacial phenotype [17, 34] and under putative modern human positive selection [8]. Considering that hedgehog signaling plays a critical role in basal progenitor expansion [35], we note the presence in this turquoise module of the outer radial glia-specific genes *IL6ST* and *STAT3* [36]. The forkhead-box transcription factor *FOXP2* is also present in RGC-cluster 2 turquoise module and associated to the ‘Neurodevelopmental delay’ ontology term. Its promoter harbors an almost fixed (>99%) mSNC. *FOXP2* is a highly conserved protein involved in language-related disorders whose evolutionary changes are particularly relevant for understanding human cognitive traits [37]. This mSNC (7:113727420) in the *FOXP2* promoter adds new evidence for a putative modern human-specific regulation of *FOXP2* together with the nearly fixed intronic SNC that affects a transcription factor-binding site [37].

While we did not detect a specific enrichment in the modules containing *SETD1A/HMT* complex components *ASH2L* or *WDR82* genes, the IPC-cluster 2 midnightblue module, which contains *SETD1A*, shows an enrichment for a β -CATENIN-containing complex (*SETD7-YAP-AXIN1- β -CATENIN* complex; CORUM:6343; h.t.: adj p -value 0.05; Suppl. Mat. Table S8) and indeed contains WNT-effector *TCF3*, which harbors nearly fixed missense mutations in modern humans [6]. *SETD1A* is known to interact with β -CATENIN [38, 39] and increase its expression to promote neural progenitor proliferation [40].

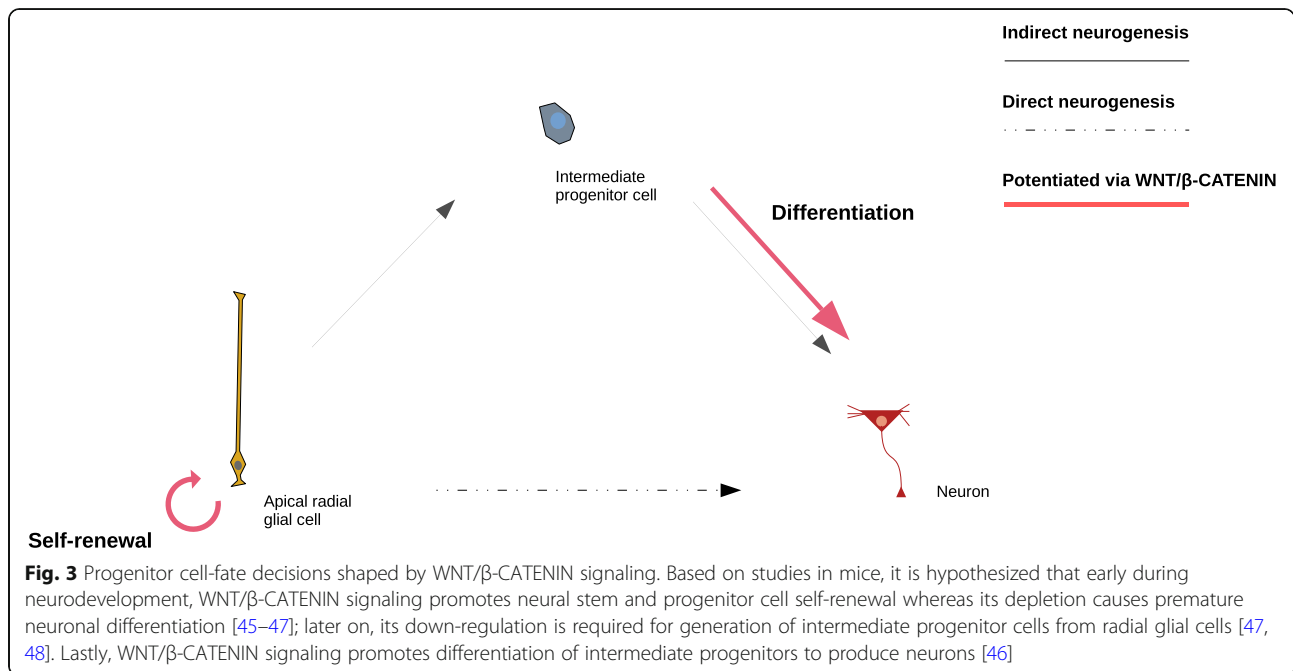
Discussion

By integrating data from paleogenomics and chromatin interaction and modification, we identified a set of genes controlled by regulatory regions that are active during early cortical development and contain single nucleotide changes that appeared in the modern human lineage after the split from the Neanderthal/Denisovan lineage. The regulatory regions reported here significantly overlap with putative modern human positively-selected regions and schizophrenia genomic loci, and control a set of genes among which we find a high number related to chromatin regulation, and most specifically the *SETD1A/HMT* complex. Regulators of chromatin dynamics are known to play key roles during cell-fate decisions through the control of specific transcriptional programs [41–43]. Both *SETD1A* and *ASH2L*, core components of the HMT complex, regulate WNT/ β -

CATENIN signaling [38–40, 44], which influences cell-fate decisions by promoting either self-maintenance or differentiation depending on the stage of progenitor differentiation (Fig. 3).

SETD1A (fixed mSNC in promoter), implicated in schizophrenia and developmental language impairment [49, 50], acts in collaboration with a histone chaperone to promote proliferation of neural progenitor cells through H3K4 trimethylation at the promoter of β -CATENIN, while its knockdown causes reduction in proliferative neural progenitor cells and an increase in cells at the cortical plate [40]. In addition, one of *SETD1A* direct targets is the WNT-effector *TCF4* [51], whose promoter also harbors a mSNC. Similarly, *ASH2L* specifically regulates WNT signaling: Conditional knockout of *ASH2L* significantly compromises the proliferative capacity of RGCs and IPCs by the time of generation of upper-layer neurons, with these progenitor cells showing a marked reduction in H3K4me3 levels and downregulation of WNT/ β -CATENIN signaling-related genes (defects that can be rescued by over-expression of β -CATENIN) [44]. Taken together, depletion of components of the *SETD1A/HMT* complex impairs the proliferative capacity of progenitor cells, altering the indirect mode of neurogenesis, with a specific regulation of the conserved WNT signaling. Interestingly, in addition to the aforementioned properties of the regulatory regions of the *SETD1A* complex components found in modern humans (overlap with modern human positively-selected regions and containing mSNCs), a recent work studying species-specific differences in chromatin accessibility using brain organoids reported that regulatory regions associated to *SETD1A* and *WDR82* were found in differentially-accessible regions in human organoids in comparison to chimpanzee organoids, with the *SETD1A* region overlapping with a human-gained histone modification signal when compared to macaques [52].

The dysfunction of chromatin regulators is among the most salient features behind causative mutations in neurodevelopmental disorders [53]. Our data highlights chromatin modifiers and remodelers that play prominent roles in neurodevelopmental disorders affecting brain growth and facial features. Along with the aforementioned chromatin regulators *PHC1* (microcephaly) and *KDM6A* (Kabuki syndrome), another paradigmatic example is the ATP-dependent chromatin remodeler *CHD8* (mSNC in enhancer), which controls neural progenitor cell proliferation through WNT-signaling related genes [54, 55]. *CHD8* is a high-risk factor for autism spectrum disorder and patients with *CHD8* mutations characteristically present macrocephaly and distinctive facial features [13]. Intriguingly, another ATP-dependent chromatin remodeler, *CHD2* (mSNC in enhancer),



presents a motif in the SETD1A promoter region containing the fixed mSNC (16:30969654; UCSC Genome Browser).

We have focused on the early stages of cortical development. While single-cell gene expression data of neural progenitor cells still remains limited, future integration of these data with other datasets covering different neo-cortical regions [56] will shed further light on modern human changes and cortical areas-specific progenitor cells. We acknowledge, in addition, that the genetic changes distinguishing modern humans and Neanderthals/Denisovans may be relevant at other stages of neurodevelopment, including the adult human brain. Progress in single-cell multi-omic technologies applied to brain organoid research will be critical to assess the impact of such changes in the diverse neural and non-neural cell-types through different developmental stages. Moreover, we excluded the examination of regulatory regions harboring Neanderthal/Denisovan changes due to the low number of high-quality genomes from Neanderthal/Denisovan individuals, which makes the determination of allele frequency in these species unreliable. We hope that the availability of a higher number of high-quality genomes for these species in the future will make such examination feasible.

Conclusions

This study complements previous research focused on protein-coding changes [4, 6] and helps extend the investigation of species-specific differences in cortical development that has so far relied on detailed comparisons between humans and non-human primates [52, 57–60].

We provide a list of new candidate genes for the study of human species-specific differences during the early stages of cortical development. The study of modern human evolutionary changes affecting chromatin regulators integrated with the examination of neurodevelopmental disorders could be a valuable entry point to understand modern human-specific brain ontogenetic trajectories.

Methods

Data processing

Integration and processing of data from different sources was performed using IPython v5.7.0. We used publicly available data from [6] of SNC in the modern human lineage (at fixed or above 90% frequency in present-day human populations) and Neanderthal/Denisovan changes. [6] analyzed high-coverage genotypes from one Denisovan and two Neanderthal individuals to report a catalog of SNC that appeared in the modern human lineage after their split from Neanderthals/Denisovans. Similarly, [6] also reported a list of SNC present in the Neanderthal/Denisovan lineages where modern humans carry the inferred ancestral allele.

For enhancer–promoter linkages, we used publicly available data from [61], based on transposase-accessible chromatin coupled to sequencing and integrated with chromatin capture via Hi-C data, from 15 to 17 post-conception weeks of the developing human cortex. A total of 92 promoters and 113 enhancers were selected as harboring mSNC and being depleted of Neanderthal/Denisovan SNC (from a total of 2574 enhancers and 1553 promoters present in the original dataset). Additionally, we completed the previous dataset filtering

annotated enhancer-gene linkages via Hi-C from the adult prefrontal cortex [62] (PsychENCODE resource portal: <http://resource.psychencode.org/>). In this case, enhancers ($n = 32,803$) were selected for further analyses if their coordinates completely overlapped with signals of active enhancers (H3K27ac) (that do not overlap with promoter signals (H3K4me3)) from the developing human cortex between 7 to 12 post-conception weeks [63]. A total of 43 enhancers, containing mSNC but free of Neanderthal/Denisovan SNC, passed this filtering. As a whole, the final integrated dataset covered regulatory regions active at early stages of human prenatal cortical development and linked to 212 genes. The coordinates (hg19 version) of the regulatory regions containing mSNC are available in the [Supplementary Material Tables S1 & S2](#).

Human positively-selected regions coordinates were retrieved from [8].

Single-cell RNA-seq analysis

The single-cell transcriptomic analysis was performed using the Seurat package v2.4 [64] in RStudio v1.1.463 (server mode).

Single-cell gene expression data was retrieved from [63] from PsychENCODE portal (<http://development.psychencode.org/#>). We used raw gene counts thresholding for cells with a minimum of 500 genes detected and for genes present at least in 10% of the total cells ($n = 762$). Data was normalized using “LogNormalize” method with a scale factor of 1,000,000. We regressed out cell-to-cell variation due to mitochondrial and cell-cycle genes (*ScaleData* function). For the latter, we used a list of genes [65] that assigns scores genes to either G1/S or G2/M phase (function *CellCycleScoring*), allowing us to reduce heterogeneity due to differences in cell-cycle phases. We further filtered cells (*FilterCells* function) setting a low threshold of 2000 and a high threshold of 9000 gene counts per cell, and a high threshold of 5% of the total gene counts for mitochondrial genes.

We assigned the label ‘highly variable’ to genes whose average expression value was between 0.5 and 8, and variance-to-mean expression level ratio between 0.5 and 5 (*FindVariableGenes* function). We obtained a total of 4261 genes for this category. Next, we performed a principal component analysis on highly variable genes and determined significance by a JackStraw analysis (*JackStraw* function). We used the first most significant principal components ($n = 13$) for clustering analysis (*FindClusters* function; resolution = 3). Data was represented in two dimensions using t-distributed stochastic neighbor embedding (*RunTSNE* function). The resulting twelve clusters were plotted using *tSNEplot* function. Cell-type assignment was based on the metadata from the original publication [63].

Weighted gene co-expression network analysis

For the gene co-expression network analysis we used the WGCNA R package [66, 67]. For each cluster of progenitor cells (RGC-1, 34 cells (15,017 genes); RGC-2, 30 cells (14,747 genes); IPC-1, 52 cells (15,790 genes); IPC-2, 41 cells (15,721 genes); IPC-3 was excluded due to low number of cells), log-transformed values of gene expression data were used as input for weighted gene co-expression network analysis. A soft threshold power was chosen (12, 12, 14, 12 for RGC-1, RGC-2, IPC-1, IPC-2 clusters, with R^2 : 0.962, 0.817, 0.961, 0.918, respectively) and a bi-weight mid-correlation applied to compute a signed weighted adjacency matrix, transformed later into a topological overlap matrix. Module detection (minimum size 200 genes) was performed using function *cutreeDynamic* (method = ‘hybrid’, deepSplit = 2), getting a total of 32, 26, 9, 23 modules for RGC-1, RGC-2, IPC-1, IPC-2, respectively (Suppl. Figs. 3 & 4).

Enrichment analysis

We ranked regulatory regions by mutation density calculating number of single nucleotide changes per regulatory region length (for those regions spanning at least 1000 base pairs). Top candidates were those ranking in the distribution within the 5% out of the total number of enhancers or promoters (Suppl. Mat Tables S3 & S4). The g:Profiler2 R package [29] was used to perform enrichment analyses (hypergeometric test; correction method ‘gSCS’; background genes: ‘only annotated genes’, *Homo sapiens*) for gene/phenotype ontology categories, biological pathways (KEGG, Reactome) and protein databases (CORUM, Human Protein Atlas) for the gene lists generated in this study. Permutation tests (10,000 permutations) were performed to evaluate enrichment of enhancers/promoters regions in different genomic regions datasets using the R package *regionR* [68]. The Hypergeometric Optimization of Motif EnRichment (HOMER) software v4.10 [69] was employed for motif discovery analysis, selecting best matches (Benjamini q -value < 0.05) of known motifs ($n = 428$; ChIP-seq-based) in our promoter and enhancer datasets.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12864-020-6706-x>.

Additional file 1. Supplementary Material.

Additional file 2. Supplementary Figures.

Abbreviations

SNC: Single-nucleotide change; mSNC: Modern human single-nucleotide change; HMT: Histone methyltransferase complex; h.t.: Hypergeometric test; RGC: Radial glial cell; IPC: Intermediate progenitor cell

Acknowledgements

We thank members of the 'Cognitive Biology of Language' group for helpful discussions.

Author's contributions

Conceptualization: C.B. & J.M.; Data Curation: J.M.; Formal Analysis: J.M.; Funding Acquisition: C.B.; Investigation: C.B. & J.M.; Methodology: C.B. & J.M.; Software: J.M.; Supervision: C.B.; Visualization: C.B. & J.M.; Writing – Original Draft Preparation: C.B. & J.M.; Writing – Review & Editing: C.B. & J.M. The authors read and approved this manuscript.

Funding

C.B. acknowledges research funds from the Spanish Ministry of Economy and Competitiveness/FEDER (grant FFI2016–78034-C2–1-P), Marie Curie International Reintegration Grant from the European Union (PIRG-GA-2009-256413), research funds from the Fundació Bosch i Gimpera, MEXT/JSPS Grant-in-Aid for Scientific Research on Innovative Areas 4903 (Evolinguistics: JP17H06379), and Generalitat de Catalunya (Government of Catalonia) – 2017-SGR-341. The funding bodies played no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

The datasets supporting the conclusions of this article (tables and code) are available in the Figshare repository, under <https://doi.org/10.6084/m9.figshare.11603478.v1> and <https://doi.org/10.6084/m9.figshare.11608074.v1>. We also made use of web-based data resources from [62] and [63] through the PsychENCODE portal (<http://development.psychencode.org/>).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Universitat de Barcelona, Gran Via de les Corts Catalanes, Barcelona, Spain.

²Universitat de Barcelona Institute of Complex Systems, Martí Franquès, Barcelona, Spain. ³Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys, Barcelona, Spain.

Received: 8 November 2019 Accepted: 30 March 2020

Published online: 16 April 2020

References

- Pääbo S. The human condition—a molecular approach. *Cell*. 2014;157(1):216–26 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S009286741301605X>.
- Hublin J-J, Neubauer S, Gunz P. Brain ontogeny and life history in Pleistocene hominins. *Philos Transact Royal Soc B: Biol Sci*. 2015;370(1663):20140062 Available from: <http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2014.0062>.
- Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science*. 2012;338(6104):222–6 Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.1224344>.
- Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*. 505(7481):43–9 Available from: <http://www.nature.com/articles/nature12886>.
- Prüfer K, de Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, et al. A high-coverage Neandertal genome from Vindija cave in Croatia. *Science*. 2017;358(6363):655–8 Available from: <http://www.sciencemag.org/lookup/doi/10.1126/science.aao1887>.
- Kuhlwilm M, Boeckx C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci Rep*. 2019;9(1) Available from: <http://www.nature.com/articles/s41598-019-44877-x>.
- King M, Wilson A. Evolution at two levels in humans and chimpanzees. *Science*. 1975;188(4184):107–16 Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.1090005>.
- Peyrégne S, Boyle MJ, Dannemann M, Prüfer K. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res*. 2017;27(9):1563–72 Available from: <http://genome.cshlp.org/lookup/doi/10.1101/gr.219493.116>.
- Petr M, Pääbo S, Kelso J, Vernot B. Limits of long-term selection against Neandertal introgression. *Proc Natl Acad Sci*. 2019;116(5):1639–44 Available from: <http://www.pnas.org/lookup/doi/10.1073/pnas.1814338116>.
- Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? *Mol Cell*. 2013;49(5):825–37 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1097276513001020>.
- Saudou F, Humbert S. The biology of huntingtin. *Neuron*. 2016;89(5):910–26 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0896627316000969>.
- Lai CSL, Fisher SE, Hurst JA, Vargha-Khadem F, Monaco AP. A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature*. 2001;413(6855):519–23 Available from: <http://www.nature.com/articles/35097076>.
- Bernier R, Golzio C, Xiong B, Stessman HA, Coe BP, Penn O, et al. Disruptive CHD8 mutations define a subtype of autism early in development. *Cell*. 2014;158(2):263–76 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0092867414007491>.
- Parras A, Anta H, Santos-Galindo M, Swarup V, Elorza A, Nieto-González JL, et al. Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing. *Nature*. 2018;560(7719):441–6 Available from: <http://www.nature.com/articles/s41586-018-0423-5>.
- Zweier C, Peippo MM, Hoyer J, Sousa S, Bottani A, Clayton-Smith J, et al. Haploinsufficiency of TCF4 causes Syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). *Am J Hum Genet*. 2007;80(5):994–1001 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S000292707609562>.
- Forrest MP, Hill MJ, Kavanagh DH, Tansey KE, Waite AJ, Blake DJ. The psychiatric risk gene transcription factor 4 (TCF4) regulates neurodevelopmental pathways associated with schizophrenia, autism, and intellectual disability. *Schizophr Bull*. 2018;44(5):1100–10 Available from: <https://academic.oup.com/schizophreniabulletin/article/44/5/1100/4700989>.
- Kalff-Suske M, Wild A, Topp J, Wessling M, Jacobsen E-M, Bornholdt D, et al. Point mutations throughout the GLI3 gene cause Greig Cephalopolysyndactyly syndrome. *Hum Mol Genet*. 1999;8(9):1769–77 Available from: <https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/8.9.1769>.
- Awad S, Al-Dosari MS, Al-Yacoub N, Colak D, Salih MA, Alkuraya FS, et al. Mutation in PHC1 implicates chromatin remodeling in primary microcephaly pathogenesis. *Hum Mol Genet*. 2013;22(11):2200–13 Available from: <https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddt072>.
- Fuentes J-J, Pritchard MA, Planas AM, Bosch A, Ferrer I, Estivill X. A new human gene from the Down syndrome critical region encodes a proline-rich protein highly expressed in fetal brain and heart. *Hum Mol Genet*. 1995;4(10):1935–44 Available from: <https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/4.10.1935>.
- Poirier K, Lebrun N, Broix L, Tian G, Saillour Y, Boscheron C, et al. Mutations in TUBG1, DYNC1H1, KIF5C and KIF2A cause malformations of cortical development and microcephaly. *Nat Genet*. 2013;45(6):639–47 Available from: <http://www.nature.com/articles/ng.2613>.
- Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014;511(7510):421–7 Available from: <http://www.nature.com/articles/nature13595>.
- Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium, BUPGEN, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, 23andMe Research Team, Grove J, Ripke S, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet*. 2019;51(3):431–44 Available from: <http://www.nature.com/articles/s41588-019-0344-8>.
- Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2018;47(D1):D1005–12. <https://doi.org/10.1093/nar/gky1120>.

24. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, et al. Extensive variation in chromatin states across humans. *Science*. 2013;342(6159):750–2 Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.1242510>.
25. Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*. 2013;342(6159):744–7 Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.1242463>.
26. McVicker G, van de Geijn B, Degner JF, Cain CE, Banovich NE, Raj A, et al. Identification of genetic variants that affect histone modifications in human cells. *Science*. 2013;342(6159):747–9 Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.1242429>.
27. Piunti A, Shilatfard A. Epigenetic balance of gene expression by Polycomb and COMPASS families. *Science*. 2016;352(6290):aad9780 Available from: <http://www.sciencemag.org/lookup/doi/10.1126/science.aad9780>.
28. Florio M, Huttner WB. Neural progenitors, neurogenesis and the evolution of the neocortex. *Development*. 2014;141(11):2182–94 Available from: <http://dev.biologists.org/cgi/doi/10.1242/dev.090571>.
29. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. G:profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*. 2007;35 Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkm226>.
30. Somel M, Franz H, Yan Z, Lorenc A, Guo S, Giger T, et al. Transcriptional neoteny in the human brain. *Proc Natl Acad Sci*. 2009;106(14):5743–8 Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.0900544106>.
31. Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, et al. Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. *Genome Res*. 2012;22(4):611–22 Available from: <http://genome.cshlp.org/cgi/doi/10.1101/gr.127324.111>.
32. Lesciotta KM, Richtsmeier JT. Craniofacial skeletal response to encephalization: how do we know what we think we know? *Am J Phys Anthropol*. 2019;168(S67):27–46 Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ajpa.23766>.
33. Miyake N, Koshimizu E, Okamoto N, Mizuno S, Ogata T, Nagai T, et al. *MLL2* and *KDM6A* mutations in patients with kabuki syndrome. *Am J Med Genet*. 2013;161(9):2234–43. <https://doi.org/10.1002/ajmg.a.36072>.
34. Adhikari K, Fuentes-Guajardo M, Quinto-Sánchez M, Mendoza-Revilla J, Camilo Chacón-Duque J, Acuña-Alonzo V, et al. A genome-wide association scan implicates *DCHS2*, *RUNX2*, *GLI3*, *PAX1* and *EDAR* in human facial variation. *Nat Commun*. 2016;7(1):11616 Available from: <http://www.nature.com/articles/ncomms11616>.
35. Wang L, Hou S, Han Y-G. Hedgehog signaling promotes basal progenitor expansion and the growth and folding of the neocortex. *Nat Neurosci*. 2016;19(7):888–96 Available from: <http://www.nature.com/articles/nn.4307>.
36. Pollen AA, Nowakowski TJ, Chen J, Retallack H, Sandoval-Espinosa C, Nicholas CR, et al. Molecular identity of human outer radial glia during cortical development. *Cell*. 2015;163(1):55–67 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0092867415011241>.
37. Fisher SE. Human genetics: the evolving story of *FOXP2*. *Curr Biol*. 2019;29(2):R65–7 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S096098221831546X>.
38. Hoffmeyer K, Raggioli A, Rudloff S, Anton R, Hierholzer A, Del Valle I, et al. Wnt/beta-catenin signaling regulates telomerase in stem cells and cancer cells. *Science*. 2012;336(6088):1549–54 Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.1218370>.
39. Salz T, Li G, Kaye F, Zhou L, Qiu Y, Huang S. *hSETD1A* regulates Wnt target genes and controls tumor growth of colorectal cancer cells. *Cancer Res*. 2014;74(3):775–86 Available from: <http://cancerres.aacrjournals.org/content/74/3/775>.
40. Li Y, Jiao J. Histone chaperone HIRA regulates neural progenitor cell proliferation and neurogenesis via beta-catenin. *J Cell Biol*. 2017;216(7):1975–92 Available from: <http://www.jcb.org/lookup/doi/10.1083/jcb.201610014>.
41. Chen T, Dent SYR. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat Rev Genet*. 2014;15(2):93–106 Available from: <http://www.nature.com/articles/nrg3607>.
42. Hirabayashi Y, Gotoh Y. Epigenetic control of neural precursor cell fate during development. *Nat Rev Neurosci*. 2010;11(6):377–88 Available from: <http://www.nature.com/articles/nrn2810>.
43. Tuoc TC, Pavlakis E, Tylkowski MA, Stoykova A. Control of cerebral size and thickness. *Cell Mol Life Sci*. 2014;71(17):3199–218 Available from: <http://link.springer.com/10.1007/s00188-014-1590-7>.
44. Li L, Ruan X, Wen C, Chen P, Liu W, Zhu L, et al. The COMPASS family protein ASH2L mediates Corticogenesis via transcriptional regulation of Wnt signaling. *Cell Rep*. 2019;28(3):698–711.e5 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2211124719308289>.
45. Chenn A, Walsh CA. Regulation of cerebral cortical size by control of cell cycle exit in neural precursors. *Science*. 2002;297(5580):365–9 Available from: <https://science.sciencemag.org/content/297/5580/365>.
46. Munji RN, Choe Y, Li G, Siegenthaler JA, Pleasure SJ. Wnt signaling regulates neuronal differentiation of cortical intermediate progenitors. *J Neurosci*. 2011;31(5):1676–87 Available from: <http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5404-10.2011>.
47. Draganova K, Zemke M, Zurkirchen L, Valenta T, Cant' u C, Okoniewski M, et al. Wnt/Beta-catenin Signaling546Regulates sequential fate decisions of murine cortical precursor cells: Beta-catenin signaling Regulates547Sequential neural fate. *Stem Cells*. 2015;33(1):170–82 Available from: <http://doi.wiley.com/10.1002/stem.1820>.
48. Mutch CA, Schulte JD, Olson E, Chenn A. Beta-catenin signaling negatively regulates intermediate progenitor population numbers in the developing cortex. *PLoS One*. 2010;5(8):e12376 Available from: <https://dx.plos.org/10.1371/journal.pone.0012376>.
49. Takata A, Xu B, Ionita-Laza I, Roos J, Gogos JA, Karayiorgou M. Loss-of-function variants in schizophrenia risk and *SETD1A* as a candidate susceptibility gene. *Neuron*. 2014;82(4):773–80 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0896627314003584>.
50. Eising E, Carrion-Castillo A, Vino A, Strand EA, Jakielski KJ, Scerri TS, et al. A set of regulatory genes co-expressed in embryonic human brain is implicated in disrupted speech development. *Mol Psychiatry*. 2019;24(7):1065–78 Available from: <http://www.nature.com/articles/s41380-018-0020-x>.
51. Mukai J, Cannavò E, Crabtree GW, Sun Z, Diamantopoulou A, Thakur P, et al. Recapitulation and reversal of schizophrenia-related phenotypes in *Setd1a*-deficient mice. *Neuron*. 2019; Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0896627319307871>.
52. Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature*. 2019;574(7778):418–22 Available from: <https://www.nature.com/articles/s41586-019-1654-9>.
53. Gabriele M, Tobon AL, D'Agostino G, Testa G. The chromatin basis of neurodevelopmental disorders: rethinking dysfunction along the molecular and temporal axes. *Prog Neuro-Psychopharmacol Biol Psychiatry*. 2018;84:306–27 Available from: <http://www.sciencedirect.com/science/article/pii/S027854617305389>.
54. Sugathan A, Biagioli M, Golzio C, Erdin S, Blumenthal I, Manavalan P, et al. *CHD8* regulates neurodevelopmental pathways associated with autism spectrum disorder in neural progenitors. *Proc Natl Acad Sci*. 2014;111(42):E4468–77 Available from: <http://www.pnas.org/lookup/doi/10.1073/pnas.1405266111>.
55. Durak O, Gao F, Kaeser-Woo YJ, Rueda R, Martorell AJ, Nott A, et al. *Chd8* mediates cortical neurogenesis via transcriptional regulation of cell cycle and Wnt signaling. *Nat Neurosci*. 2016;19(11):1477–88 Available from: <http://www.nature.com/articles/nn.4400>.
56. Nowakowski TJ, Bhaduri A, Pollen AA, Alvarado B, Mostajo-Radji MA, Di Lullo E, et al. Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science*. 2017;358(6368):1318–23 Available from: <http://www.sciencemag.org/lookup/doi/10.1126/science.aap8809>.
57. Mora-Bermúdez F, Badsha F, Kanton S, Camp JG, Vernot B, Köhler K, et al. Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. *Musacchio a, editor. eLife*. 2016 Sep;5:e18683. <https://doi.org/10.7554/eLife.18683>.
58. Otani T, Marchetto MC, Gage FH, Simons BD, Livesey FJ. 2D and 3D stem cell models of primate cortical development identify species-specific differences in progenitor behavior contributing to brain size. *Cell Stem Cell*. 2016;18(4):467–80 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1934590916001090>.
59. Marchetto MC, Hrvov-Mihic B, Kerman BE, Yu DX, Vadodaria KC, Linker SB, et al. Species-specific maturation profiles of human, chimpanzee and bonobo neural cells. *Zoghbi HY, Arlotta P, editors. eLife*. 2019;8:e37527. <https://doi.org/10.7554/eLife.37527>.
60. Pollen AA, Bhaduri A, Andrews MG, Nowakowski TJ, Meyerson OS, Mostajo-Radji MA, et al. Establishing cerebral organoids as models of human-specific brain evolution. *Cell*. 2019;176(4):743–756.e17 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0092867419300509>.

61. de la Torre-Ubieta L, Stein JL, Won H, Opland CK, Liang D, Lu D, et al. The dynamic landscape of open chromatin during human cortical neurogenesis. *Cell*. 2018;172(1–2):289–304.e18 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0092867417314940>.
62. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Ashley-Koch AE, Crawford GE, Garrett ME, Song L, Safi A, Johnson GD, et al, editors. *Science*. 2018;362(6420) Available from: <https://science.sciencemag.org/content/362/6420/eaat8464>.
63. Li M, Santpere G, Imamura Kawasawa Y, Evgrafov OV, Gulden FO, Pochareddy S, et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*. 2018;362(6420):eaat7615 Available from: <http://www.sciencemag.org/lookup/doi/10.1126/science.aat7615>.
64. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 2018;36:411. <https://doi.org/10.1038/nbt.4096>.
65. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. 2016;352(6282):189–96 Available from: <http://www.sciencemag.org/cgi/doi/10.1126/science.aad0501>.
66. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559. <https://doi.org/10.1186/1471-2105-9-559>.
67. Langfelder P, Horvath S. Fast *r* functions for robust correlations and hierarchical clustering. *J Stat Softw*. 2012;46(11) Available from: <http://www.jstatsoft.org/v46/i11/>.
68. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*. 2015;btv562 Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv562>.
69. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*. 2010;38(4):576–89 Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1097276510003667>.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Chapter 3

A multi-layered integrative analysis reveals a cholesterol metabolic program in outer radial glia with implications for human brain evolution

Published as:


Moriano, J., Leonardi, O. Vitriolo, A., Testa, G., & Boeckx, C. 2023. A multi-layered integrative analysis reveals a cholesterol metabolic program in outer radial glia with implications for human brain evolution. *bioRxiv*

doi:[10.1101/2023.06.23.546307](https://doi.org/10.1101/2023.06.23.546307)

1 A multi-layered integrative analysis reveals a 2 cholesterol metabolic program in outer radial glia 3 with implications for human brain evolution

4 Juan Moriano ^{1,2} , Oliviero Leonardi ³, Alessandro Vitriolo ^{3,4}, Giuseppe Testa
5 ^{3,4}, Cedric Boeckx ^{1,2,5,6} 

6 ¹University of Barcelona; ²University of Barcelona Institute of Complex Systems;
7 ³Human Technopole, Viale Rita Levi-Montalcini 1, 20157, Milan, Italy; ⁴Department of
8 Oncology and Hemato-Oncology, University of Milan, Via Santa Sofia 9, 20122, Milan,
9 Italy; ⁵University of Barcelona Institute of Neurosciences; ⁶Catalan Institute for
10 Research and Advanced Studies (ICREA)

11  **For correspondence:**
jmoriano@ub.edu,
cedric.boeckx@ub.edu

Present address: University
of Barcelona, Gran Via 585,
08007 Barcelona, Spain

Data availability: Code is
available at
[https://github.com/jjaa-mp/
MultiLayered_IndirectNeuro/](https://github.com/jjaa-mp/MultiLayered_IndirectNeuro/)
and datasets supporting the
conclusions of this work are
available as Supplementary
Material

Funding: Generalitat de
Catalunya (2021-SGR-313;
FI-SDUR 2020); Spanish
Ministry of Science and
Innovation
(PID2019-107042GB-I00)

Competing interests: The
author declare no competing
interests.

12 Abstract

13 The definition of molecular and cellular mechanisms contributing to evolutionary divergences in
14 brain ontogenetic trajectories is essential to formulate hypotheses about the emergence of our
15 species. Yet the functional dissection of evolutionary modifications derived in the *Homo sapiens*
16 lineage at an appropriate level of granularity remains particularly challenging. Capitalizing on
17 recent single-cell sequencing efforts that have massively profiled neural stem cells from the
18 developing human cortex, we develop an integrative computational framework in which we
19 perform (i) trajectory inference and gene regulatory network reconstruction, (ii)
20 (pseudo)time-informed non-negative matrix factorization for learning the dynamics of gene
21 expression programs, and (iii) paleogenomic analysis for a higher-resolution mapping of the
22 regulatory landscape where our species acquired derived mutations in comparison to our closest
23 relatives. We provide evidence for cell type-specific activation and regulation of gene expression
24 programs during indirect neurogenesis. In particular, our analysis uncovers a zinc-finger
25 transcription factor, KLF6, as a key regulator of a cholesterol metabolic program specifically in
26 outer radial glia. Our strategy allows us to further probe whether the (semi)discrete gene
27 expression programs identified have been under selective pressures in our species lineage. A
28 cartography of the regulatory landscape impacted by *Homo sapiens*-derived transcription factor
29 binding site disruptions reveals signals of selection clustering around regulatory regions
30 associated with *GLI3*, a well-known regulator of the radial glial cell cycle. As a whole, our study
31 contributes to the evidence of significant changes impacting metabolic pathways in recent
32 human brain evolution.

33 Introduction

34 A large number of studies has unveiled genetic, molecular and cellular features that contribute to
35 species-specific mechanisms of corticogenesis in the primate lineage. These comprise, but are not
36 limited to, transcriptomic divergence, emergence of novel genes, substitutions in regulatory ele-
37 ments, control of the timing of neural proliferation and differentiation, or progenitor diversity and
38 abundance (some recent comprehensive reviews include Pinson and Huttner, 2021; Libé-Philippot
39 and Vanderhaeghen, 2021; Pollen, Kilik, et al., 2023; Vanderhaeghen and Polleux, 2023). Addi-

41 tionally, following the availability of genomes from extinct species most closely related to us, the
42 elucidation of the molecular underpinnings of unique aspects of brain organization in *Homo sapi-*
43 *ens*, going beyond sheer brain size, is now on the research horizon (Pääbo, 2014), and suggestive
44 evidence for developmental differences is already available (Trujillo et al., 2021; Mora-Bermúdez,
45 Kanis, et al., 2022; Stepanova et al., 2021; Pinson, Xing, et al., 2022).

46 The large scale and high resolution afforded by single-cell sequencing technologies, coupled
47 with increasingly powerful computational approaches, have significantly contributed to our under-
48 standing of the identity, heterogeneity and developmental progression of neural progenitors. Yet,
49 substantial gaps exist in our knowledge of the regulatory mechanisms implicated in neural progen-
50 itor proliferation and differentiation during corticogenesis, and how these mechanisms may have
51 been modified over the course of human evolution.

52 During neurogenesis, two main proliferative regions can be identified in the dorsal telencephalon.
53 The ventricular zone is populated by ventricular radial glia (vRG), which serve as a scaffold for the
54 growing neocortex as well as a stem cell pool capable of self-renewal and differentiation (Silbereis
55 et al., 2016). And, the subventricular zone (SVZ), which subsequently emerges and expands due
56 to the asymmetric division of vRG and the self-renewal capacity of basal progenitors sustained
57 over a prolonged period (Silbereis et al., 2016). Two main types of basal progenitors can be distin-
58 guished: outer radial glial cells (oRG), which retain similar features to vRG, present distinctive mor-
59 phologies linked to their self-renewal capacity and typically express markers such as *HOPX* (Pollen,
60 Nowakowski, et al., 2015; Kalebic and Huttner, 2020); and intermediate progenitor cells (IPC), short-
61 lived progenitors with characteristic multipolar morphologies and which express *EOMES* (Pollen,
62 Nowakowski, et al., 2015; Pebworth et al., 2021).

63 Neurogenesis from basal progenitors, as opposed to the direct route from vRG to neuron, is re-
64 ferred to as indirect neurogenesis, and is thought to be responsible for the generation of the vast
65 majority of upper layer neurons (Lui, Hansen, and Kriegstein, 2011). Indeed, the developmental
66 period for supragranular layer neuron generation coincides with the appearance of a discontin-
67 uous radial glia scaffold where the SVZ remains as the main proliferative niche (Nowakowski et
68 al., 2016). There is evidence that the neocortical expansion in the primate lineage that most dra-
69 matically affected cortical upper layer neurons, and species-specific features of brain organization,
70 are intimately connected to the regulatory mechanisms that govern the behavior and modes of
71 division of neural progenitor cells (Rakic, 1995; Kriegstein, Noctor, and Martínez-Cerdeño, 2006).

72 Here we seek to provide a high-resolution characterization of gene regulatory networks at play
73 during indirect neurogenesis and ask whether there is evidence of evolutionary modifications of
74 the (semi)discrete gene expression programs emerging from the modular nature of the regula-
75 tory networks we identified. To do so, we leverage an integrative computational framework in
76 which to perform (i) trajectory inference and gene regulatory network reconstruction, (ii) inference
77 of the dynamics of gene expression programs via the implementation of a novel (pseudo)time-
78 informed non-negative matrix factorization method, and (iii) a paleogenomic analysis yielding a
79 higher-resolution mapping of the regulatory landscape where our species acquired derived single
80 nucleotide mutations in comparison to our closest relatives, both extinct and extant, for which
81 high-coverage genomes are available.

82 Using this framework, we resolve the bifurcation tree defining apical progenitor differentiation
83 towards either outer radial glia or intermediate progenitor cells and characterize waves of gene ex-
84 pression programs activated differentially among the neural lineages leading to each basal progen-
85 itor subtype. Among cell type-specific transcription factor-gene interactions, we uncover a previ-
86 ously unnoticed transcription factor, *KLF6*, as a putative master regulator of a cholesterol metabolic
87 program specific to the differentiation route leading to outer radial glia. An evolutionary-informed
88 analysis of transcription factor binding site disruptions leads to the hypothesis of a human-specific
89 regulatory modification of the *KLF6*-mTOR signaling axis in outer radial glia, with an important role
90 played by transcription factor *GLI3*, for which we identified changes associated with signals of pos-
91 itive selection in our species.

92 Results

93 Inferring neural progenitor states during indirect neurogenesis from the develop- 94 ing human cortex

95 Exploiting the potential of high-throughput single-cell sequencing to capture intermediate cellular
96 states during neural cell differentiation, we first sought to characterize the main axis of variation of
97 neural progenitor cells from the developing human cortex at around mid-gestation (Trevino et al.,
98 2021) (Figure 1A). Principal Component analysis (PCA) reveals a marked distinction among cell clus-
99 ters: the first principal component discriminates among progenitor types, that is, radial glial cells
100 and intermediate progenitors, while the second principal component captures the differentiation
101 state, from ventricular radial glia to basal progenitors (see Figure 1). Among genes that contribute
102 the most to each axis, we find markers of progenitor subtypes: e.g., *VIM* and *FOS* for ventricular
103 radial glia, *HOPX* and *PTPRZ1* for outer radial glia, or *EOMES* and *SSTR2* for intermediate progenitor
104 cells (see Figure 1C). Coherently, a differential expression analysis on a coarse clustering identifies
105 well-known markers for each subtype (Figure 1 S1A). Samples from different batches intermix in
106 the low dimensional space, confirming the absence of a significant contribution of technical arti-
107 facts (Figure 1 S1A).

108 To test our ability to reconstruct the apical-to-basal neural lineage trajectories, we performed
109 principal graph learning and computed a force-directed graph where we projected the inferred tree
110 of principal points (Methods & Materials). We obtained a bifurcating tree that resolves the molecu-
111 lar continuum describing the progression of ventricular radial glia and branching into either outer
112 radial glia or intermediate progenitor fates. The expression of the aforementioned marker genes
113 recapitulates the expected dynamics along pseudotime (Figure 1F; as well as that of genes whose
114 expression trajectories significantly change along the inferred tree, see Figure 1G), confirming the
115 differentiation progression through intermediate cellular states.

116 We obtained similar results when an independent dataset was projected into the low dimen-
117 sional space obtained before via PCA (Polioudakis et al., 2019); Figure 1 S1B to F). This provides
118 an ideal setting in which to test the validity of our results with time-matched samples around post-
119 conception week 16, a developmental stage with active proliferation in both germinal zones and
120 at around the transition from continuous to discontinuous radial glia scaffold (Nowakowski et al.,
121 2016).

122 A pseudotime-informed non-negative matrix factorization to identify dynamic gene 123 expression programs

124 We next sought to characterize how gene expression programs unfold as indirect neurogenesis
125 proceeds. A key analytical challenge associated with high-throughput single cell profiling is the
126 ability to extract meaningful patterns from high-dimensional datasets. To overcome this obstacle,
127 we developed a two-step computational strategy aimed at recovering the dynamics of gene expres-
128 sion programs during neural progenitor cell differentiation (Methods & Materials). Our approach
129 consists of:

- 130 1. A pseudotime-informed non-negative matrix factorization (piNMF) as the core algorithm to
131 capture the underlying structure of a high-dimensional dataset, explicitly accounting for the
132 continuous nature of gene expression trajectories through pseudotime, building on recent
133 computational advances on NMF using parametrizable functions (Hautecoeur and Glineur,
134 2020); and
- 135 2. An iterative strategy where stable gene expression programs are recovered by performing K-
136 means clustering over multiple replicates of the matrix factorization core algorithm (following
137 strategy in Kotliar et al., 2019), thereby addressing the non-uniqueness problem of matrix
138 factorization approximation methods.

139 Our strategy departs from the standard NMF (hereafter, stdNMF), where matrix decomposi-

140 tion is achieved through a linear combination of vectors that does not model continuous signals,
141 such as dynamically changing gene expression trajectories. We evaluated the performance of both
142 piNMF and stdNMF approaches on four dominant gene expression programs inferred across cell
143 types and datasets (Methods & Materials; Figure 2A,B and Figure 2 S2). Both approaches recover
144 programs linked to cell cluster identities, which is expected since cell type signatures significantly
145 contribute to the variation detected in single-cell data. However, we observe that expression pro-
146 grams at intermediate states towards basal progenitor clusters are not clearly defined by stdNMF,
147 while piNMF finely resolves a sequential activation of expression programs (Figure 2A). A compar-
148 ison of statistically significant genes associated to each expression program using multiple least
149 squares regression reveals a higher congruence in gene module membership for programs linked
150 to vRG and oRG cell clusters (especially for outer radial glia, with 79% overlap; 0.35% for IPC) than
151 for transient expression programs (<25%; see Figure 2B). In line with this, we find that exclusive,
152 top-significant Gene Ontology (GO) terms in transient expression programs captured by piNMF
153 provide a better characterization of key biological processes, with terms directly relevant such as
154 neuroepithelial differentiation, neurogenesis or cerebral cortex absent in the stdNMF analysis (std-
155 NMF instead returns more generic terms related to cell-cycle and chromatin organization; see Fig-
156 ure 2 S1).

157 **A cholesterol metabolic program activated in the radial glial branch**
158 A comparison of expression modules between oRG or IPC clusters inferred via piNMF reveals neu-
159 ral cell biology-specific features. Congruently with the reported roles of gap junctions in coupling
160 radial glial cells (Lo Turco and Kriegstein, 1991), we find GO terms related to cell adhesion and gap
161 junction in the radial glia branch. Similarly, exclusively for the late expression programs (modules
162 3 and 4) of the radial glia branch, we observe terms related to glia identity such as glia cell projec-
163 tion or glial cell differentiation, as well as terms related to extracellular matrix, critical for radial
164 glia stemness (Fietz et al., 2012; Pollen, Nowakowski, et al., 2015). Among the exclusive terms over-
165 represented in the IPC branch we find G1 phase, including a key regulator of basal progenitor G1
166 phase-length cyclin D1 (Lange, Huttner, and Calegari, 2009), cell-cell signaling and Notch signaling
167 (Kawaguchi et al., 2008), as well as axon and cell projection terms (in agreement with a reported
168 activation of axogenesis-related genes in basal progenitors in mouse (Bedogni and Hevner, 2021))
169 (Supplementary Table ST1). These results indicate that the piNMF implemented here successfully
170 captures relevant molecular processes during neural cell differentiation.

171 Prominently, the module that is activated last in pseudotime and that pertains to the acquisition
172 of oRG identity returns an over-representation of genes involved in cholesterol metabolism (Figure
173 2D). For instance, we observe the activation of the expression of several enzymes of the cholesterol
174 biosynthesis pathway, such as the 3-hydroxy-3-methylglutaryl-coenzyme A (HMG-CoA) synthase 1,
175 which participates in a condensation reaction previous to the production of the cholesterol pre-
176 cursor mevalonate; or the mevalonate pyrophosphate decarboxylase (MVD), which catalyzes the
177 production of isoprenes for cholesterol synthesis. While the interplay of cholesterol metabolism
178 and neural progenitor cells still awaits systematic exploration (Namba et al., 2021), previous studies
179 using mice have revealed important roles for cholesterol in the context of cortical radial thickness
180 and neural stem cell proliferation and differentiation (Saito et al., 2009; Nourse et al., 2022; Corbeil
181 et al., 2010). Importantly, the prominence of cholesterol metabolism in the oRG cluster, absent
182 in IPC cluster gene expression modules, is replicated when analyzing an independent dataset (Po-
183 lioudakis et al., 2019) and additionally cross-validated by GO terms that are also captured by the
184 standard NMF despite gene module composition differences (ST1 and 2).

185 **A *KLF6*-centered regulatory network for the activation of a cholesterol metabolism** 186 **program in human radial glia**

187 We next proceeded to the identification of key regulators of gene expression programs active dur-
188 ing neural progenitor cell fate dynamics. We performed a gene regulatory network reconstruction

189 using the *CellOracle* software (Kamimoto et al., 2023). First, we identified replicated signals across
190 single-cell ATAC-seq studies on the developing human brain in order to create a brain atlas of open
191 chromatin regions (Methods & Materials). Second, we retained confident TF-target gene links from
192 the open chromatin region atlas for each cell cluster, based on a machine learning-based regres-
193 sion analysis on the single-cell gene expression data (Methods & Materials).

194 We evaluated the prominence of transcription factors and genes within the reconstructed net-
195 works for each progenitor subtype cluster according to the following network connectivity mea-
196 sures (as proposed in Kamimoto et al., 2023): eigenvector centrality, for overall relevance of a
197 given gene in a network according to the quality of its connections to other genes, and between-
198 ness centrality, i.e., the influence of a given gene in the transfer of information within a network.
199 Consistently across network measures and comparatively among cell clusters, we find the zinc
200 finger-containing transcription factor *KLF6* as one of the top-ranked genes in radial glial cells (Fig-
201 ure 3A,B and Figure 3 S2A). This is consistent with the gene's association to a super-interactive
202 promoter in radial glia (Song et al., 2020), but not in intermediate progenitor cells. Within radial
203 glia, *KLF6* occupies a more prominent position in the oRG cluster (these results were replicated in
204 an independent dataset (Polioudakis et al., 2019); Figure 3 S2A,B).

205 To gain further insight into the cell cluster-specific regulatory network associated with *KLF6*, we
206 compared its target genes in vRG and oRC cell clusters. *KLF6* targets in vRG are most significantly
207 related to biological processes that include responsiveness to abiotic stimulus and organic sub-
208 stances, regulation of apoptosis, neurogenesis or cell migration. By contrast, in the oRG cluster,
209 the *KLF6* transcriptional network is significantly over-represented in genes linked to cholesterol
210 and steroid biosynthesis, as indicated by GO terms such as cholesterol metabolism, regulation of
211 cholesterol biosynthesis by SREBF, and steroid biosynthesis or steroid metabolic process (Figure
212 3C,D; ST3). We performed a similar analysis on an independent dataset (Polioudakis et al., 2019)
213 and although we did not obtain a clear discrimination for *KLF6* roles in radial glia cell subtypes (with
214 few terms related to steroids in radial glia (ST3)), we examined the *KLF6* transcriptional network
215 reported in (Polioudakis et al., 2019), reconstructed using an independent GRN inference method,
216 and reported to be over-represented in outer radial glia and endothelial cell clusters, and here too
217 an enrichment for cholesterol metabolism emerged (ST3).

218 We find *KLF6* target genes across the four sequentially activated gene expression programs
219 detected by piNMF, and specifically enzymes of the cholesterol biosynthetic pathway in the latest-
220 activated module in oRG. As expected, *KLF6* targets present in piNMF modules are enriched in
221 cholesterol metabolism exclusively in the latest oRG module (ST4). Lastly, in agreement with the
222 reported roles of *KLF6* as a regulator of cholesterol metabolism via activation of mTOR signaling
223 and sterol regulatory element binding transcription factors (Syafuruddin et al., 2019), we detect
224 the mTOR signaling-related platelet-derived growth factor receptor *PDGFRB* and insulin-like-growth
225 factor binding protein *IGFBP2* as well as the GO term 'activation of gene expression by SREBF' in
226 the late piNMF module 4 (ST4).

227 Taken together, our results reveal a previously unnoticed transcription factor, *KLF6*, acting as a
228 central node for the activation of a cholesterol metabolic program in human radial glia.

229 **A paleogenomic interrogation of regulatory regions active during human cortico-** 230 **genesis**

231 In light of recent work mentioned in the introduction showing how some protein-coding muta-
232 tions (virtually) fixed across contemporary human populations but absent in closely related extinct
233 species affect various aspects of neural progenitor cell behavior, we decided to take advantage of
234 our comprehensive atlas of open chromatin regions active during human corticogenesis presented
235 above and focus on the still less well studied mutations in the regulatory regions of the genome,
236 aiming to identify points of divergence among closely related species that achieved similar brain
237 sizes (VanSickle, Cofran, and Hunt, 2020), but likely via distinct ontogenies (Hublin, Neubauer, and
238 Gunz, 2015), reflected in different neurocranial shapes.

239 To do so, we first isolated a set of regulatory regions that contain high-frequency *Homo sapiens*-
240 derived variants but crucially where the Neanderthals/Denisovans carry the ancestral allele (found
241 in non-human primate genomes used as reference). We call these 'regulatory islands', and defined
242 such regions in terms of a genomic window of 3,000 base pairs around each variant where the
243 Neanderthal/Denisovan homolog regions did not acquire species-specific, derived variants (Figure
244 4A; Methods & Materials). This led to the identification of a total of 4836 "regulatory islands" linked
245 to 4797 genes, complementing and extending recent efforts on regulatory variants derived in our
246 lineage (Moriano and Boeckx, 2020; Weiss et al., 2021; McArthur et al., 2022).

247 A substantial proportion of top marker genes for the previously characterized piNMF expres-
248 sion programs are found associated to regulatory islands, with a more pronounced abundance in
249 late, relative to early, modules in the oRG branch, while a more even distribution is observed in the
250 IPC branch (Figure 4B; for both datasets studied here, (Trevino et al., 2021) and (Polioudakis et al.,
251 2019); Figure 4 and ST6). Among the genes linked to regulatory islands we find key oRG markers
252 such as *HOPX*, *PTPRZ1*, *LIFR* or *MOXD1* (Pollen, Nowakowski, et al., 2015). We note that some of the
253 regulatory islands exhibit additional special properties in the context of recent human evolution
254 (ST6): this is the case of a region linked to *PTPRZ1* and another linked to *RB1CC1*, both found in
255 genomic region depleted of archaic introgression (so-called large "introgression deserts") (Chen
256 et al., 2020). Both genes are direct *KLF6* targets specifically in the oRG program uncovered by our
257 analysis above; of note, we have also identified a regulatory island associated to *KLF6*. Other reg-
258 ulatory islands are associated with signals of positive selection in the *sapiens* lineage compared to
259 extinct hominins (Peyr gne et al., 2017). This is the case for interacting regulators (T. Yang et al.,
260 2002; Sun et al., 2005) for cholesterol biosynthesis such as *SCAP* (which additionally carries a fixed
261 derived missense mutation in *Homo sapiens* (Kuhlwilm and Boeckx, 2019)) and *SEC24D*. It is also
262 the case for two regulatory islands mapping linked to *GLI3*. As a matter of fact, *GLI3*-associated reg-
263 ulator islands are the only ones found to be associated with signals of positive selection affecting
264 a transcription factor included in our *Cell Oracle* analysis.

265 Differential transcription factor binding analysis exhibits signals of positive selection in
266 *GLI3* regulatory islands

267 Differential transcription factor binding plays a key role in the divergence of gene regulation across
268 species (Villar, Flicek, and Odom, 2014; Zhang, Fang, and Huang, 2023), and indeed *Homo* species-
269 specific regulatory variants have been associated to differential gene expression in cell-line models
270 (Weiss et al., 2021). For this reason we examined whether variants found in regulatory islands
271 implicate disruptions of transcription factor binding sites (TFBS) by implementing the motifbreakR
272 predictive tool (S. G. Coetzee, G. A. Coetzee, and Hazelett, 2015).

273 With this approach we aimed to identify statistically significant relations involving TFs with
274 overall reduced, or increased, binding affinity in regulatory islands, and in a manner fine-grained
275 enough to discriminate between TFs impacting the regulation of genes found in early vs late mod-
276 ules in the piNMF expression programs discussed above (ST6). TFs with the highest number of
277 increased binding affinity sites are associated with regulation of the adaptive response to hypoxia
278 and various metabolic processes including lipid metabolism (*HIF1A*, *ARNT*), and include a promi-
279 nent downstream target of *KLF6* in the regulation of cholesterol metabolism: *BHLHE40* (Syafuruddin
280 et al., 2019). Regulatory islands affected by differential *BHLHE40* binding include target genes such
281 as the aforementioned *GLI3* as well as *ITGB8*, whose role in PI3K-AKT-mTOR signaling in (outer)
282 radial glia in humans has been highlighted in two independent studies (Mora-Berm dez, Badsha,
283 et al., 2016; Pollen, Bhaduri, et al., 2019). Another transcription factor controlling cholesterol home-
284 ostasis, *SREBF2*, exhibits differential binding affinity for a regulatory island linked to *PALMD*, which
285 plays a specific role in basal progenitor proliferation (Kalebic, Gilardi, et al., 2019) and is among
286 the very few genes that have accumulated derived mutations in our lineage but none in the Nean-
287 derthal/Denisovan genomes (Kuhlwilm and Boeckx, 2019).

288 Our analysis reveals differential binding affinity sites for *KLF6*, including reduced affinity affect-

289 ing a regulatory island linked to *SHROOM3*, a well-studied marker of apical/ventral progenitors. Our
290 analysis also predicts a KLF6-associated regulatory variant altering a *GLI3* TFBS (chr10:3978704-G-C,
291 hg19 genome version), with higher affinity in *Homo sapiens* when compared to the ancestral variant
292 found in Neanderthal/Denisovan genomes (Figure 4C; ST7). While it is not surprising to find this
293 mutual regulation of cholesterol and *sonic hedgehog* signaling (Blassberg and Jacob, 2017), even in
294 the context of basal progenitors (L. Wang, Hou, and Han, 2016), we find this differential binding
295 affinity by *GLI3* particularly intriguing in the context of the present study. *GLI3* is a critical regulator
296 of the dorsoventral cell fate specification and the switch between proliferative and differentiative
297 radial glia divisions (in different model systems (Hasenpusch-Theil et al., 2018; Fleck et al., 2022)).

298 We found *GLI3* as one of the genes whose expression trajectory significantly changes through
299 pseudotime, and our piNMF analysis places *GLI3* prominently at the juncture between early and
300 late radial glia modules (program 2). Consistent with this, among GO terms marking the beginning
301 of the late piNMF modules (oRG states) we find "hedgehog offstate" (ST4). In addition, the regu-
302 latory islands linked to *GLI3* and associated with positive selection already mentioned above are
303 associated with increased binding affinity for genes like *ARID3A* and *LHX2* (and decreased affinity for
304 *NKX2-1*, a ventral forebrain marker) (Figure 4C; ST7). Both *ARID3A* and *LHX2* are known to modulate
305 the cell cycle and the tempo of cortical neurogenesis in a β -catenin-dependent manner (Saadat,
306 2013; Hsu et al., 2015). Both TFs have been linked to the regulation mTOR pathway, and this is also
307 the case for *GLI3* too (e.g., loss of *GLI3* is reported to activate mTORC1 signaling in muscle satellite
308 cells (Brun et al., 2022), consistent with the transition between early and late RG programs in our
309 piNMF analysis).

310 It is noteworthy that the *GLI3* variants within regulatory islands under putative positive se-
311 lection have ClinVar-associated phenotypes (Landrum et al., 2018), with the minor (ancestral) al-
312 lele linked to Greig cephalopolysyndactyly syndrome (OMIM:175700) and Pallister-Hall syndrome
313 (OMIM:146510), which affect brain size and craniofacial traits among other clinical features. Val-
314 idating the impact of these changes in an experimental setting is an important research direc-
315 tion emerging from this analysis. We observe in this context that within the *KLF6* transcriptional
316 networks in our analysis one finds prominent *GLI3* targets relevant for the specification of dorsal
317 telencephalic progenitors (Fleck et al., 2022), such as: *HES1*, *HES4* or *HES5*, as well as *CTNBN1*. In ad-
318 dition, experimental perturbation of *GSK3 β* , a kinase that integrates multiple signaling pathways
319 (including hedgehog and WNT- β -catenin signaling in mice neural progenitors (Kim et al., 2009)),
320 specifically affects cholesterol metabolism and indeed *KLF6* expression coincident with the emer-
321 gence of the oRG lineage in human cortical organoids (López-Tobón et al., 2019).

322 Discussion

323 Previous large-scale single-cell studies have extensively characterized neural cells from the devel-
324 oping human brain. However, the molecular definition of the lineage tree relating apical progeni-
325 tors to basal progenitor populations, as part of an intricate web of complex lineage relationships,
326 has remained elusive. By implementing an integrative computational framework for the joint in-
327 vestigation of different biological layers of the cell using high-throughput single-cell data, we char-
328 acterized gene expression programs sequentially activated during progenitor cell progression and
329 identified key transcriptional regulators shedding light onto central processes of neural progenitor
330 cell fate dynamics, and evolutionary modifications thereof.

331 Our findings uncover *KLF6* transcription factor as a central node in human radial glia transcrip-
332 tional networks. *KLF6* is a member of the zinc finger-containing family of transcription factors
333 resembling *Drosophila* protein Krüppel (Dang, Pevsner, and V. W. Yang, 2000) and whose role in
334 human neurogenesis has to date remained largely undescribed. *KLF6* has been associated to a
335 super-interactive promoter specifically in radial glia (Song et al., 2020) and its targets during neo-
336 cortical development were previously reported to be enriched in oRG (Polioudakis et al., 2019),
337 consistent with our findings based on GRN reconstruction and piNMF. We identified several en-

zymes implicated in cholesterol biosynthesis under the KLF6 transcriptional control, prominently during the acquisition of outer radial glia identity. Previous studies in other model systems have also reported similar gene expression programs regulated by KLF6 related to lipid homeostasis (Syafuruddin et al., 2019; Z. Wang et al., 2018). Future work is required to elucidate the roles of cholesterol metabolism in outer radial glia proliferation and neurogenesis, particularly in light of clinical association of *KLF6* to glioblastoma (Masilamani et al., 2017), where sustained cholesterol synthesis impacts tumor cell growth (Kambach et al., 2017).

The metabolic control of neural progenitor cell behavior significantly contributes to species-specific features of brain evolution (Namba et al., 2021; Iwata et al., 2023), and experimental evidence already points to significant changes impacting various metabolic pathways in our recent evolution (after the split from our closest extinct relatives) (Stepanova et al., 2021; Pinson, Xing, et al., 2022). Our evolutionary-informed analysis of transcription factor binding site disruptions contributes to this emerging picture by highlighting modifications clustering around cholesterol metabolism. In addition, our study highlights the relevance of mutations affecting *GLI3*. Not only did we infer a differential regulation of *KLF6* by *GLI3*, we also uncovered regulatory islands associated with signals of positive selection predicted to impact *GLI3* expression during cortical development. We find it noteworthy that some of the variants defining the regulatory island around *GLI3* are among the most recent derived high-frequency *GLI3* changes in our lineage (Kuhlwilm and Boeckx, 2019), and are predicted to have emerged between 200 and 300kya (Andirkó et al., 2022), a significant period in our recent evolutionary history (Hublin, Ben-Ncer, et al., 2017; Schlebusch et al., 2017; Skoglund et al., 2017). Also, in light of our findings, future research may explore further the promising interplay between the primary cilia and *GLI3* activity in the regulation of cell cycle length and cortical size (Wilson et al., 2012), considering as well the evolutionary relevant role of mTOR signaling in ciliary dynamics, impacting particularly basal progenitors (Heurck et al., 2023), and between cholesterol accessibility and the regulation of hedgehog signaling in the membrane of the primary cilium (Kinnebrew et al., 2019).

Our approach illustrates the relevance of paleogenomes in adding temporal precision to important differences that comparisons between humans and other great apes already revealed (Pollen, Kilik, et al., 2023), in particular here the role of mTOR signaling in human cortical development (Pollen, Bhaduri, et al., 2019). At a more general level, our work adds to the mounting evidence for the importance of regulatory regions in modifying developmental programs in the course of (recent) human evolution (Peyr gne et al., 2017; Moriano and Boeckx, 2020; Weiss et al., 2021; Gokhman et al., 2020; Mangan et al., 2022; Keough et al., 2023; Kaplow et al., 2023).

Our work also shows how paleogenomics offers the potential to probe questions about brain evolution that go beyond traits that may be recoverable from the (traditional) fossil record, such as overall adult brain size or shape. Our evolution-oriented analysis invites the hypothesis that important modifications impacting upper-layers of the neocortex took place relatively recently in our history. The evidence presented here involving differential regulation of cholesterol signaling in outer radial glia, together with independent evidence concerning changes affecting genes specifically involved in basal progenitor proliferation (such as *PALMD* (Kuhlwilm and Boeckx, 2019; Kalebic, Gilardi, et al., 2019) or *TKTL1* (Pinson, Xing, et al., 2022)), as well as upper-layer neuron markers like *SATB2* (Weiss et al., 2021), points to the need to probe the nature of associative, cortico-cortical connections characteristic of upper-layer neuronal ensembles further.

Methods & Materials

Single-cell RNA-seq data processing

Raw single cell RNA-seq datasets from selected studies were processed using Seurat 4.2.0, guided by best practices of single cell analysis (Luecken and Theis, 2019). Seurat objects were created from raw count matrices and retention of high quality cells was based on the following cell attributes: total counts, expressed genes, percentage of mitochondrial gene counts and percentage of zero

387 counts, requiring a distribution of values within three median absolute deviations for each attribute
388 and per batch. Actively dividing cells were filtered out based on *TOP2A* expression. To jointly ana-
389 lyze samples from different batches, as well as data from both Trevino et al., 2021 and Polioudakis
390 et al., 2019, in a shared low dimensional space, we performed normalization and integration using
391 Seurat dedicated functions *SCTransform* and *IntegrateData* before computing principal component
392 analysis. A common processing was implemented for inferring the branch trajectories and for
393 gene regulatory network reconstruction (see below): retaining genes with expression in at least 50
394 cells, normalization of cell counts to equal median of counts per cell before normalization, selec-
395 tion of 4000 highly variable genes based on Seurat variance-stabilizing transformation algorithm
396 (Hafemeister and Satija, 2019), followed by re-normalization and log-transformation. Coarse clus-
397 tering was performed using Leiden algorithm and resolution parameter to 0.1. Logistic regression
398 was used to identify differentially expressed genes. Cell cluster annotation was based on both
399 differential expression analysis and available annotations from the original publications.

400 Complementarily, we performed single-cell trajectory reconstruction using python package *sc-*
401 *Fates* (Faure et al., 2023) on normalized, log-transformed count matrices. A force-directed graph
402 was drawn using our previously computed PCA coordinates for initialization. Then we used the
403 Palantir software (Setty et al., 2019) included in the *scFates* toolkit to generate a diffusion space for
404 tree learning using the ElPiGraph algorithm. Pseudotime was calculated using FOS gene expres-
405 sion for root selection and the genes that significantly change in expression along the inferred tree
406 were identified using the *scFates* cubic spline regression function.

407 **Gene regulatory network inference and analysis**

408 Gene regulatory network reconstruction was performed following the computational framework of
409 *CellOracle* software, combining single-cell ATAC-seq and RNA-seq data modalities for transcription
410 factor-target genes inference.

411 In order to build an atlas of open chromatin regions active during human cortical development,
412 we selected as reference the singleome ATAC-seq dataset from Trevino et al., 2021, containing the
413 highest number of ATAC-seq peaks, and required a minimum of 50% overlap with open chromatin
414 signals from either one of the following datasets: multiome ATAC-seq data from Trevino et al., 2021,
415 ATAC-seq datasets from Markenscoff-Papadimitriou et al., 2020 and Torre-Ubieta et al., 2018. As
416 the reference dataset does not contain signals for the X and Y chromosomes, we included these
417 data as available in Markenscoff-Papadimitriou et al., 2020 and Torre-Ubieta et al., 2018. At total of
418 392961 regulatory regions (hg38 genome version) were used for downstream analyses. We then
419 built regulatory region-gene associations based on genomic proximity and literature curated regu-
420 latory domains (McLean et al., 2010). Next, we scanned each regulatory region for transcription fac-
421 tors motifs using the Hocomoco database version 11 (Kulakovskiy et al., 2018). The resulting tran-
422 scription factor-regulatory region-gene associations represent the raw gene regulatory network
423 for the machine learning-based regression analysis to impute cluster-specific GRNs (Kamimoto et
424 al., 2023). Of the two algorithms available in the *CellOracle* software, we chose the bagging ridge
425 regression model, as it consistently reported better scores for network degree distribution (Figure
426 S1). Cluster-specific TF-target gene interactions were obtained by filtering by a p-value threshold
427 of 0.001 for connection strength and a maximum of 2000 links per cluster. An evaluation of such
428 GRNs was performed on the basis of the centrality measures, including betweenness centrality
429 and eigenvector centrality (as proposed in Kamimoto et al., 2023).

430 **Pseudotime-informed non-negative matrix factorization**

431 We implemented a matrix factorization analysis to learn the dynamics of gene expression pro-
432 grams dependent on pseudotime from single-cell data. Non-negative matrix factorization consists
433 in the decomposition of a matrix of n vectors with non-negative values into two lower rank, non-
434 negative matrices: the pattern matrix containing basis vectors and the coefficient matrix with the
435 coefficients of the non-negative linear combination of the basis vectors, aiming to minimize:

$$d(Y, AX) \tag{1}$$

436 where d is the distance (by a given measure) between the original matrix and the reconstruc-
437 tion AX . As our inquiry deals with cellular differentiation events, we sought to decompose a high-
438 dimensional single cell dataset accounting for the dynamic nature of gene expression trajectories
439 through pseudotime. As the core algorithm, we computed the matrix factorization following the
440 original work of Hautecoeur and Glineur, 2020, where the approximation is now:

$$y_i(t) \approx \sum_j^r a_j(t)x_{ji} \tag{2}$$

441 where each vector of y is a function dependent on time t , a contains a set of r non-negative
442 functions, and x contains the non-negative coefficient values, for a given factorization rank r and
443 $1 \leq j \leq r, 1 \leq i \leq n$. As with other factorization methods, there is no a priori knowledge of the
444 factorization rank (i.e. expected number of patterns in the data), and thus k must be provided by
445 the user; measures of stability and error (see below) can guide this selection. Here we chose four
446 expression programs as a neat balance between stability across branches and datasets and reso-
447 lution of semi-discrete modules along pseudotime (see Figure 2 and S2). We used degree 3 splines
448 as the set of functions to model gene expression trajectories, selecting the number of knots (ob-
449 taining intervals where to fit trajectories) to 4 (a low number avoids overfitting and better captures
450 global trends). The algorithm to solve the factorization problem is based on Hierarchical Alternat-
451 ing Least Squares (implemented in Hautecoeur and Glineur, 2020), and a maximum number of
452 iterations of 10^4 and tolerance 10^{-10} were set as stopping criteria.

453 Given that NMF is a matrix approximation method, we followed the iterative and clustering
454 strategies presented in Kotliar et al., 2019 as an extended algorithm to recover stable gene expres-
455 sion modules. Matrix decompositions from the core algorithm presented above were computed
456 over 750 iterations per factorization rank to obtain replicates that were then clustered via KMeans
457 clustering based on Euclidean distance to obtain consensus values for the pattern and coefficient
458 matrices. Measures of stability and error of the matrix reconstruction were calculated using silhou-
459 ette scores and the Frobenius norm of approximation, respectively, following Kotliar et al., 2019.
460 Additionally, in order to statistically associate genes to gene expression programs, marker genes
461 for each module were identified using the normalized z-score gene expression value of each gene
462 for multiple least squares regression against the programs in the pattern matrix, as implemented
463 in Kotliar et al., 2019.

464 **Paleogenomic analysis**

465 We made use of a paleogenomic dataset (Kuhlwilm and Boeckx, 2019) that catalogs segregating
466 sites between *Homo sapiens* and high quality genomes from two Neanderthals and one Denisovan
467 individuals (Meyer et al., 2012; Prüfer, Racimo, et al., 2014; Prüfer, Filippo, et al., 2017), where an-
468 cestralty was inferred from publicly available multiple genome alignments (Paten et al., 2008) or,
469 when this information was not available, from the macaque reference genome (Yan et al., 2011).
470 Allele frequency was determined from the dbSNP database build 147 (Sherry et al., 2001) and a 90%
471 allele frequency threshold was set to retain high-frequency variants for further analyses (Kuhlwilm
472 and Boeckx, 2019). In the search for regulatory regions that might have been under selection in
473 recent *Homo sapiens* evolution and differentially impact gene expression, we intersected the reg-
474 ulatory regions from our open chromatin region brain atlas with *Homo sapiens*-derived variants
475 where the Neanderthals/Denisovans carry the ancestral allele (using the bedtools suite (Quinlan
476 and Hall, 2010)); additionally, to identify genomic regions that may encapsulate *Homo*-specific reg-
477 ulatory mechanisms, we required for each variant to be contained within a genomic window of
478 at least 3000bp where the Neanderthal/Denisovan homolog regions did not accumulated lineage-
479 specific derived changes. A total of $n=4836$ "regulatory islands" were identified and associated to

480 4797 genes. To evaluate disruptions of transcription factor binding sites, we generated a set of
481 genomic coordinates of variants sitting within regulatory islands using a unique identifier based
482 on genomic coordinates and allele information. Differences in transcription factor binding affinity
483 were computed applying the information content method from the motifbreakR package (S. G.
484 Coetzee, G. A. Coetzee, and Hazelett, 2015) and using position weighted matrices annotated in
485 the Hocomoco motif collection (Kulakovskiy et al., 2018) (consistent with our GRN reconstruction
486 analysis). A significance threshold was set to $1e-4$ and an even background nucleotide distribution
487 was assumed. Redundant motifs were dropped and the resulting TF-variant associations further
488 filtered by retaining only those with a predicted affinity difference falling in the 4th quantile of the
489 distribution. Finally, an enrichment score was computed for each TF based on the number of
490 strong and total hits identified. GO enrichment analyses were performed on the TF identified as
491 described above (using the same Hocomoco motif collection as custom reference set). Analyses
492 were performed with the TopGO package (Alexa, Rahnenfuhrer, et al., 2010) using the following
493 parameters: 'weight01' as algorithm, 'Fisher' as statistics, 8 as 'nodeSize' and 3 as 'minTerms'; a
494 p-value < 0.05 and an enrichment > 1 were set as thresholds to select significant GO terms.

495 **Limitations of this study**

496 The (pseudo)temporal ordering of gene expression states from single-cell data presented here al-
497 lows us to interpret cell differentiation as a molecular continuum, but it remains to be seen how
498 closely this recapitulates the transcriptional dynamics of lineage progression *in vivo*. Additionally,
499 the process of indirect neurogenesis studied here idealizes away from what is a much more com-
500 plex network of lineage relationships among neural progenitor subtypes. The reconstruction and
501 recovery of regulatory networks and expression programs rely on the identification of a set of
502 transcription factors and highly variable genes that only partially represent the higher complexity
503 of the cells. This complexity is even more manifest when the temporal differences among neural
504 progenitors during the long human gestational period is taken into account. Lastly, future experi-
505 mental work is required to validate the predictions derived from the paleogenomic interrogation
506 of regulatory variants presented here.

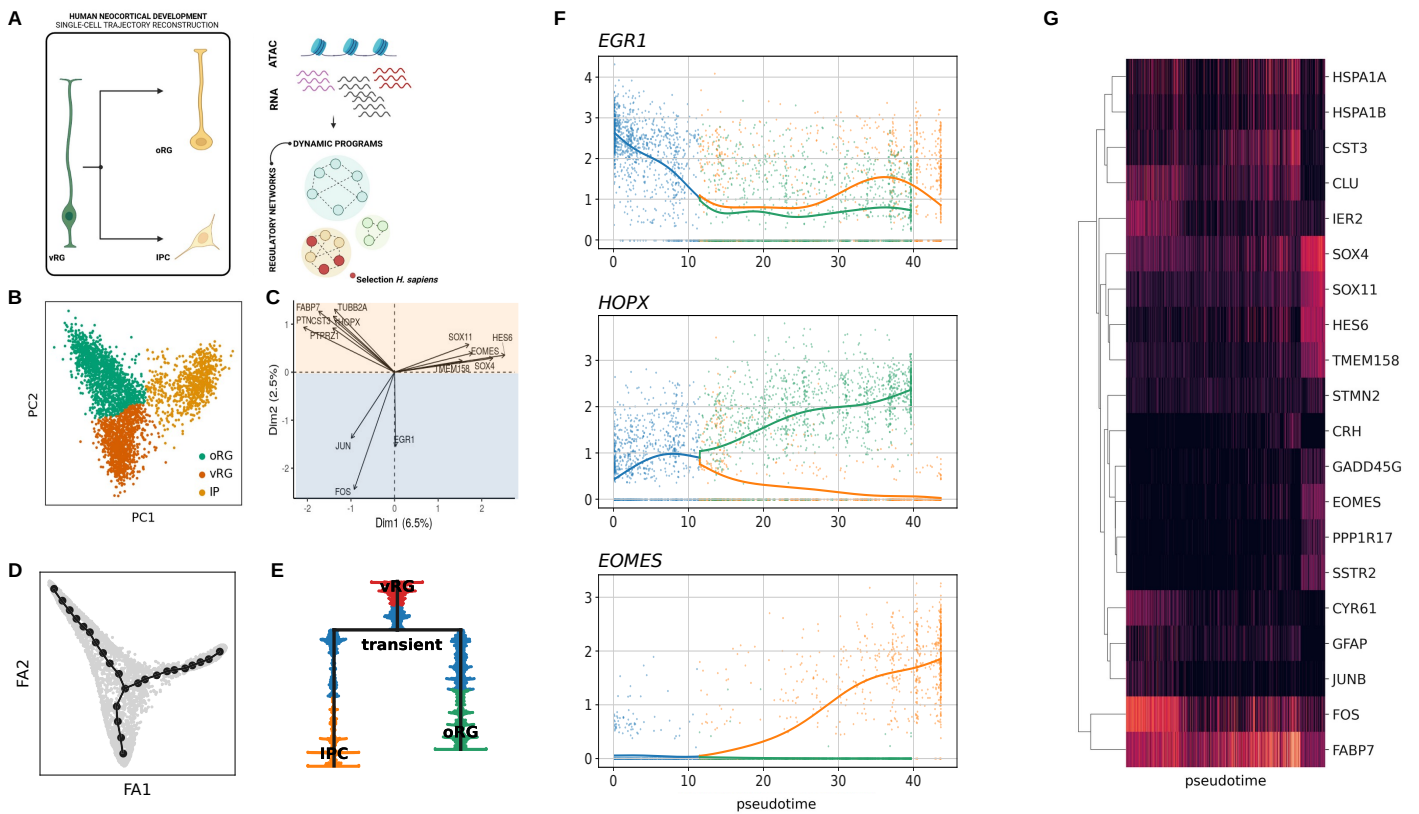



Figure 1. Resolving the tree of neural progenitor cell differentiation during human corticogenesis. A) Schematic of analyses implemented in this paper: single-cell trajectory reconstruction of basal progenitor generation, for the inference and recovery of gene regulatory networks and expression programs, illuminated by paleogenomic analysis. This subfigure was created using [BioRender.com](https://www.biorender.com). B) and C) Identifying the main axis of variation using principal component analysis is a powerful strategy to characterize the heterogeneity and transcriptional dynamics of progenitor cells (as shown for instance in a comprehensive study in mice (Mukhtar et al., 2022)). Here, we performed PCA on a single-cell dataset of human neural progenitors, which allowed the discrimination of radial glia and intermediate progenitor cell subtypes (coarse clustering, B). Top gene loadings with known markers of neural progenitor subtypes are shown in C). D) and E) Inferred tree of principal points and associated dendrogram capturing the hierarchy of neural cell lineage relationships as inferred from single-cell data. F) Expression trajectory along pseudotime of three marker genes for ventricular radial, outer radial glia and intermediate progenitor cell clusters. G) Heatmap with representative genes whose trajectories significantly change as pseudotime progresses.

Figure 1—figure supplement 1. Differential expression and complementary analysis on an independent dataset.

507 Acknowledgment

508 We are grateful to Cécile Hautecoeur for providing help and insights into the non-negative matrix
509 factorization methods. The format of this preprint is based on the LaPreprint template ([https://
510 github.com/roaldarbol/lapreprint](https://github.com/roaldarbol/lapreprint)) by Mikkel Roald-Arbøl .

511 Author contributions

512 Conceptualization: J.M. and C.B.; Methodology: J.M., O.L. and A.V.; Data curation: J.M., O.L. and
513 C.B.; Visualization: J.M., O.L. and C.B.; Investigation: J.M., O.L. and C.B.; Writing—original draft: J.M,
514 O.L. and C.B.; Writing—review and editing: J.M., O.L., A.V., G.T. and C.B.; Supervision: G.T. and C.B.;
515 Funding acquisition: G.T. and C.B.

516 Figures

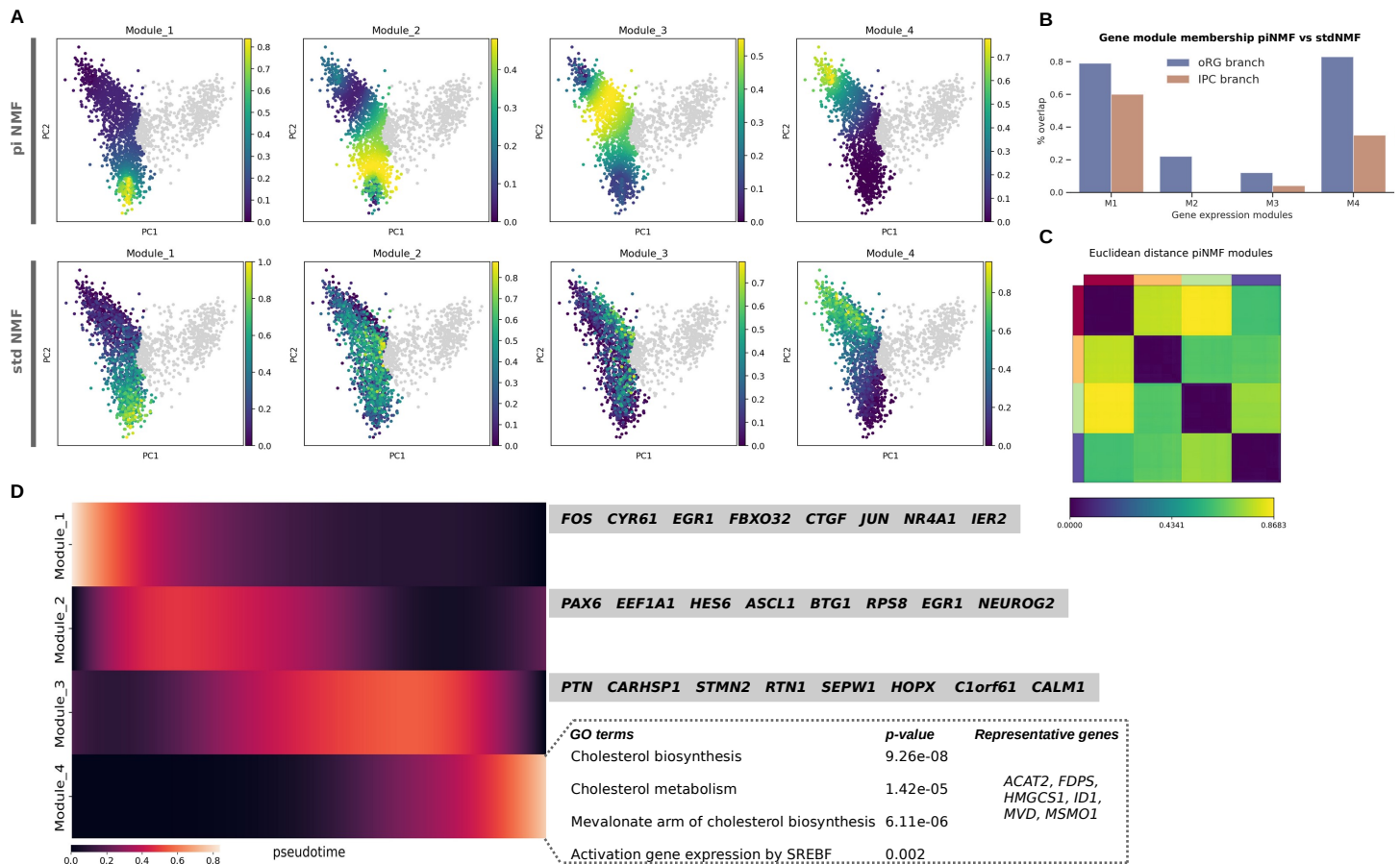


Figure 2. Pseudotime-informed non-negative matrix factorization recovers a sequential activation of gene expression programs. A) Comparatively in PCA plots, piNMF is able to resolve expression programs transiently activated for the lineage branch leading to oRG cluster (same for the IPC branch, see Figure S2), while stdNMF does not recover such clear patterns from the data. B) Genes assigned to modules at the extreme of the lineage tree (vRG and either oRG or IPC) are shared in higher percentage when compared to modules 2 and 3, confirming main differences among NMFs algorithms pertain to the transient activation of expression programs along the tree. C) The high values on the euclidean distance among the four gene expression programs supports, along with the stability and error measures (see Figure S2), the factorization rank selection. D) Heatmap depicting the sequential activation of expression programs in the radial glia branch, with marker genes for each module and, for module 4, representative GO terms highlighted in the main text.

Figure 2—figure supplement 1. Comparison of GO terms captured by NMF methods for transiently activated modules

Figure 2—figure supplement 2. Non-negative matrix factorizations on the IPC branch and factorization rank selection.

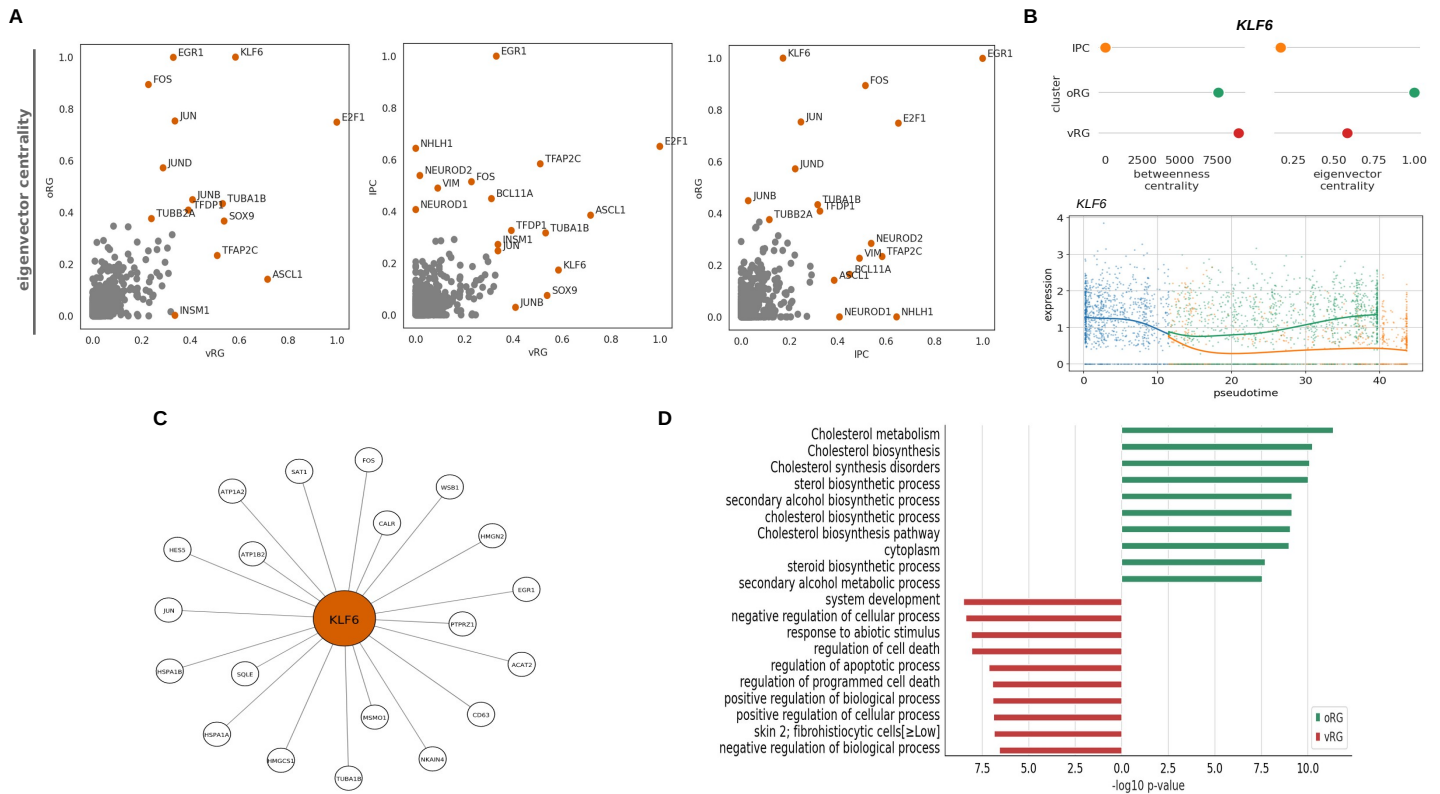


Figure 3. Gene regulatory network reconstruction from human neural progenitor single-cell data A) Pairwise comparisons of eigenvector centrality values among single-cell progenitor cell clusters, highlighting top 10 genes in each cluster. Some differentially expressed genes for vRG cell cluster retain some level of expression in basal progenitors and are indeed present among top 10 genes for different GRN connectivity measures across clusters (see also Figure S2); this is the case of *EGR1*, *FOS* or *JUN*. In addition to *KLF6* for oRG, others genes that are more prominently associated to specific clusters include *ASCL1*, *SOX9*, *TFAP2C* for vRG when compared to oRG, or neuron differentiation-related basic helix-loop-helix transcription factors *NEUROD1*, *NEUROD2* and *NHLH1* between IPC and RG clusters, consistent also with the closer transcriptomic similarity of IP cells to excitatory neurons (Bhaduri et al., 2021). B) *KLF6* networks measures across single-cell clusters, with a marked contrast between IPC cluster and RG clusters, and most prominently as central node in outer radial glia (eigenvector centrality). Below, *KLF6* expression along pseudotime, showing upregulation in oRG and downregulation in IPC. C) Top representative genes by network weight among *KLF6* target genes. D) Top GO terms associated to *KLF6* targets in outer radial glia and ventricular radial glia, with prominence of cholesterol metabolism in outer radial glia. Cholesterol metabolism GO terms only appears for vRG cluster *KLF6* targets if lowering the p-value threshold above 0.01 (see also ST3).

Figure 3—figure supplement 1. Evaluation of gene regulatory networks across algorithms and datasets.

Figure 3—figure supplement 2. Networks measures (eigenvector centrality and betweenness centrality) for two independent datasets.

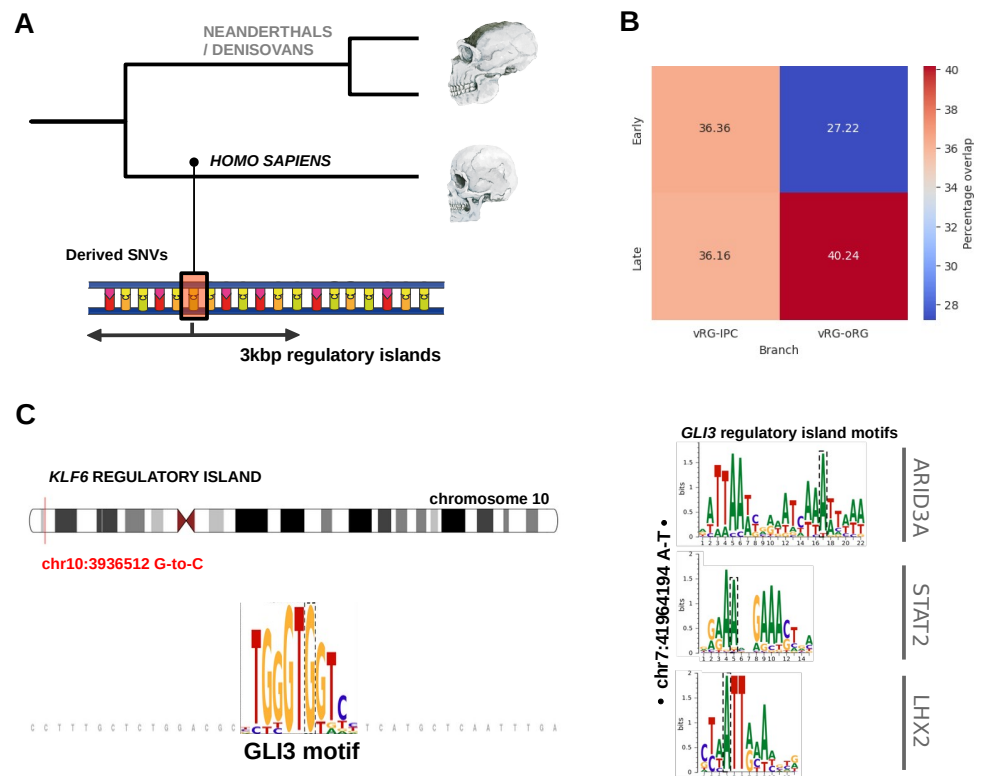


Figure 4. Paleogenomic analysis of regulatory variants A) Building on the brain atlas of open chromatin regions, regulatory islands were defined as 3kbp-length regions where *Homo sapiens* acquired derived alleles in comparison to Neanderthals and Denisovans (carrying the ancestral version found in chimpanzees). Human skulls are modified images from (Theofanopoulou et al., 2017). B) Genes associated to the regulatory islands are found in piNMF modules detected for both vRG to IPC and to oRG branches, with more pronounced abundance on the oRG lineage. C) Predicted TF differential regulation of *KLF6* by *GLI3* (left) and cluster of motifs within a *GLI3* regulatory island under positive selection (right), affected by *Homo sapiens*-derived single nucleotide variants.

References

- 517
518 Aibar, Sara et al. (2017). "SCENIC: single-cell regulatory network inference and clustering". In: *Nature*
519 *Methods* 14.11, pp. 1083–1086. DOI: [10.1038/nmeth.4463](https://doi.org/10.1038/nmeth.4463).
- 520 Alexa, Adrian, Jorg Rahnenfuhrer, et al. (2010). "topGO: enrichment analysis for gene ontology". In:
521 *R package version 2.0*, p. 2010.
- 522 Andirkó, Alejandro et al. (2022). "Temporal mapping of derived high-frequency gene variants sup-
523 ports the mosaic nature of the evolution of Homo sapiens". In: *Scientific Reports* 12.1, p. 9937.
524 DOI: [10.1038/s41598-022-13589-0](https://doi.org/10.1038/s41598-022-13589-0).
- 525 Bedogni, Francesco and Robert F. Hevner (2021). "Cell-Type-Specific Gene Expression in Developing
526 Mouse Neocortex: Intermediate Progenitors Implicated in Axon Development". In: *Frontiers in*
527 *Molecular Neuroscience* 14. DOI: [10.3389/fnmol.2021.686034](https://doi.org/10.3389/fnmol.2021.686034).
- 528 Bhaduri, Aparna et al. (2021). "An atlas of cortical arealization identifies dynamic molecular signa-
529 tures". In: *Nature* 598.7879, pp. 200–204. DOI: [10.1038/s41586-021-03910-8](https://doi.org/10.1038/s41586-021-03910-8).
- 530 Blassberg, Robert and John Jacob (2017). "Lipid metabolism fattens up hedgehog signaling". In: *BMC*
531 *Biology* 15, p. 95. DOI: [10.1186/s12915-017-0442-y](https://doi.org/10.1186/s12915-017-0442-y).
- 532 Brun, Caroline E. et al. (2022). "Gli3 regulates muscle stem cell entry into GAlert and self-renewal".
533 In: *Nature Communications* 13.1. Number: 1 Publisher: Nature Publishing Group, p. 3961. DOI:
534 [10.1038/s41467-022-31695-5](https://doi.org/10.1038/s41467-022-31695-5).
- 535 Butler, Andrew et al. (2018). "Integrating single-cell transcriptomic data across different conditions,
536 technologies, and species". In: *Nature Biotechnology* 36.5, pp. 411–420. DOI: [10.1038/nbt.4096](https://doi.org/10.1038/nbt.4096).
- 537 Chen, Lu et al. (2020). "Identifying and Interpreting Apparent Neanderthal Ancestry in African Indi-
538 viduals". In: *Cell* 180.4, 677–687.e16. DOI: [10.1016/j.cell.2020.01.012](https://doi.org/10.1016/j.cell.2020.01.012).
- 539 Coetzee, Simon G., Gerhard A. Coetzee, and Dennis J. Hazelett (2015). "motifbreakR: an R/Biocon-
540 ductor package for predicting variant effects at transcription factor binding sites". In: *Bioinfor-*
541 *matics* 31.23, pp. 3847–3849. DOI: [10.1093/bioinformatics/btv470](https://doi.org/10.1093/bioinformatics/btv470).
- 542 Corbeil, Denis et al. (2010). "The intriguing links between prominin-1 (CD133), cholesterol-based
543 membrane microdomains, remodeling of apical plasma membrane protrusions, extracellular
544 membrane particles, and (neuro)epithelial cell differentiation". In: *FEBS Letters* 584.9, pp. 1659–
545 1664. DOI: [10.1016/j.febslet.2010.01.050](https://doi.org/10.1016/j.febslet.2010.01.050).
- 546 Dang, Duyen T., Jonathan Pevsner, and Vincent W. Yang (2000). "The biology of the mammalian
547 Krüppel-like family of transcription factors". In: *The international journal of biochemistry & cell*
548 *biology* 32.11-12, pp. 1103–1121.
- 549 Faure, Louis et al. (2023). "scFates: a scalable python package for advanced pseudotime and bifurca-
550 tion analysis from single-cell data". In: *Bioinformatics* 39.1, btac746. DOI: [10.1093/bioinformatics/btac746](https://doi.org/10.1093/bioinformatics/btac746).
- 551
552 Fietz, Simone A. et al. (2012). "Transcriptomes of germinal zones of human and mouse fetal neo-
553 cortex suggest a role of extracellular matrix in progenitor self-renewal". In: *Proceedings of the*
554 *National Academy of Sciences of the United States of America* 109.29, pp. 11836–11841. DOI: [10.1073/pnas.1209647109](https://doi.org/10.1073/pnas.1209647109).
- 555
556 Fleck, Jonas Simon et al. (2022). "Inferring and perturbing cell fate regulomes in human brain
557 organoids". In: *Nature*, pp. 1–8. DOI: [10.1038/s41586-022-05279-8](https://doi.org/10.1038/s41586-022-05279-8).
- 558 Gokhman, David et al. (2020). "Differential DNA methylation of vocal and facial anatomy genes in
559 modern humans". In: *Nature Communications* 11.1, p. 1189. DOI: [10.1038/s41467-020-15020-6](https://doi.org/10.1038/s41467-020-15020-6).
- 560 Hafemeister, Christoph and Rahul Satija (2019). "Normalization and variance stabilization of single-
561 cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1,
562 p. 296. DOI: [10.1186/s13059-019-1874-1](https://doi.org/10.1186/s13059-019-1874-1).
- 563 Hasenpusch-Theil, Kerstin et al. (2018). "Gli3 controls the onset of cortical neurogenesis by regu-
564 lating the radial glial cell cycle through Cdk6 expression". In: *Development (Cambridge, England)*
565 145.17, dev163147. DOI: [10.1242/dev.163147](https://doi.org/10.1242/dev.163147).

- 566 Hautecoeur, Cécile and François Glineur (2020). "Nonnegative Matrix Factorization over Contin-
567 uous Signals using Parametrizable Functions". In: *Neurocomputing* 416, pp. 256–265. DOI: [10.1016/j.neucom.2019.11.109](https://doi.org/10.1016/j.neucom.2019.11.109).
- 568
- 569 Heurck, Roxane Van et al. (2023). "CROCCP2 acts as a human-specific modifier of cilia dynamics
570 and mTOR signaling to promote expansion of cortical progenitors". In: *Neuron* 111.1. Publisher:
571 Elsevier, 65–80.e6. DOI: [10.1016/j.neuron.2022.10.018](https://doi.org/10.1016/j.neuron.2022.10.018).
- 572 Hsu, Lea Chia-Ling et al. (2015). "Lhx2 regulates the timing of β -catenin-dependent cortical neuro-
573 genesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.39,
574 pp. 12199–12204. DOI: [10.1073/pnas.1507145112](https://doi.org/10.1073/pnas.1507145112).
- 575 Hublin, Jean-Jacques, Abdelouahed Ben-Ncer, et al. (2017). "New fossils from Jebel Irhoud, Morocco
576 and the pan-African origin of Homo sapiens". In: *Nature* 546.7657. Number: 7657 Publisher:
577 Nature Publishing Group, pp. 289–292. DOI: [10.1038/nature22336](https://doi.org/10.1038/nature22336).
- 578 Hublin, Jean-Jacques, Simon Neubauer, and Philipp Gunz (2015). "Brain ontogeny and life history
579 in Pleistocene hominins". In: *Philosophical Transactions of the Royal Society B: Biological Sciences*
580 370.1663. Publisher: Royal Society, p. 20140062. DOI: [10.1098/rstb.2014.0062](https://doi.org/10.1098/rstb.2014.0062).
- 581 Iwata, Ryohei et al. (2023). "Mitochondria metabolism sets the species-specific tempo of neuronal
582 development". In: *Science* 379.6632, eabn4705. DOI: [10.1126/science.abn4705](https://doi.org/10.1126/science.abn4705).
- 583 Kalebic, Nereo, Carlotta Gilardi, et al. (2019). "Neocortical Expansion Due to Increased Proliferation
584 of Basal Progenitors Is Linked to Changes in Their Morphology". In: *Cell Stem Cell* 24.4, 535–
585 550.e9. DOI: [10.1016/j.stem.2019.02.017](https://doi.org/10.1016/j.stem.2019.02.017).
- 586 Kalebic, Nereo and Wieland B. Huttner (2020). "Basal Progenitor Morphology and Neocortex Evo-
587 lution". In: *Trends in Neurosciences* 43.11. Publisher: Elsevier, pp. 843–853. DOI: [10.1016/j.tins.2020.07.009](https://doi.org/10.1016/j.tins.2020.07.009).
- 588
- 589 Kambach, Diane M. et al. (2017). "Disabled cell density sensing leads to dysregulated cholesterol
590 synthesis in glioblastoma". In: *Oncotarget* 8.9, pp. 14860–14875. DOI: [10.18632/oncotarget.14740](https://doi.org/10.18632/oncotarget.14740).
- 591 Kamimoto, Kenji et al. (2023). "Dissecting cell identity via network inference and in silico gene per-
592 turbation". In: *Nature* 614.7949, pp. 742–751. DOI: [10.1038/s41586-022-05688-9](https://doi.org/10.1038/s41586-022-05688-9).
- 593 Kaplow, Irene M. et al. (2023). "Relating enhancer genetic variation across mammals to complex
594 phenotypes using machine learning". In: *Science* 380.6643. Publisher: American Association for
595 the Advancement of Science, eabm7993. DOI: [10.1126/science.abm7993](https://doi.org/10.1126/science.abm7993).
- 596 Kawaguchi, Ayano et al. (2008). "Single-cell gene profiling defines differential progenitor subclasses
597 in mammalian neurogenesis". In: *Development* 135.18, pp. 3113–3124. DOI: [10.1242/dev.022616](https://doi.org/10.1242/dev.022616).
- 598 Keough, Kathleen C. et al. (2023). "Three-dimensional genome rewiring in loci with human acceler-
599 ated regions". In: *Science (New York, N.Y.)* 380.6643, eabm1696. DOI: [10.1126/science.abm1696](https://doi.org/10.1126/science.abm1696).
- 600 Kim, Woo-Yang et al. (2009). "GSK-3 is a master regulator of neural progenitor homeostasis". In:
601 *Nature Neuroscience* 12.11, pp. 1390–1397. DOI: [10.1038/nn.2408](https://doi.org/10.1038/nn.2408).
- 602 Kinnebrew, Maia et al. (2019). "Cholesterol accessibility at the ciliary membrane controls hedgehog
603 signaling". In: *eLife* 8. Ed. by Duoia Pan, Marianne E Bronner, and Stacey K Ogden, e50051. DOI:
604 [10.7554/eLife.50051](https://doi.org/10.7554/eLife.50051).
- 605 Kotliar, Dylan et al. (2019). "Identifying gene expression programs of cell-type identity and cellular
606 activity with single-cell RNA-Seq". In: *eLife* 8. Ed. by Alfonso Valencia et al., e43803. DOI: [10.7554/eLife.43803](https://doi.org/10.7554/eLife.43803).
- 607
- 608 Kriegstein, Arnold, Stephen Noctor, and Verónica Martínez-Cerdeño (2006). "Patterns of neural
609 stem and progenitor cell division may underlie evolutionary cortical expansion". In: *Nature Re-
610 views Neuroscience* 7.11, pp. 883–890. DOI: [10.1038/nrn2008](https://doi.org/10.1038/nrn2008).
- 611 Kuhlwil, Martin and Cedric Boeckx (2019). "A catalog of single nucleotide changes distinguishing
612 modern humans from archaic hominins". In: *Scientific Reports* 9.1, p. 8463. DOI: [10.1038/s41598-019-44877-x](https://doi.org/10.1038/s41598-019-44877-x).
- 613
- 614 Kulakovskiy, Ivan V et al. (2018). "HOCOMOCO: towards a complete collection of transcription fac-
615 tor binding models for human and mouse via large-scale ChIP-Seq analysis". In: *Nucleic Acids
616 Research* 46.Database issue, pp. D252–D259. DOI: [10.1093/nar/gkx1106](https://doi.org/10.1093/nar/gkx1106).

- 617 Landrum, Melissa J. et al. (2018). "ClinVar: improving access to variant interpretations and support-
618 ing evidence". In: *Nucleic Acids Research* 46.D1, pp. D1062–D1067. DOI: [10.1093/nar/gkx1153](https://doi.org/10.1093/nar/gkx1153).
- 619 Lange, Christian, Wieland B. Huttner, and Federico Calegari (2009). "Cdk4/CyclinD1 Overexpression
620 in Neural Stem Cells Shortens G1, Delays Neurogenesis, and Promotes the Generation and
621 Expansion of Basal Progenitors". In: *Cell Stem Cell* 5.3, pp. 320–331. DOI: [10.1016/j.stem.2009.05.
622 026](https://doi.org/10.1016/j.stem.2009.05.026).
- 623 Libé-Philippot, Baptiste and Pierre Vanderhaeghen (2021). "Cellular and Molecular Mechanisms
624 Linking Human Cortical Development and Evolution". In: *Annual Review of Genetics* 55.1, pp. 555–
625 581. DOI: [10.1146/annurev-genet-071719-020705](https://doi.org/10.1146/annurev-genet-071719-020705).
- 626 Lo Turco, Joseph J. and Arnold Kriegstein (1991). "Clusters of Coupled Neuroblasts in Embryonic
627 Neocortex". In: *Science* 252.5005, pp. 563–566. DOI: [10.1126/science.1850552](https://doi.org/10.1126/science.1850552).
- 628 López-Tobón, Alejandro et al. (2019). "Human Cortical Organoids Expose a Differential Function of
629 GSK3 on Cortical Neurogenesis". In: *Stem Cell Reports* 13.5, pp. 847–861. DOI: [10.1016/j.stemcr.
630 2019.09.005](https://doi.org/10.1016/j.stemcr.2019.09.005).
- 631 Luecken, Malte D and Fabian J Theis (2019). "Current best practices in single-cell RNA-seq analysis:
632 a tutorial". In: *Molecular Systems Biology* 15.6, e8746. DOI: [10.15252/msb.20188746](https://doi.org/10.15252/msb.20188746).
- 633 Lui, Jan H., David V. Hansen, and Arnold Kriegstein (2011). "Development and evolution of the hu-
634 man neocortex". In: *Cell* 146.1, pp. 18–36. DOI: [10.1016/j.cell.2011.06.030](https://doi.org/10.1016/j.cell.2011.06.030).
- 635 Mangan, Riley J. et al. (2022). "Adaptive sequence divergence forged new neurodevelopmental en-
636 hancers in humans". In: *Cell* 185.24. Publisher: Elsevier, 4587–4603.e23. DOI: [10.1016/j.cell.2022.
637 10.016](https://doi.org/10.1016/j.cell.2022.10.016).
- 638 Markenscoff-Papadimitriou, Eirene et al. (2020). "A Chromatin Accessibility Atlas of the Developing
639 Human Telencephalon". In: *Cell* 182.3, 754–769.e18. DOI: [10.1016/j.cell.2020.06.002](https://doi.org/10.1016/j.cell.2020.06.002).
- 640 Masilamani, A. P. et al. (2017). "KLF6 depletion promotes NF-κB signaling in glioblastoma". In: *Oncog-
641 ene* 36.25, pp. 3562–3575. DOI: [10.1038/onc.2016.507](https://doi.org/10.1038/onc.2016.507).
- 642 McArthur, Evonne et al. (2022). *Reconstructing the 3D genome organization of Neanderthals reveals
643 that chromatin folding shaped phenotypic and sequence divergence*. Pages: 2022.02.07.479462 Sec-
644 tion: New Results. DOI: [10.1101/2022.02.07.479462](https://doi.org/10.1101/2022.02.07.479462).
- 645 McLean, Cory Y. et al. (2010). "GREAT improves functional interpretation of cis-regulatory regions".
646 In: *Nature Biotechnology* 28.5, pp. 495–501. DOI: [10.1038/nbt.1630](https://doi.org/10.1038/nbt.1630).
- 647 Meyer, Matthias et al. (2012). "A high-coverage genome sequence from an archaic Denisovan indi-
648 vidual". In: *Science (New York, N.Y.)* 338.6104, pp. 222–226. DOI: [10.1126/science.1224344](https://doi.org/10.1126/science.1224344).
- 649 Mora-Bermúdez, Felipe, Farhath Badsha, et al. (2016). "Differences and similarities between human
650 and chimpanzee neural progenitors during cerebral cortex development". In: *eLife* 5. Ed. by
651 Andrea Musacchio. Publisher: eLife Sciences Publications, Ltd, e18683. DOI: [10.7554/eLife.18683](https://doi.org/10.7554/eLife.18683).
- 652 Mora-Bermúdez, Felipe, Philipp Kanis, et al. (2022). "Longer metaphase and fewer chromosome
653 segregation errors in modern human than Neanderthal brain development". In: *Science Ad-
654 vances* 8.30. Publisher: American Association for the Advancement of Science, eabn7702.
- 655 Moriano, Juan and Cedric Boeckx (2020). "Modern human changes in regulatory regions implicated
656 in cortical development". In: *BMC Genomics* 21.1, p. 304. DOI: [10.1186/s12864-020-6706-x](https://doi.org/10.1186/s12864-020-6706-x).
- 657 Mukhtar, Tanzila et al. (2022). "Temporal and sequential transcriptional dynamics define lineage
658 shifts in corticogenesis". In: *The EMBO Journal* 41.24. Publisher: John Wiley & Sons, Ltd, e111132.
659 DOI: [10.15252/emboj.2022111132](https://doi.org/10.15252/emboj.2022111132).
- 660 Namba, Takashi et al. (2021). "Metabolic Regulation of Neocortical Expansion in Development and
661 Evolution". In: *Neuron* 109.3, pp. 408–419. DOI: [10.1016/j.neuron.2020.11.014](https://doi.org/10.1016/j.neuron.2020.11.014).
- 662 Nourse, Jamison L. et al. (2022). "Piezo1 regulates cholesterol biosynthesis to influence neural stem
663 cell fate during brain development". In: *The Journal of General Physiology* 154.10, e202213084.
664 DOI: [10.1085/jgp.202213084](https://doi.org/10.1085/jgp.202213084).
- 665 Nowakowski, Tomasz J. et al. (2016). "Transformation of the Radial Glia Scaffold Demarcates Two
666 Stages of Human Cerebral Cortex Development". In: *Neuron* 91.6, pp. 1219–1227. DOI: [10.1016/
667 j.neuron.2016.09.005](https://doi.org/10.1016/j.neuron.2016.09.005).

- 668 Pääbo, Svante (2014). "The Human Condition—A Molecular Approach". In: *Cell* 157.1, pp. 216–226.
669 DOI: [10.1016/j.cell.2013.12.036](https://doi.org/10.1016/j.cell.2013.12.036).
- 670 Paten, Benedict et al. (2008). "Genome-wide nucleotide-level mammalian ancestor reconstruction".
671 In: *Genome Research* 18.11, pp. 1829–1843. DOI: [10.1101/gr.076521.108](https://doi.org/10.1101/gr.076521.108).
- 672 Pebworth, Mark-Phillip et al. (2021). "Human intermediate progenitor diversity during cortical de-
673 velopment". In: *Proceedings of the National Academy of Sciences* 118.26. Publisher: Proceedings
674 of the National Academy of Sciences, e2019415118. DOI: [10.1073/pnas.2019415118](https://doi.org/10.1073/pnas.2019415118).
- 675 Peyrégne, Stéphane et al. (2017). "Detecting ancient positive selection in humans using extended
676 lineage sorting". In: *Genome Research* 27.9, pp. 1563–1572. DOI: [10.1101/gr.219493.116](https://doi.org/10.1101/gr.219493.116).
- 677 Pinson, Anneline and Wieland B. Huttner (2021). "Neocortex expansion in development and evolu-
678 tion—from genes to progenitor cell biology". In: *Current Opinion in Cell Biology*. Differentiation
679 and development 73, pp. 9–18. DOI: [10.1016/j.ceb.2021.04.008](https://doi.org/10.1016/j.ceb.2021.04.008).
- 680 Pinson, Anneline, Lei Xing, et al. (2022). "Human TKTL1 implies greater neurogenesis in frontal neo-
681 cortex of modern humans than Neanderthals". In: *Science* 377.6611, eabl6422. DOI: [10.1126/
682 science.abl6422](https://doi.org/10.1126/science.abl6422).
- 683 Polioudakis, Damon et al. (2019). "A Single-Cell Transcriptomic Atlas of Human Neocortical Devel-
684 opment during Mid-gestation". In: *Neuron* 103.5, 785–801.e8. DOI: [10.1016/j.neuron.2019.06.011](https://doi.org/10.1016/j.neuron.2019.06.011).
- 685 Pollen, Alex A., Aparna Bhaduri, et al. (2019). "Establishing Cerebral Organoids as Models of Human-
686 Specific Brain Evolution". In: *Cell* 176.4, 743–756.e17. DOI: [10.1016/j.cell.2019.01.017](https://doi.org/10.1016/j.cell.2019.01.017).
- 687 Pollen, Alex A., Umut Kilik, et al. (2023). "Human-specific genetics: new tools to explore the molec-
688 ular and cellular basis of human evolution". In: *Nature Reviews Genetics*, pp. 1–25. DOI: [10.1038/
689 s41576-022-00568-4](https://doi.org/10.1038/s41576-022-00568-4).
- 690 Pollen, Alex A., Tomasz J. Nowakowski, et al. (2015). "Molecular Identity of Human Outer Radial Glia
691 during Cortical Development". In: *Cell* 163.1, pp. 55–67. DOI: [10.1016/j.cell.2015.09.004](https://doi.org/10.1016/j.cell.2015.09.004).
- 692 Prüfer, Kay, Cesare de Filippo, et al. (2017). "A high-coverage Neandertal genome from Vindija Cave
693 in Croatia". In: *Science (New York, N.Y.)* 358.6363, pp. 655–658. DOI: [10.1126/science.aao1887](https://doi.org/10.1126/science.aao1887).
- 694 Prüfer, Kay, Fernando Racimo, et al. (2014). "The complete genome sequence of a Neanderthal
695 from the Altai Mountains". In: *Nature* 505.7481, pp. 43–49. DOI: [10.1038/nature12886](https://doi.org/10.1038/nature12886).
- 696 Quinlan, Aaron R. and Ira M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing ge-
697 nomic features". In: *Bioinformatics* 26.6, pp. 841–842. DOI: [10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033).
- 698 Rakic, Pasko (1995). "A small step for the cell, a giant leap for mankind: a hypothesis of neocortical
699 expansion during evolution". In: *Trends in Neurosciences* 18.9, pp. 383–388. DOI: [10.1016/0166-
700 2236\(95\)93934-P](https://doi.org/10.1016/0166-2236(95)93934-P).
- 701 Saadat, Khandakar A. S. M. (2013). "[Role of ARID3A in E2F target gene expression and cell growth]".
702 In: *Kokubyo Gakkai Zasshi. The Journal of the Stomatological Society, Japan* 80.1, pp. 15–20.
- 703 Saito, Kanako et al. (2009). "Ablation of cholesterol biosynthesis in neural stem cells increases their
704 VEGF expression and angiogenesis but causes neuron apoptosis". In: *Proceedings of the National
705 Academy of Sciences* 106.20, pp. 8350–8355. DOI: [10.1073/pnas.0903541106](https://doi.org/10.1073/pnas.0903541106).
- 706 Schlebusch, Carina M. et al. (2017). "Southern African ancient genomes estimate modern human
707 divergence to 350,000 to 260,000 years ago". In: *Science* 358.6363. Publisher: American Associ-
708 ation for the Advancement of Science, pp. 652–655. DOI: [10.1126/science.aao6266](https://doi.org/10.1126/science.aao6266).
- 709 Setty, Manu et al. (2019). "Characterization of cell fate probabilities in single-cell data with Palantir".
710 In: *Nature Biotechnology* 37.4. Number: 4 Publisher: Nature Publishing Group, pp. 451–460. DOI:
711 [10.1038/s41587-019-0068-4](https://doi.org/10.1038/s41587-019-0068-4).
- 712 Sherry, S. T. et al. (2001). "dbSNP: the NCBI database of genetic variation". In: *Nucleic Acids Research*
713 29.1, pp. 308–311. DOI: [10.1093/nar/29.1.308](https://doi.org/10.1093/nar/29.1.308).
- 714 Silbereis, John C. et al. (2016). "The Cellular and Molecular Landscapes of the Developing Human
715 Central Nervous System". In: *Neuron* 89.2, pp. 248–268. DOI: [10.1016/j.neuron.2015.12.008](https://doi.org/10.1016/j.neuron.2015.12.008).
- 716 Skoglund, Pontus et al. (2017). "Reconstructing Prehistoric African Population Structure". In: *Cell*
717 171.1. Publisher: Elsevier, 59–71.e21. DOI: [10.1016/j.cell.2017.08.049](https://doi.org/10.1016/j.cell.2017.08.049).

- 718 Song, Michael et al. (2020). "Cell-type-specific 3D epigenomes in the developing human cortex". In:
719 *Nature* 587.7835, pp. 644–649. DOI: [10.1038/s41586-020-2825-4](https://doi.org/10.1038/s41586-020-2825-4).
- 720 Stepanova, Vita et al. (2021). "Reduced purine biosynthesis in humans after their divergence from
721 Neandertals". In: *Elife* 10. Publisher: eLife Sciences Publications, Ltd, e58741.
- 722 Sun, Li-Ping et al. (2005). "Insig required for sterol-mediated inhibition of Scap/SREBP binding to
723 COPII proteins in vitro". In: *The Journal of Biological Chemistry* 280.28, pp. 26483–26490. DOI:
724 [10.1074/jbc.M504041200](https://doi.org/10.1074/jbc.M504041200).
- 725 Syafruddin, Saiful E. et al. (2019). "A KLF6-driven transcriptional network links lipid homeostasis
726 and tumour growth in renal carcinoma". In: *Nature Communications* 10.1, p. 1152. DOI: [10.1038/
727 s41467-019-09116-x](https://doi.org/10.1038/s41467-019-09116-x).
- 728 Theofanopoulou, Constantina et al. (2017). "Self-domestication in Homo sapiens: Insights from
729 comparative genomics". In: *PLOS ONE* 12.10, pp. 1–23. DOI: [10.1371/journal.pone.0185306](https://doi.org/10.1371/journal.pone.0185306).
- 730 Torre-Ubieta, Luis de la et al. (2018). "The Dynamic Landscape of Open Chromatin during Human
731 Cortical Neurogenesis". In: *Cell* 172.1, 289–304.e18. DOI: [10.1016/j.cell.2017.12.014](https://doi.org/10.1016/j.cell.2017.12.014).
- 732 Trevino, Alexandro E. et al. (2021). "Chromatin and gene-regulatory dynamics of the developing
733 human cerebral cortex at single-cell resolution". In: *Cell* 184.19, 5053–5069.e23. DOI: [10.1016/j.
734 cell.2021.07.039](https://doi.org/10.1016/j.cell.2021.07.039).
- 735 Trujillo, Cleber A et al. (2021). "Reintroduction of the archaic variant of NOVA1 in cortical organoids
736 alters neurodevelopment". In: *Science* 371.6530. Publisher: American Association for the Ad-
737 vancement of Science, eaax2537.
- 738 Vanderhaeghen, Pierre and Franck Polleux (2023). "Developmental mechanisms underlying the
739 evolution of human cortical circuits". In: *Nature Reviews Neuroscience* 24.4, pp. 213–232. DOI:
740 [10.1038/s41583-023-00675-z](https://doi.org/10.1038/s41583-023-00675-z).
- 741 VanSickle, Caroline, Zachary Cofran, and David Hunt (2020). "Did Neandertals have large brains?
742 Factors affecting endocranial volume comparisons". In: *American Journal of Physical Anthropology*
743 173.4, pp. 768–775. DOI: [10.1002/ajpa.24124](https://doi.org/10.1002/ajpa.24124).
- 744 Villar, Diego, Paul Flicek, and Duncan T. Odom (2014). "Evolution of transcription factor binding in
745 metazoans — mechanisms and functional implications". In: *Nature Reviews Genetics* 15.4. Num-
746 ber: 4 Publisher: Nature Publishing Group, pp. 221–233. DOI: [10.1038/nrg3481](https://doi.org/10.1038/nrg3481).
- 747 Wang, Lei, Shirui Hou, and Young-Goo Han (2016). "Hedgehog signaling promotes basal progenitor
748 expansion and the growth and folding of the neocortex". In: *Nature neuroscience* 19.7, pp. 888–
749 896. DOI: [10.1038/nn.4307](https://doi.org/10.1038/nn.4307).
- 750 Wang, Zimei et al. (2018). "KLF6 and STAT3 co-occupy regulatory DNA and functionally synergize to
751 promote axon growth in CNS neurons". In: *Scientific Reports* 8.1, p. 12565. DOI: [10.1038/s41598-
752 018-31101-5](https://doi.org/10.1038/s41598-018-31101-5).
- 753 Weiss, Carly V et al. (2021). "The cis-regulatory effects of modern human-specific variants". In: *eLife*
754 10. Ed. by Patricia J Wittkopp, e63713. DOI: [10.7554/eLife.63713](https://doi.org/10.7554/eLife.63713).
- 755 Wilson, Sandra L. et al. (2012). "Primary cilia and Gli3 activity regulate cerebral cortical size". In:
756 *Developmental Neurobiology* 72.9, pp. 1196–1212. DOI: <https://doi.org/10.1002/dneu.20985>.
- 757 Yan, Guangmei et al. (2011). "Genome sequencing and comparison of two nonhuman primate an-
758 imal models, the cynomolgus and Chinese rhesus macaques". In: *Nature Biotechnology* 29.11,
759 pp. 1019–1023. DOI: [10.1038/nbt.1992](https://doi.org/10.1038/nbt.1992).
- 760 Yang, Tong et al. (2002). "Crucial Step in Cholesterol Homeostasis: Sterols Promote Binding of SCAP
761 to INSIG-1, a Membrane Protein that Facilitates Retention of SREBPs in ER". In: *Cell* 110.4. Pub-
762 lisher: Elsevier, pp. 489–500. DOI: [10.1016/S0092-8674\(02\)00872-3](https://doi.org/10.1016/S0092-8674(02)00872-3).
- 763 Zhang, Xinru, Bohao Fang, and Yi-Fei Huang (2023). "Transcription factor binding sites are fre-
764 quently under accelerated evolution in primates". In: *Nature Communications* 14.1. Number: 1
765 Publisher: Nature Publishing Group, p. 783. DOI: [10.1038/s41467-023-36421-3](https://doi.org/10.1038/s41467-023-36421-3).

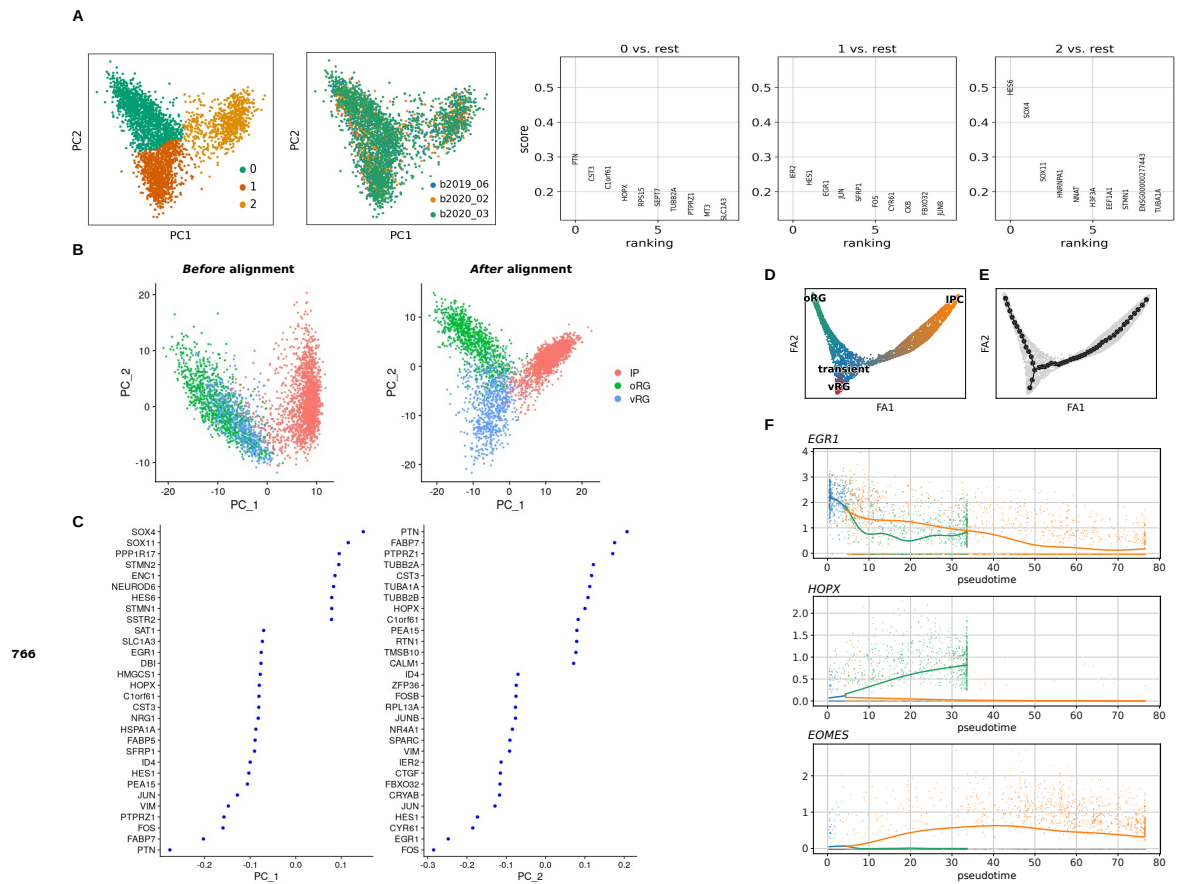


Figure 1—figure supplement 1. A) A coarse clustering (leiden algorithm; resolution 0.1) was used for differential gene expression analysis (logistic regression), which captured known markers for each progenitor subtype (0: oRG; 1:vRG; 2:IPC). Additionally, samples from different batches aggregate after normalization and integration Butler et al., 2018. B) A comparable dataset from (Polioudakis et al., 2019) was used to cross-validate findings obtained with the reference dataset Trevino et al., 2021. Polioudakis et al., 2019 dataset was processed similarly under Seurat analytical framework and projected into a shared low dimensional space, which allowed the discrimination of progenitor subtypes as main axes of variation via principal component analysis. C) Genes that most contribute to the first two principal component analysis in the shared low dimensional space. D) and E) Force-directed graph of neural progenitors from Polioudakis et al., 2019 dataset and projected principal tree on the force-directed graph, respectively. F) Recapitulation of the expected dynamics for three marker genes as pseudotime progresses.

Top 10 unique GO terms pseudotime-informed NMF – Modules 2 & 3

oRG_branch_M2	oRG_branch_M3	IPC_branch_M2	IPC_branch_M3
neurogenesis	nervous system development	cell projection organization	nervous system development
brain development	generation of neurons	plasma membrane-bounded cell proj.	system development
anatomical structure development	neurogenesis	cerebellum; molecular layer - neuropil[Low]	multicellular organism development
central nervous system development	neuron differentiation	extracellular region	anatomical structure development
columnar/cuboidal epithelial cell differentiation	multicellular organism development	cerebral cortex; neuropil[High]	neurogenesis
neuroepithelial cell differentiation	system development	structural constituent of cytoskeleton	developmental process
forebrain development	membrane	protein binding	generation of neurons
cell population proliferation	anatomical structure development	plasma membrane bounded cell projection	neuron differentiation
central nervous system neuron differentiation	extracellular region	Dysgenesis of the basal ganglia	multicellular organismal process
positive regulation of cell population proliferation	neuron projection development	cellular component morphogenesis	regulation of cellular process

Top 10 unique GO terms standard NMF – Modules 2 & 3

oRG_branch_M2	oRG_branch_M3	IPC_branch_M2	IPC_branch_M3
chromatin organization	Cell Cycle	regulation of cellular process	DNA metabolic process
Abnormality of the palpebral fissures	DNA metabolic process	regulation of biological process	DNA replication
Abnormal lip morphology	DNA replication	positive regulation of developmental process	Cell Cycle
Abnormality of the philtrum	Cell Cycle, Mitotic	chromatin	Cell Cycle, Mitotic
Abnormal upper lip morphology	chromosome	regulation of cell differentiation	DNA-templated DNA replication
Thick eyebrow	chromosome organization	Factor: sp4	cellular response to DNA damage stimulus
Cryptorchidism	cell cycle	Factor: ZXDL	Retinoblastoma gene in cancer
Abnormal eyebrow morphology	cellular response to DNA damage stimulus	biological regulation	DNA repair
Facial hypertrichosis	DNA-templated DNA replication	Factor: ETF	DNA strand elongation
hsa-miR-21-5p	DNA repair	regulation of developmental process	chromosome

Figure 2—figure supplement 1. Gene expression modules 2 and 3 captured by piNMF are sequentially activated as pseudotime progresses towards basal progenitor cell clusters. GO terms associated to these modules, for either oRG or IP cell clusters, belong to cardinal biological processes relevant for neural progenitor differentiation (upper table), while stdNMF does not fully resolve transient gene expression programs and GO terms are more generic (bottom table).

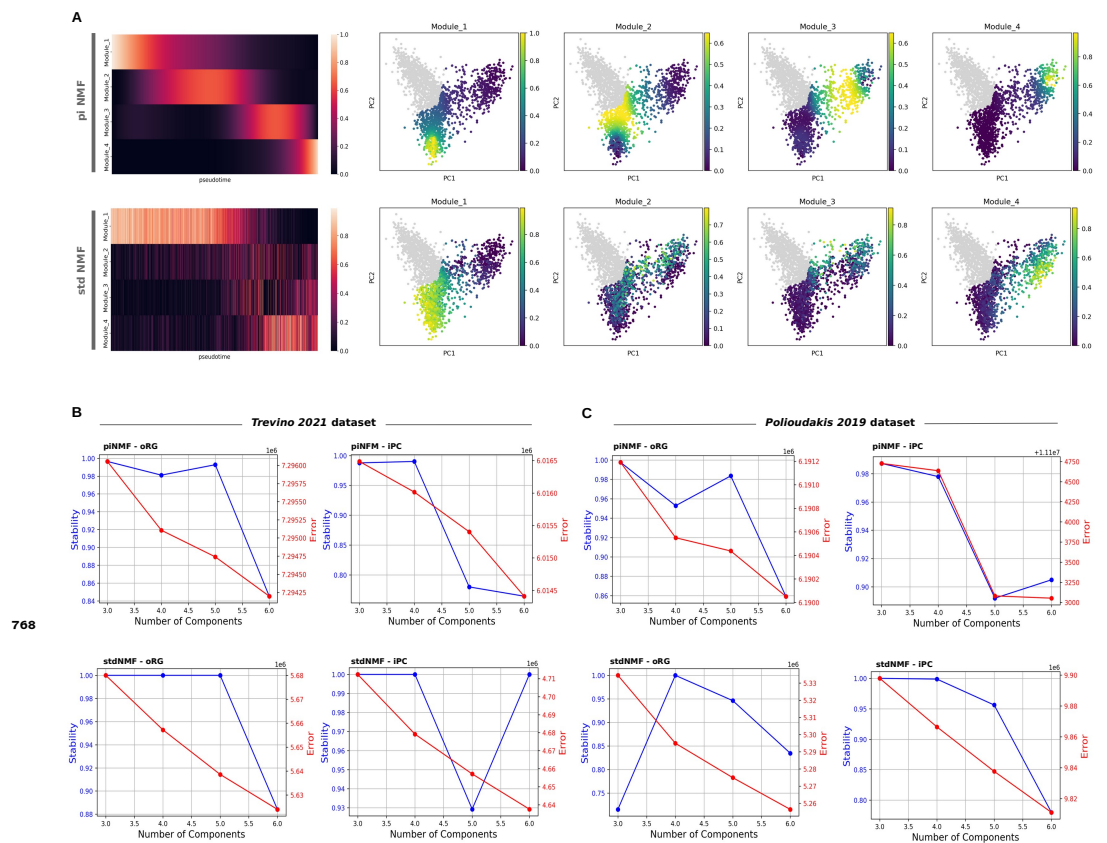


Figure 2—figure supplement 2. A) Similarly to the analysis on the oRG branch, piNMF better captures the continuous nature of gene expression programs activated along pseudotime on the IPC branch (see particularly heatmaps on the left), in contrast to stdNMF, specially for transient modules 2 and 3. B) and C) Factorization rank selection can be guided by a stability measure (silhouette score) of the resulting components (K-means clustering) over many replicates, and an error metric (Frobenius norm) to evaluate the distance between the original matrix and the NMF approximation. We observed, across branches (vRG to either oRG or IPC), datasets (from Trevino et al., 2021 and Polioudakis et al., 2019) and NMF algorithms (pseudotime-informed and standard NMF) factorization rank 4 as a reasonable selection allowing cross-evaluations, according to high stability and decreasing error. As there is not definitive solution for factorization rank selection, a detailed examination of the modules recovered is always required.

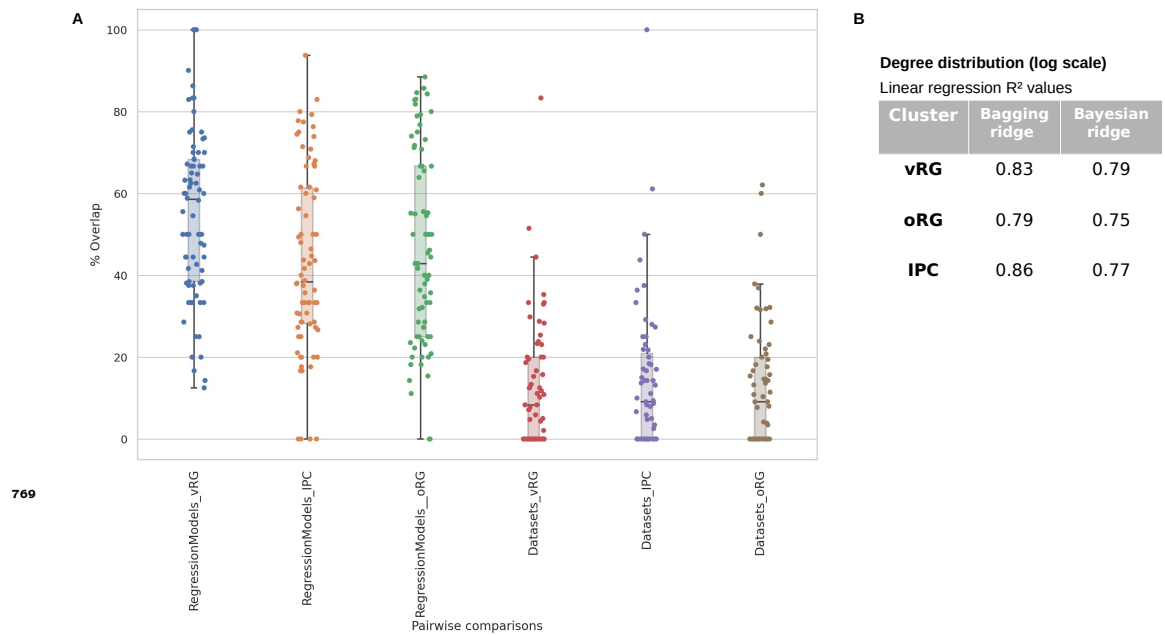


Figure 3—figure supplement 1. A) Significant overlaps (hypergeometric test; ST5) but substantial variability are detected in the TF-target gene pairs recovered by two machine learning-based Regression Models, bagging ridge and bayesian ridge algorithms from the *CellOracle* software (Kamimoto et al., 2023), when applied to the reference dataset Trevino et al., 2021 (between 43% to 55% depending on the cell cluster). More pronounced differences (overlaps between 12% to 14%) are observed when contrasting GRN Datasets: TF-target gene pairs obtained with *CellOracle* software compared to the regulatory networks (regulons) reported in Polioudakis et al., 2019, a comparable dataset based on *SCENIC* as GRN software (Aibar et al., 2017). B) Among *CellOracle* regression models, the bagging ridge model reports higher linear regression-based R² values for the degree distribution of the networks (log scale), and it was our choice for GRN analysis.

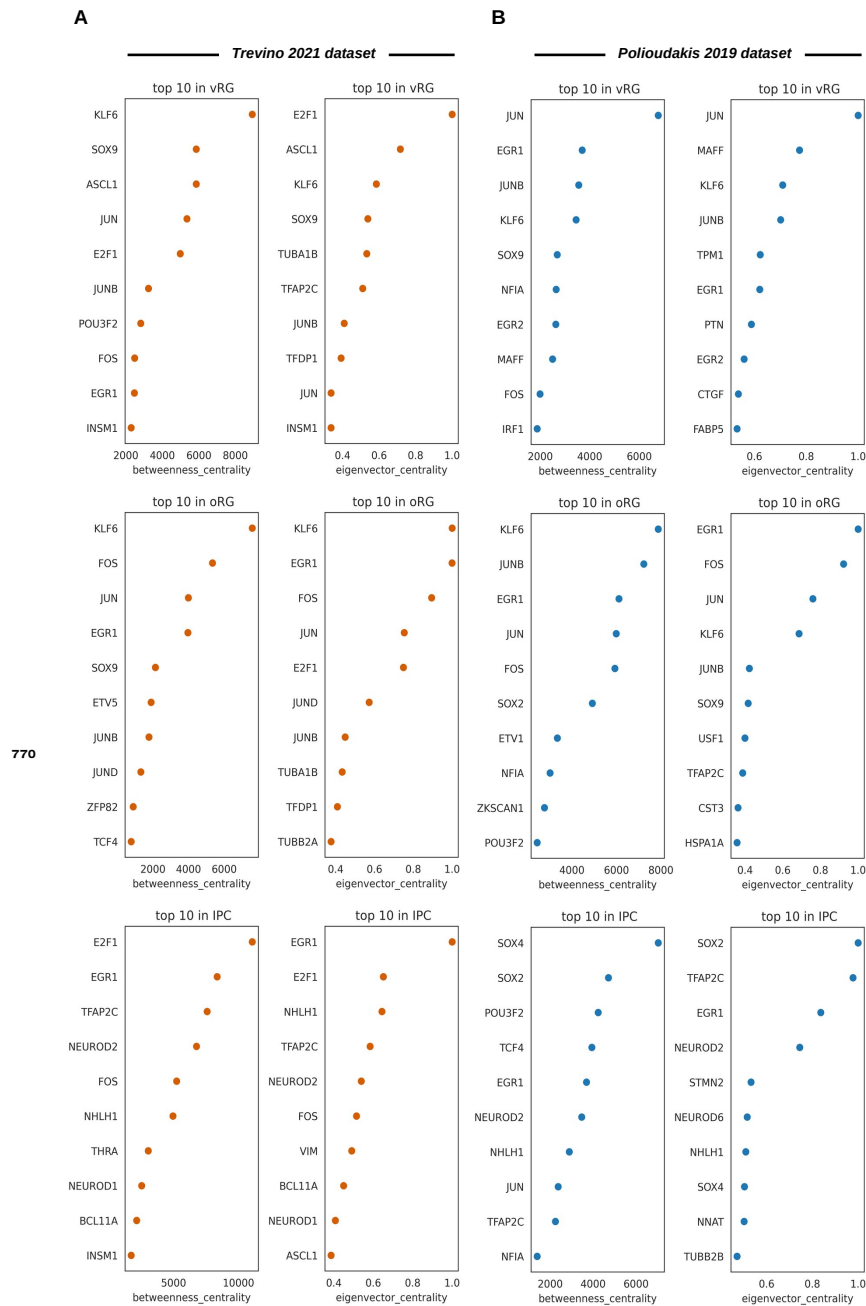


Figure 3—figure supplement 2. Networks measures (eigenvector centrality and betweenness centrality) for two independent datasets: A) Dataset from Trevino et al., 2021 and B) Dataset from Polioudakis et al., 2019. Genes identified as top 10 in both datasets include *KLF6*, *EGR1*, *JUN*, or *FOS* for radial glial clusters and *NHLH1*, *TFAP2C* or *NEUROD2* in intermediate progenitor clusters.

Chapter 4

A Brain Region-Specific Expression Profile for Genes Within Large Introgression Deserts and Under Positive Selection in *Homo sapiens*

Published as:

Buisán, R., Moriano, J., Andirkó, A., & Boeckx, C. 2022. A brain region-specific expression profile for genes within large introgression deserts and under positive selection in *Homo sapiens*. *Frontiers in Cell and Developmental Biology*
doi:[10.3389/fcell.2022.824740](https://doi.org/10.3389/fcell.2022.824740)



A Brain Region-Specific Expression Profile for Genes Within Large Introgression Deserts and Under Positive Selection in *Homo sapiens*

Raül Buisan^{1†}, Juan Moriano^{1,2†}, Alejandro Andirkó^{1,2} and Cedric Boeckx^{1,2,3*}

¹Universitat de Barcelona, Barcelona, Spain, ²Universitat de Barcelona Institute of Complex Systems, Barcelona, Spain, ³Catalan Institute for Research and Advanced Studies (ICREA), Barcelona, Spain

OPEN ACCESS

Edited by:

Elena Taverna,
Human Technopole, Italy

Reviewed by:

David E. MacHugh,
University College Dublin, Ireland
Fabrizio Mafessoni,
Weizmann Institute of Science, Israel

*Correspondence:

Cedric Boeckx
cedric.boeckx@ub.edu

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Stem Cell Research,
a section of the journal
Frontiers in Cell and Developmental
Biology

Received: 29 November 2021

Accepted: 04 April 2022

Published: 26 April 2022

Citation:

Buisan R, Moriano J, Andirkó A and
Boeckx C (2022) A Brain Region-
Specific Expression Profile for Genes
Within Large Introgression Deserts and
Under Positive Selection in
Homo sapiens.
Front. Cell Dev. Biol. 10:824740.
doi: 10.3389/fcell.2022.824740

Analyses of ancient DNA from extinct hominins have provided unique insights into the complex evolutionary history of *Homo sapiens*, intricately related to that of the Neanderthals and the Denisovans as revealed by several instances of admixture events. These analyses have also allowed the identification of introgression deserts: genomic regions in our species that are depleted of “archaic” haplotypes. The presence of genes like *FOXP2* in these deserts has been taken to be suggestive of brain-related functional differences between *Homo* species. Here, we seek a deeper characterization of these regions and the specific expression trajectories of genes within them, taking into account signals of positive selection in our lineage. Analyzing publicly available transcriptomic data from the human brain at different developmental stages, we found that structures outside the cerebral neocortex, in particular the cerebellum, the striatum and the mediodorsal nucleus of the thalamus show the most divergent transcriptomic profiles when considering genes within large introgression deserts and under positive selection.

Keywords: *Homo sapiens*, deserts of introgression, positive selection, cerebellum, striatum, thalamus, gene expression

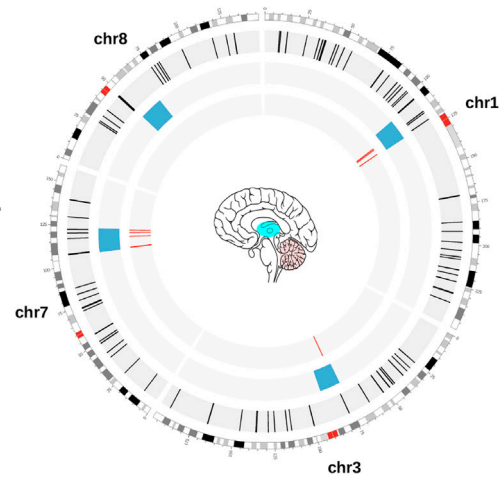
1 INTRODUCTION

The availability of high-coverage genomes from our closest extinct relatives, the Neanderthals and Denisovans, constitutes a significant advance in the range of questions one can ask about the deep history of our species (Meyer et al., 2012; Prüfer et al., 2014; Prüfer et al., 2017; Mafessoni et al., 2020). One of the main themes emerging from this progress is interbreeding. In recent years, a fairly large number of admixture events between Neanderthals, Denisovans and Sapiens populations have been postulated. A recent review (Bergström et al., 2021) considers that at least four such events are supported by strong evidence.

While it is important to ask whether our species benefited from these admixture events (so-called adaptive introgression, where alleles inherited from other hominins rose to high frequency as a result of positive selection after gene flow), it is also worth examining regions of the genomes that are depleted of alleles resulting from gene flow from other hominins (Sankararaman et al., 2016; Vernot et al., 2016; Chen et al., 2020; Skov et al., 2020; Rinker et al., 2020). Such regions are called introgression deserts (sometimes also “genomic islands of divergence/speciation” (Wang et al., 2020) and have now been identified in a range of species (Fontseré et al., 2019).

TABLE 1 | Genomic coordinates used in this study. Large deserts were retrieved from (Chen et al., 2020), and positively-selected regions from (Peyr egne et al., 2017) (see Section 4). The circo plot on the right shows the distribution of our regions of interest: Blue boxes: deserts of introgression; Red lines: positively-selected regions within deserts of introgression. Colored regions within the brain represent structures that figure prominently in this study.

Set	chr	Start	End
Large deserts	1	105400000	120600000
	3	74100000	89300000
	7	106200000	123200000
	8	49400000	66500000
Positively-selected regions within deserts	1	113427676	113560554
	1	114641362	114645248
	1	119322276	119387279
	3	77027847	77034264
	7	106877730	107233808
	7	116762909	116773234
	7	120147456	120174406
	7	122320035	122406480

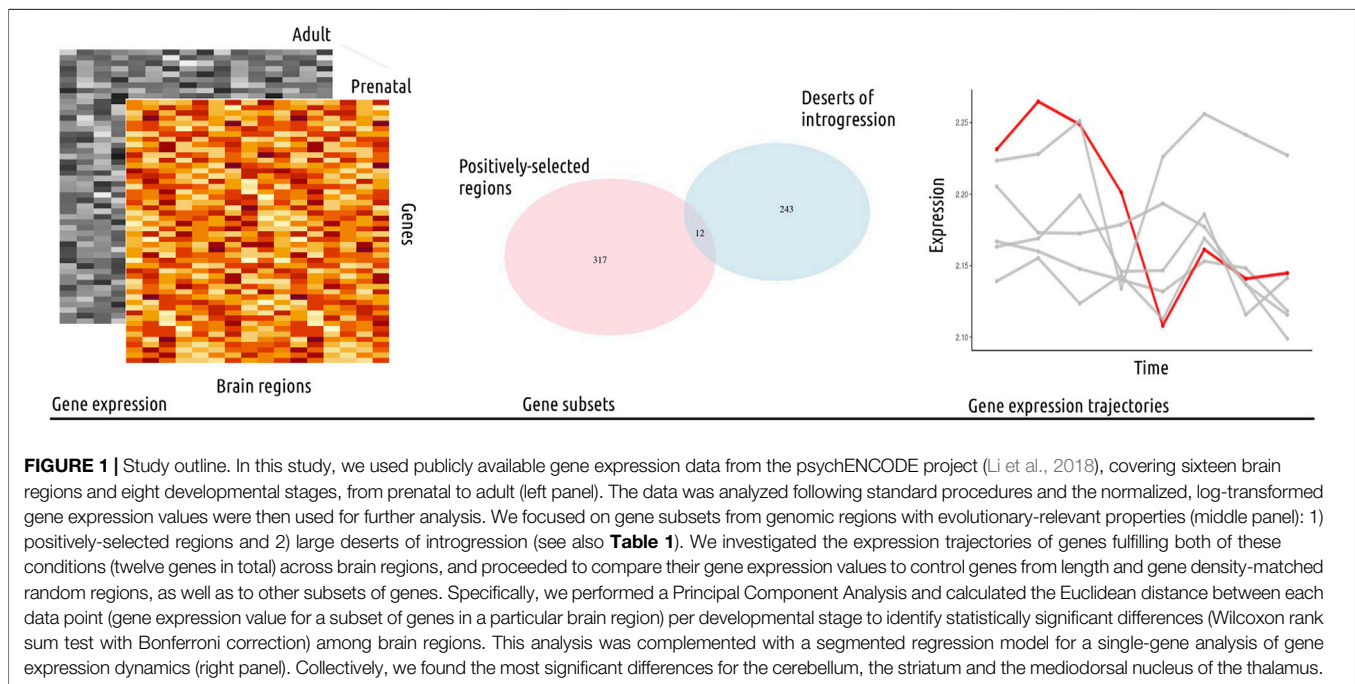


There are multiple reasons why genetic differences that arose after the divergence of populations may not be well tolerated (Wolf and Akey, 2018): there could be negative selection on “archaic” variants (deleterious changes on the “archaic” lineage), or positive selection on human-specific variants (adaptive changes on the human lineage), or it may be due to drift. It is reasonable to expect, and indeed has been shown, that the X chromosome constitutes such a desert region [not only in our species (Kuhlwilm et al., 2019; Martin and Jiggins, 2017)]. This could be due to repeated selective sweeps on this chromosome: genes involved in reproduction on this chromosome might act as strong reproductive barriers between populations (Fontser e et al., 2019).

In the case of modern humans, other genomic regions are devoid of Neanderthal and Denisovan introgression, for reasons that are perhaps less obvious, and therefore worth investigating further. A recent study (Chen et al., 2020) identifies four large deserts depleted of Neanderthal introgression, partially overlapping with a previous independent study (Vernot et al., 2016). As pointed out in (Kuhlwilm, 2018; Wolf and Akey, 2018), since it is likely that there were several different pulses of gene flow between us and our closest relatives (Iasi et al., 2021), the depletion observed in these four regions must have been reinforced repeatedly, and given the size of the deserts, it is reasonable that the “archaic” haplotype was purged within a short time after the gene flow event, as predicted by mathematical modeling on whole-genome simulations (Veller et al., 2021), and as evidenced in the analysis of genome-wide data from the earliest Late Pleistocene modern humans known to have been recovered in Europe (Hajdinjak et al., 2021).

The presence of *FOXP2*, a gene known for its role in language (Lai et al., 2001; Fisher, 2019), in one of these large deserts has attracted attention (Kuhlwilm, 2018), as it raises the possibility that the incompatibility between *Homo sapiens* and other hominin in such persistent introgression deserts may point to (subtle, but real) cognitive/behavioral differences. Indeed, the presence in such deserts of not only *FOXP2* but also other genes like *ROBO1*, *ROBO2*, and *EPHA3*, all independently associated with language traits (St Pourcain et al., 2014; Wang et al., 2015; Eising et al., 2021; Mekki et al., 2022), together with an earlier observation in Vernot et al. (2016) that genes within large deserts are significantly enriched in the developing cerebral cortex and in the adult striatum, suggest a possible point of entry into some of the most distinctive aspects of the human condition (P aabo, 2014). Such considerations, combined with independent evidence that introgressed Neanderthal alleles show significant downregulation in brain regions (McCoy et al., 2017), motivated us to focus on the brain in this study.

Specifically, we focused on the four largest genomic regions that resisted “archaic” introgression reported in (Chen et al., 2020), jointly with the most comprehensive catalog to date of signals of positive selection in our lineage (Peyr egne et al., 2017) (see Table 1), a combination that, to our knowledge, has not been previously studied in detail. Here, we tested if the genes that fulfill these two conditions (falling within large deserts of introgression and being under positive selection) follow particular (brain-region) expression trajectories that significantly deviate from that of other subsets of genes with evolutionary relevance or from control genomic regions. We characterized the gene expression dynamics (including genes falling within either



deserts of introgression or positively-selection regions alone) by analyzing transcriptomic data from several brain regions encompassing multiple developmental stages from prenatal to adulthood. This dataset allows for greater resolution than the Allen Brain Atlas data used in (Vernot et al., 2016), especially at early stages of development (see **Figure 1**). Three of the brain regions under study showed marked transcriptomic divergence (i.e., a statistically significant difference when compared to all other regions, based on the Principal Component Analysis-derived Euclidean distances): the cerebellum, the striatum and the thalamus. Among the genes at the intersection of regions under positive selection and large deserts of introgression, we found *CADPS2*, *ROBO2*, or *SYT6*, involved in neurotrophin release, axon guidance and neuronal proliferation, and known to be expressed in the brain regions our analysis highlights.

2 RESULTS

2.1 Genes in Large Deserts of Introgression Have Different Expression Levels Relative to the Rest of the Genome

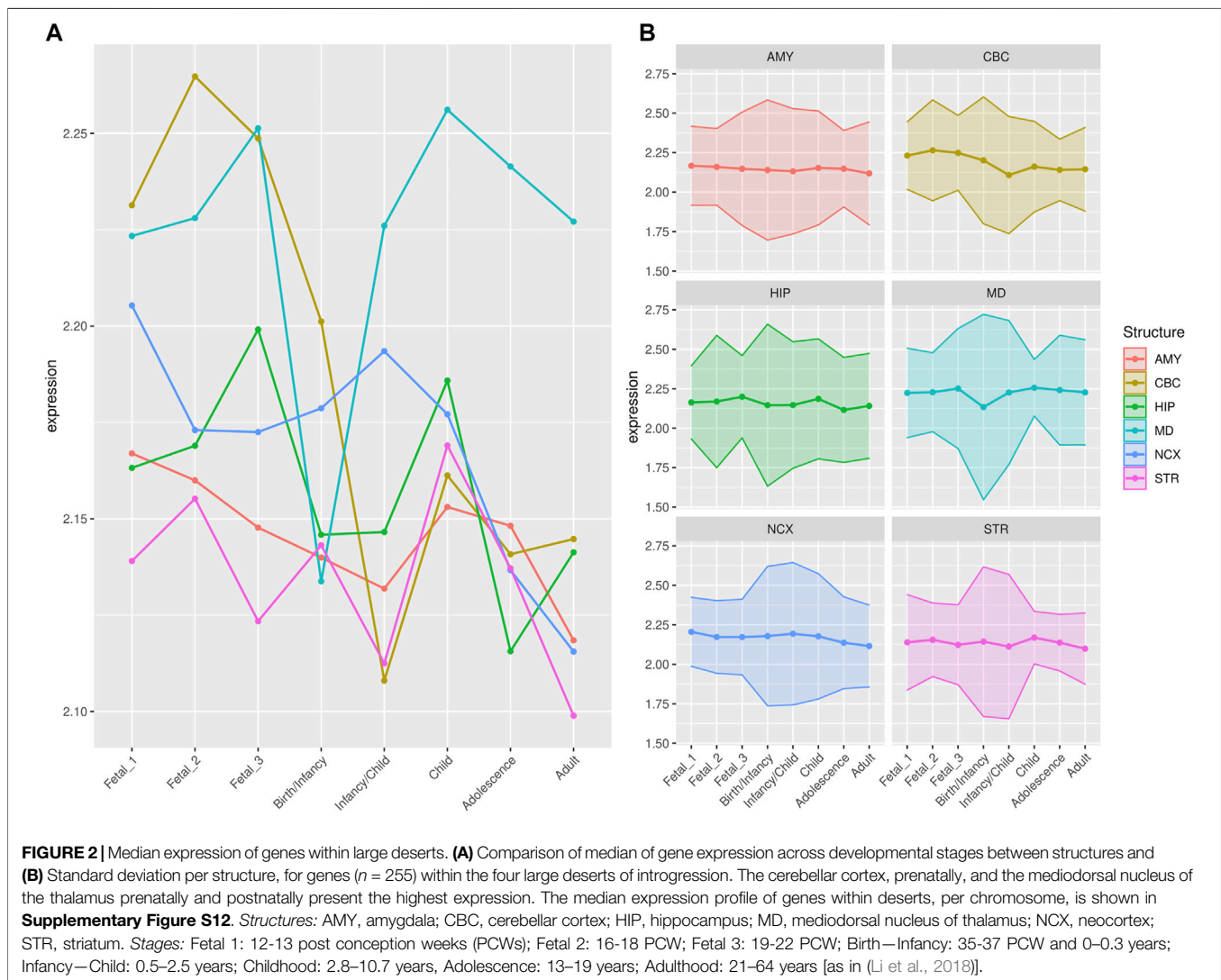
We set out to understand whether the mean expression of genes in large deserts of introgression (Chen et al., 2020) and the positively selected regions within them (extracted from (Peyrégne et al., 2017)) is significantly different compared to the rest of the genome, using publicly available transcriptomic data from the human brain (Li et al., 2018). To this end, we selected random regions of the genome ($n = 1,000$), excluding the large deserts, of the same average length (i.e., 15 million base-pairs), with a possible deviation of 1 million base-pairs to account for the length variability between different deserts of

introgression. To avoid genomic regions with low genetic density that might skew the results, the randomized areas were required to hold at least as many genes (265) as the desertic regions reported in (Chen et al., 2020).

The mean expression of genes lying in random regions of the genome was summarized for each brain structure (and log₂-transformed). A repeated-measures two-way ANOVA shows that the mean expression of both sets of these regions is significantly different from the rest of the genome ($p < 0.01$ for both sets). A post-hoc pairwise ANOVA (with Bonferroni correction) shows the difference between a gene expression value in a brain region as derived from the control set and that obtained from the genes in our two sets of interest is significant for most structures. An outlier's Grubbs test shows that the structures with the highest and lowest mean gene expression values in large deserts of introgression and the positively-selected windows within them fall inside the expected range of variability given the data ($p > 0.01$).

2.2 The Cerebellar Cortex, the Striatum and the Thalamus Show Divergent Transcriptomic Profiles When Considering Genes Within Large Deserts of Introgression and Under Positive Selection

We then investigated the temporal progression of the expression of genes within large deserts of introgression and putative positively-selected regions analyzing RNA-seq data of different human brain regions at different developmental stages (Li et al., 2018). We found that the median expression of genes within large deserts and positively-selected regions is higher than those present in deserts alone, the former peaking at prenatal stages in neocortical areas and decreasing later on. Outside the cerebral

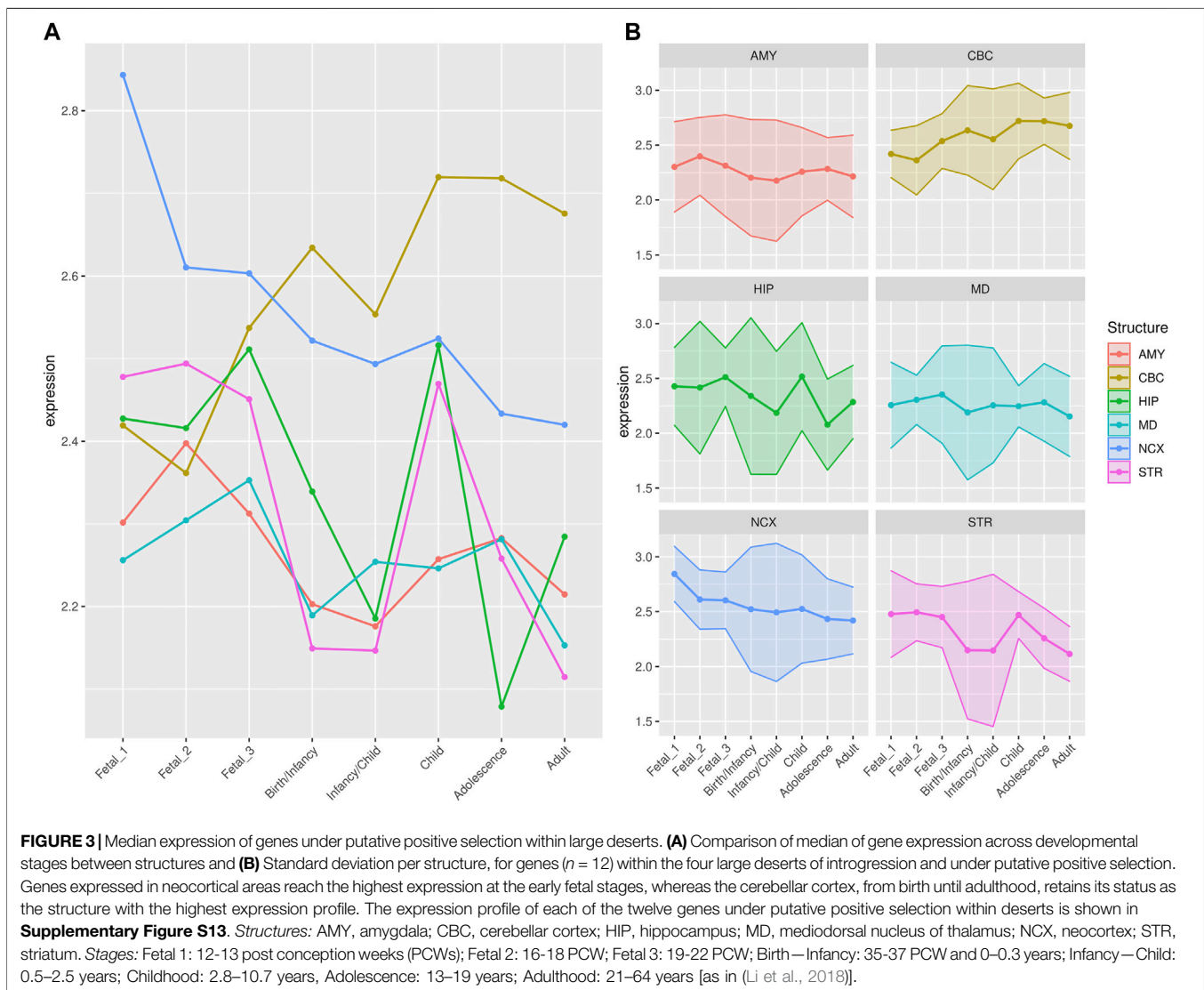


neocortex, this pronounced prenatal peak is not observed and, specifically for the cerebellar cortex, the expression profile of these genes increases before birth and reaches the highest median expression from childhood to adulthood in comparison to the rest of structures (see **Supplementary Figure S1**).

In order to statistically evaluate the differences observed for each structure and developmental stage (see **Figures 2, 3**), we performed a Principal Component Analysis and calculated the pairwise Euclidean distances between brain regions for each developmental stage using statistically significant principal components ($p < 0.05$) as assessed using the JackStraw analysis implemented in Seurat (Butler et al., 2018). For genes within large deserts of introgression overlapping putative positively-selected regions, we performed dimensionality reduction on the first two principal components. Due to the low number of genes at this intersection ($n = 12$), the second principal component did not report statistical significance. The sum of the percentage of variance explained by first and second components is around 50%. The transcriptomic profile of a brain

region in a given developmental stage was considered “divergent” if the expression value of the subset of genes under consideration was significantly different ($p < 0.01$) in that region when compared to all other regions (performing a Wilcoxon rank sum test with Bonferroni correction).

For genes that reside in the deserts of introgression under consideration, the cerebellum stands out as the structure with the most divergent transcriptomic profile at postnatal stages, from childhood to adulthood (**Figure 4**). For genes under positive selection that are also found within introgression deserts, the cerebellum still remains as the most transcriptomically divergent structure postnatally (birth/infancy, childhood, adolescence and adulthood; see the caption of **Figure 2** for the specific time points associated to each developmental stage). Moreover, prenatally, the cerebellum again (fetal stages 1 and 2) and the mediodorsal nucleus of the thalamus (fetal stage 1; see **Supplementary Figures S2, S3**) exhibit the most significant differences in the pairwise comparisons. Previous research found that genes within large deserts are over-represented in the striatum at adolescence and

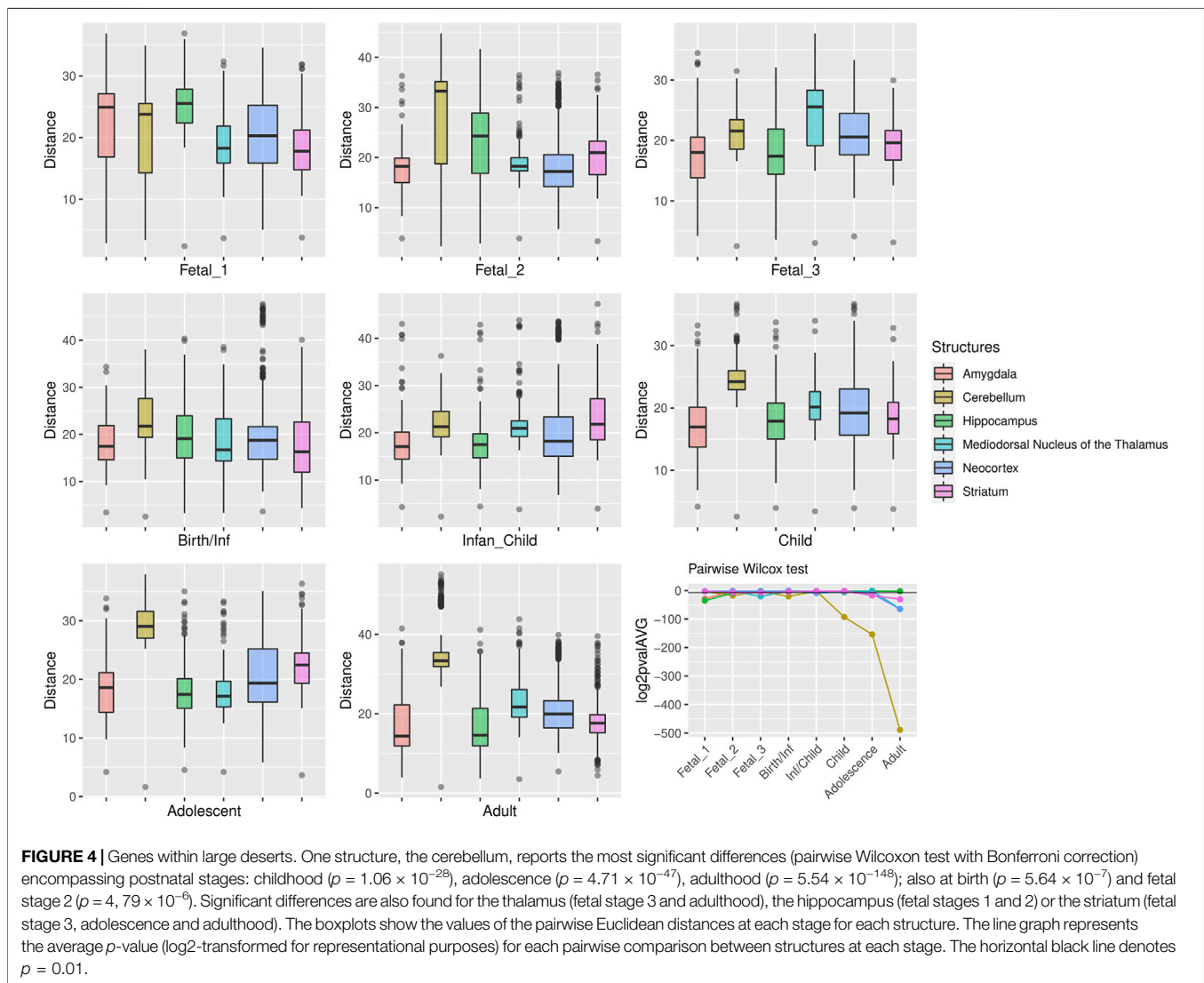


adult stages (Vernot et al., 2016). In agreement with this finding, we found that the transcriptomic profile of the striatum for genes within large deserts is significantly different at adolescence and adulthood but also at fetal stage 3, while for genes within deserts under putative positive selection, significant differences are found at infancy and adolescence (see **Figures 4, 5**, and **Supplementary Figure S4**). Lastly, to disentangle the effect of set of genes within specific chromosomes, we also evaluated the expression dynamics of genes within large deserts of introgression for each of the four chromosomal regions separately (a corresponding evaluation of the twelve genes under putative positive selection within deserts is presented in the next section). Overall, and in agreement with the previous observations, the cerebellum (at perinatal and later postnatal stages for the four chromosomes) and the striatum (at adulthood for three out of four chromosomes, and childhood for one chromosomal region) are found as the most transcriptomically divergent structures. The transcriptomic profile of the mediodorsal nucleus of the thalamus was also

found to be statistically different at fetal stages for chromosome 1 and chromosome 8 (see **Supplementary Figures S5–S8**).

For the sake of comparison, we note that a similar profile postnatally was obtained for the cerebellum when subsetting for genes under positive selection not present within large introgression deserts (marked differences from childhood to adulthood; see **Supplementary Figure S9**). When evaluating the global expression profile ($n = 9,358$ genes), the cerebellum shows statistically significant differences also at postnatal stages (birth, infancy, childhood and adulthood) and the mediodorsal nucleus of the thalamus at fetal stage 3 and adulthood (see **Supplementary Figure S10**). All p -values can be found in the Supplementary files.

The trajectories of expression across developmental stages in genes within large deserts of introgression might be affected by positive selection. To control for this, we analyzed the contrast between a control group of genes not under positive selection but

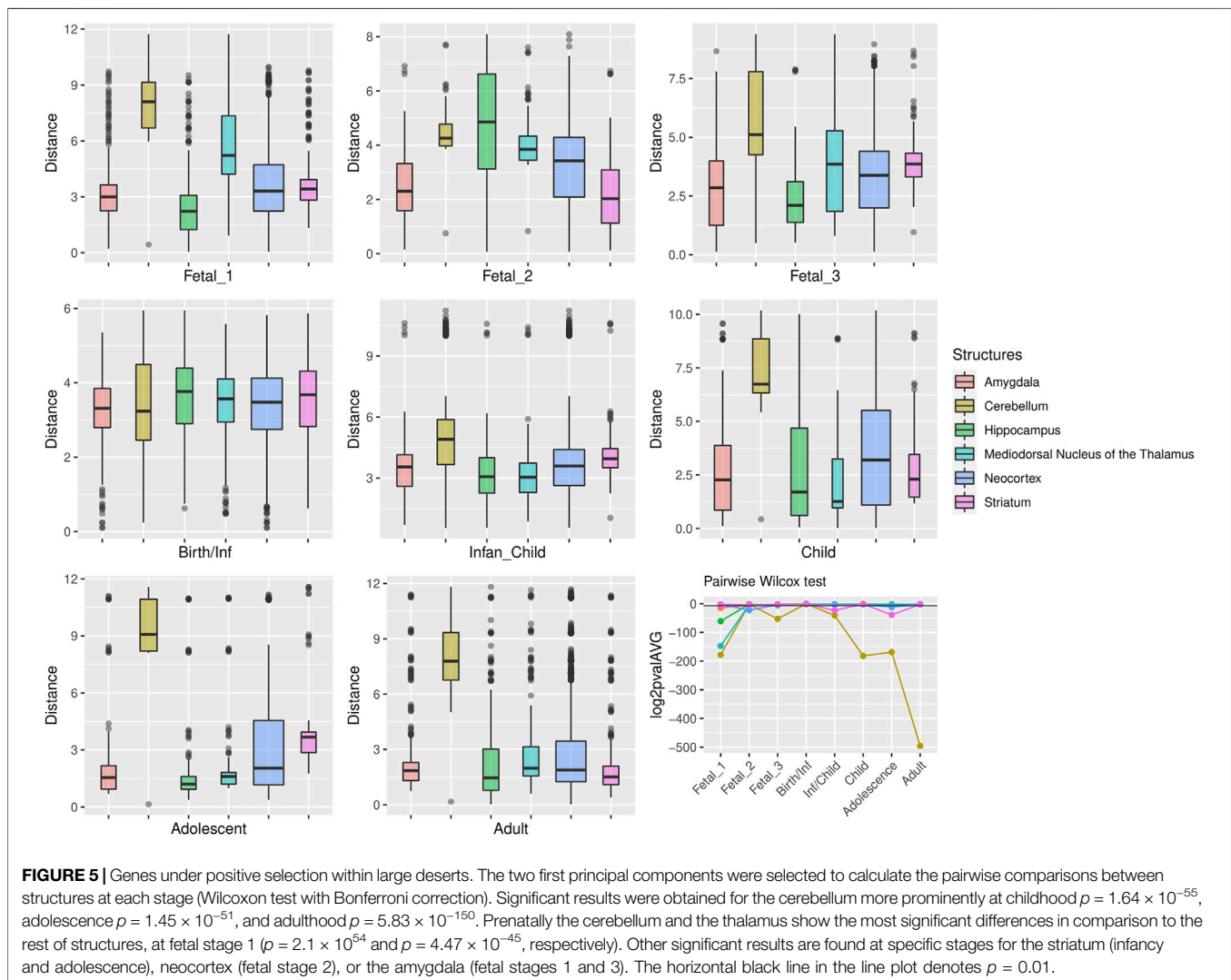


within deserts of introgression compared to those under positive selection in these same regions. We found that, within large deserts of introgression, genes under positive selection have an overall lower expression than those in regions not under positive selection ($p = 0.0007$, Kruskal-Wallis test). A linear regression model predicts that this effect is not structure-specific ($p = 0.655$), and that overall variability in the data is not explained by between-structure differences ($p = 0.9904$, ANOVA test between fitted models that do and do not include brain regions as a variable). Expression linked to specific developmental stages diverges significantly between genes under positive selection and those that are not (0.0001, linear regression). However, a post-hoc TUKEY test (corrected for repeated measures, **Supplementary Figure S11**) reveals that this difference holds only at the fetal stages. In portions of large deserts not under selection, the fetal period of development is significantly different from most posterior stages, while in genes under selective

pressures only the first fetal stage is significantly different from post-fetal stages (with a significance threshold of $p < 0.05$).

2.3 Gene-specific Expression Trajectories of Genes in the Overlapping Desertic and Positively-Selected Regions

As described in **section 4**, we included in our analyses any outlier present in the set of genes that are either within the four large deserts of introgression or under putative positive selection within large deserts, due to their potential evolutionary relevance. To evaluate in more detail the expression of specific genes, we focused on the specific trajectories of genes at the intersection of large deserts and positively-selected regions ($n = 12$ genes; **Supplementary Figure S13**), and performed a segmented regression analysis (using the Trendy package (Bacher et al., 2018)) filtering out genes with an adjusted R^2

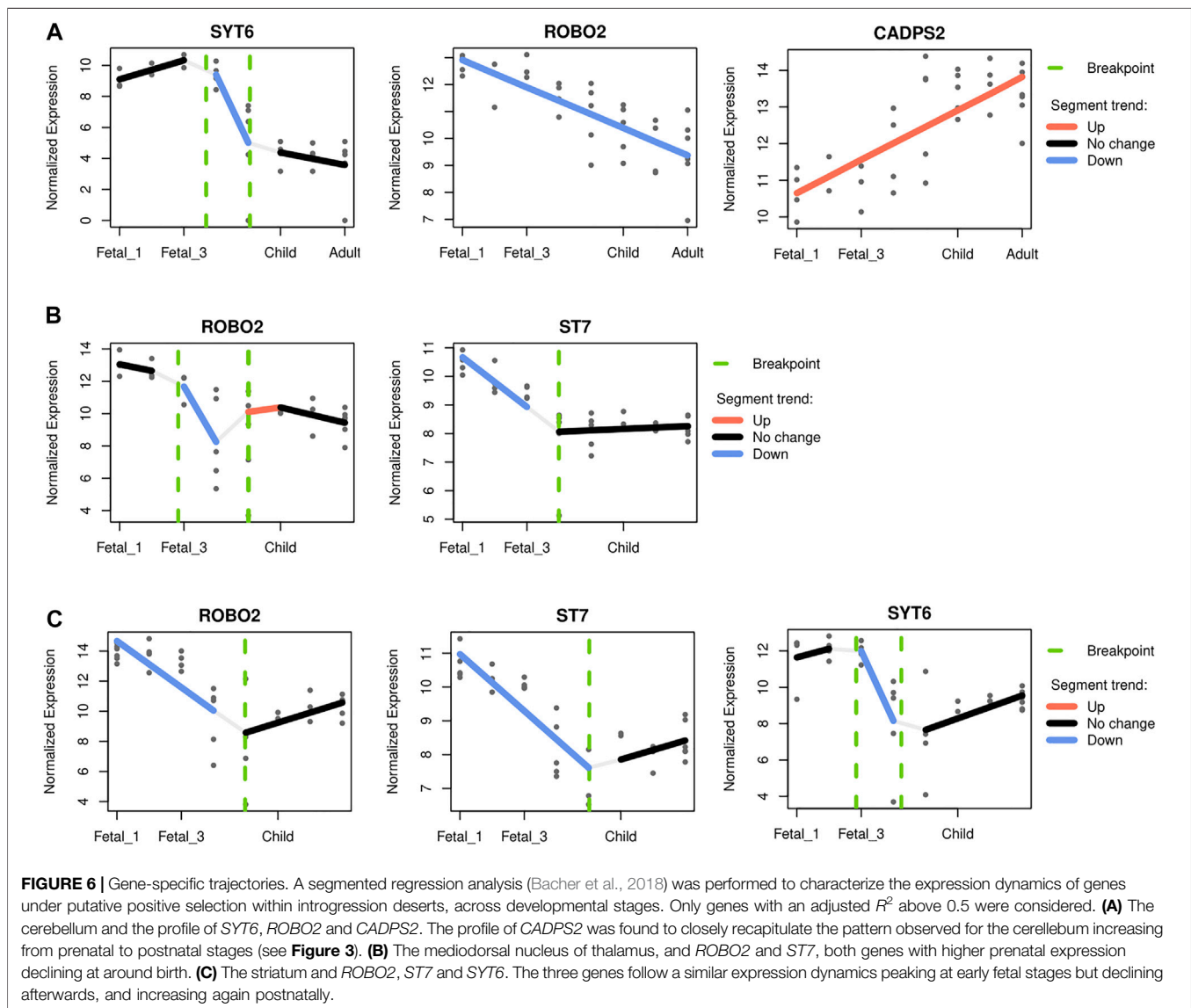


less than 0.5. As our analysis showed a marked increase of transcriptomic divergence at different developmental stages for the cerebellum, the striatum and the mediodorsal nucleus of the thalamus, we decided to focus on these structures.

For the cerebellum, *CADPS2* (chromosome 7) expression is the one that most closely mimics the observed pattern, with highest postnatal expression and a marked increase of its expression around birth and infancy (R^2 0.56; see **Supplementary Figure S13** and **Figure 6A**). This Ca^{2+} -dependent activator protein is known to regulate exocytosis in granule cells, particularly neurotrophic factors BDNF and NT-3 release, and its knockout disrupts normal cerebellar development and causes an autistic-like behavioral phenotype in mice (Sadakata et al., 2007; Sadakata et al., 2014). In addition, decreasing expression through developmental stages was also found for *SYT6* and *ROBO2* (chromosome 1 and 3 respectively; R^2 0.76 and 0.60; see **Figure 6A**). Two other genes, *KCND2* and *ST7* (both in chromosome 7), exhibited comparatively high expression postnatally, but did not pass the adjusted R^2 threshold (**Supplementary Figure S13**).

Regarding the thalamus, two genes within the overlapping desertic and positively-selected regions could be fitted with an adjusted R^2 higher than 0.5: *ROBO2* and *ST7*. Both genes show higher expressions at prenatal stages, followed by a steady decline at around birth (R^2 0.65 and 0.61, respectively; see **Figure 6B**). The roles of Robo2 in the thalamus have been studied as a receptor of the Slit/Robo signaling pathway which is critically involved in axon guidance. Indeed, Robo2 is highly expressed in the dorsal thalamus and cerebral cortex in the embryonic mouse brain and, in cooperation with Robo1, is required for the proper development of cortical and thalamic axonal projections (López-Bendito et al., 2007).

Lastly, for the striatum, three genes within the overlapping desertic and positively-selected regions could be fitted with an adjusted R^2 higher than 0.5. *ST7*, *ROBO2* and *SYT6* follow a V-shape profile with higher expression at prenatal stages, a decrease around birth, and increasing levels during later postnatal stages ($R^2 = 0.75, 0.57, \text{ and } 0.53$, respectively; see **Figure 6C**). While the role of *ST7* in neurodevelopment remains to be elucidated, Robo2 is a receptor of the Slit/Robo



signaling pathway which is critically involved in axon guidance (López-Bendito et al., 2007), but also in the proliferation and differentiation of neural progenitors with possible different roles in dorsal and ventral telencephalon (Andrews et al., 2008; Borrell et al., 2012). *Syt6* is another synapse-related gene expressed in the developing basal ganglia (Long et al., 2009), and in fact linked to the distinctive expression profile of this structure (Konopka et al., 2012). Additionally, *Syt6* shows a similar expression profile in the cerebellar cortex although at lower levels (see **Figure 6C**), a region where *Syt6* has been found, in mice, to be differentially expressed in a *Cadps2* knockout background (Sadakata et al., 2017).

3 DISCUSSION

There are two main findings to take away from our study: the importance of structures beyond the cerebral neocortex in the attempt to characterize some of the most derived features of our

species' brain, and the fact that some of the strongest effects in these regions takes place at early stages of development. In this way our work provides complementary evidence for the perinatal globularization phase as a species-specific ontogenic innovation (Gunz et al., 2010), and also provides new evidence for the claim that brain regions outside the neocortex (cerebellum, thalamus, striatum) significantly contribute to this phenotype (Boeckx and Benítez-Burraco, 2014; McCoy et al., 2017; Neubauer et al., 2018; Gunz et al., 2019; Weiss et al., 2021).

To our knowledge this is the first study to reveal the effect of the cerebellum in the context of large introgression deserts. For the striatum, previous studies have already highlighted the relevance of this structure: genes carrying Neanderthal-derived changes and expressed in the striatum during adolescence exhibit a higher McDonald-Kreitman ratio (Mafessoni et al., 2020). In addition, using a different range of introgressed regions and gene expression data from the Allen Brain Atlas (with lower temporal resolution than the database used in this study), it had already

been noted (Vernot et al., 2016) that genes within large deserts are significantly enriched in the striatum at adolescence and adult stages, which converges with the life stages highlighted from our analysis using the most recent report of genomic regions depleted of archaic variants (Chen et al., 2020).

Naturally, the functional effects of these divergent developmental profiles for the cerebellum, the prenatal thalamus or the striatum remain to be understood, particularly in the context of the possible differences among *Homo*-species concerning regulation of the genes highlighted in this study. This is especially relevant in light of emerging evidence that selection against DNA introgression is stronger in regulatory regions (Vilgalys et al., 2021), which in addition have been found to be over-represented in putative positively-selected regions in *Homo sapiens* (Peyrégne et al., 2017; Petr et al., 2019). The fact that early developmental stages are critical holds the promise of using brain organoid technology to probe the nature of these differences, since such *in vitro* techniques best track these earliest developmental windows (Muchnik et al., 2019; Mostajo-Radji et al., 2020; Kyrousi and Cappello, 2020). Our level of analysis (mRNA-seq data, informed by paleogenetic studies) can be complemented with other *omics* data to finely resolve cell-type specificities of the genes considered here across brain areas, as with the use of single-cell RNA-seq data, or to infer gene regulatory networks (from differentially accessible and methylated regions and chromatin immunoprecipitation data) that underlie the divergent gene expression trajectories observed.

The fact that *FOXP2* expression is known to be particularly high in the brain regions highlighted here (Lai et al., 2003) may help shed light on why *FOXP2* is found in one of the large introgression deserts in modern human genomes. As pointed out in (Kuhlwilm, 2018), this portion of chromosome 7 is not a desert for introgression in other great apes, nor did it act as a barrier for gene flow from *Sapiens* into *Neanderthals*. As such, it may indeed capture something genuinely specific about our species.

4 METHODS

Analyses were performed using R (R Core Team, 2019). Putative positively-selected regions were retrieved from the extended set of sweep regions in Peyrégne et al. (2017), built from two independent recombination maps using a Hidden Markov-based model applied to African and Neanderthal/Denisovan genomes. Coordinates for (large) deserts of introgression were retrieved from Chen et al. (2020), and genes within these two sets of regions were obtained using the BioMart R package version 2.42.1 (Durinck et al., 2009), using the respective genomic region coordinates as input and filtering by protein-coding genes.

mRNA-seq analysis. Publicly available transcriptomic data of the human brain at different developmental stages was retrieved from (Li et al., 2018) and analyzed using R (full code can be found at <https://github.com/jjaa-mp/desertsHomo>). Reads per kilo base per million mapped reads (RPKM) normalized counts were log-transformed and then subsetted to select genes either in large deserts of introgression or in both deserts and putative positively-

selected regions. The complete log-transformed, RPKM normalized count matrix was subsetted to select genes with median expression value > 2 , as in (Li et al., 2018), while no median filtering was employed for the subsets of genes within deserts and positively-selected regions, due to the potential relevance of the outliers in these specific regions for the purposes of our study. To assess transcriptomic variability between brain regions accounted for by genes either in large deserts or in deserts and positively-selected regions, we performed principal component analysis and calculated the pairwise Euclidean distances between brain regions for each dataset [following (Li et al., 2018)]. We then statistically evaluated such differences at each developmental stage using pairwise Wilcoxon tests with Bonferroni correction. Significant differences were considered if $p < 0.01$. Our analysis based on statistically significant principal components did not make it possible for us to use the Allen Brain Atlas data for comparisons with the psychENCODE project dataset used in this study, due to the more limited resolution, especially at prenatal stages, offered by the former.

To evaluate the expression profile of genes from our regions of interest in comparison to other regions of the human genome, we generated sets of random regions of the same length and gene density (that do not overlap with the genomic coordinates of deserts on introgression). These served as control regions for comparisons of mean expression values using two-way repeated measures ANOVA, implemented in R. ANOVA tests were performed taking mean expression values as dependent value, with structure names as subject identifiers and the different regions of interest (datasource) as between-subjects factor variable. Posthoc tests were performed similarly but with the mean expression data grouped by the datasource, obtaining an ANOVA table for each structure, with a Bonferroni correction to account for repeated measures. The stage-version of the ANOVA grouped subject identifiers by stage. Two Kruskal-Wallis tests were used, one designed to detect whether non positively-selected genes in deserts of introgression have different mean expression levels than genes that are both in deserts and in positively-selected windows; and the second to determine whether any particular brain structure has a particularly different expression mean than the rest, regardless of selection. We also used two two-level linear mixed-effects regression models, to compare non-positively selected genes and positively selected genes within introgression deserts. These models consist of repeated measures of expression on different brain structures in three different groups: control, deserts of introgression, and deserts with selection signals. The same model applies when stages are taken into account, replacing structure identifiers. Tukey's test was then used to fit the model.

Gene-specific expression trajectories. The R package Trendy version 1.8.2 (Bacher et al., 2018) was used to perform segmented regression analysis and characterize the expression trajectories of genes within both deserts of introgression and putative positively-selected regions (12 genes). The normalized RPKM values [from (Li et al., 2018)] in the form of a gene-by-time samples matrix was used to fit each gene expression trajectory to an optimal segmented regression model. Genes were considered if their

adjusted R^2 was >0.5 . In addition, a maximum number of breakpoints (significant changes in gene expression trajectory) was set at 3, minimum number of samples in each segment at 2, and minimum mean expression, 2.

The permutation tests using gene expression data from (Li et al., 2018) were done using the regioneR package version 1.26.1 (Gel et al., 2016) at $n = 1,000$.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <https://github.com/jjaa-mp/desertsHomo>.

AUTHOR CONTRIBUTIONS

Conceptualization: CB, AA, JM and RB; Data Curation: AA, JM and RB; Formal Analysis: AA, JM and RB; Visualization: CB, AA, JM and RB; Writing—Original Draft Preparation: CB, AA, JM, RB; Writing—Review and Editing: CB, AA, JM and RB; Supervision: CB; Funding Acquisition: CB.

REFERENCES

- Andrews, W., Barber, M., Hernandez-Miranda, L. R., Xian, J., Rakic, S., and Sundaresan, V. (2008). The Role of Slit-Robo Signaling in the Generation, Migration and Morphological Differentiation of Cortical Interneurons. *Develop. Biol.* 313, 648–658. doi:10.1016/j.ydbio.2007.10.052
- Bacher, R., Leng, N., Chu, L.-F., Ni, Z., Thomson, J. A., Kendzierski, C., et al. (2018). Trendy: Segmented Regression Analysis of Expression Dynamics in High-Throughput Ordered Profiling Experiments. *BMC Bioinformatics* 19, 380. doi:10.1186/s12859-018-2405-x
- Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L., and Skoglund, P. (2021). Origins of Modern Human Ancestry. *Nature* 590, 229–237. doi:10.1038/s41586-021-03244-5
- Boeckx, C. A., and Benitez-Burraco, A. (2014). The Shape of the Human Language-Ready Brain. *Front. Psychol.* 5. doi:10.3389/fpsyg.2014.00282
- Borrell, V., Cárdenas, A., Ciceri, G., Galcerán, J., Flames, N., Pla, R., et al. (2012). Slit/Robo Signaling Modulates the Proliferation of Central Nervous System Progenitors. *Neuron* 76, 338–352. doi:10.1016/j.neuron.2012.08.003
- Butler, A., Hoffman, P., Smibert, P., Papalexis, E., and Satija, R. (2018). Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species. *Nat. Biotechnol.* 36, 411–420. doi:10.1038/nbt.4096
- Chen, L., Wolf, A. B., Fu, W., Li, L., and Akey, J. M. (2020). Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. *Cell* 180, 677–687. e16. doi:10.1016/j.cell.2020.01.012
- Durinck, S., Spellman, P. T., Birney, E., and Huber, W. (2009). Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package biomaRt. *Nat. Protoc.* 4, 1184–1191. doi:10.1038/nprot.2009.97
- Eising, E., Mirza-Schreiber, N., de Zeeuw, E. L., Wang, C. A., Truong, D. T., Allegrini, A. G., et al. (2021). *Genome-wide Association Analyses of Individual Differences in Quantitatively Assessed reading- and Language-Related Skills in up to 34,000 People*. *bioRxiv* doi:10.1101/2021.11.04.466897
- Fisher, S. E. (2019). Human Genetics: The Evolving Story of FOXP2. *Curr. Biol.* 29, R65–R67. doi:10.1016/j.cub.2018.11.047
- Fontserè, C., Manuel, M. d., Marques-Bonet, T., and Kuhlwillm, M. (2019). Admixture in Mammals and How to Understand its Functional Implications. *BioEssays* 41, 1900123. doi:10.1002/bies.201900123
- Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., and Malinverni, R. (2016). regioneR: an R/Bioconductor Package for the

FUNDING

CB acknowledges support from the Spanish Ministry of Science and Innovation (grant PID2019-107042GB-I00), MEXT/JSPS Grant-in-Aid for Scientific Research on Innovative Areas #4903 (Evolinguistics: JP17H06379), Generalitat de Catalunya (2017-SGR-341), and the support of a 2020 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation. JM acknowledges financial support from the Departament d'Empresa i Coneixement, Generalitat de Catalunya (FI-SDUR 2020). AA acknowledges financial support from the Spanish Ministry of Economy and Competitiveness and the European Social Fund (BES-2017-080366). Funding bodies take no responsibility for the opinions, statements and contents of this project, which are entirely the responsibility of its authors.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcell.2022.824740/full#supplementary-material>

- Association Analysis of Genomic Regions Based on Permutation Tests. *Bioinformatics* 32, 289–291. doi:10.1093/bioinformatics/btv562
- Gunz, P., Neubauer, S., Maureille, B., and Hublin, J.-J. (2010). Brain Development after Birth Differs between Neanderthals and Modern Humans. *Curr. Biol.* 20, R921–R922. doi:10.1016/j.cub.2010.10.018
- Gunz, P., Tilot, A. K., Wittfeld, K., Teumer, A., Shapland, C. Y., van Erp, T. G., et al. (2019). Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity. *Curr. Biol.* 29, 120–127. e5. doi:10.1016/j.cub.2018.10.065
- Hajdinjak, M., Mafessoni, F., Skov, L., Vernot, B., Hübner, A., Fu, Q., et al. (2021). Initial Upper Palaeolithic Humans in Europe Had Recent Neanderthal Ancestry. *Nature* 592, 253–257. doi:10.1038/s41586-021-03335-3
- Iasi, L. N. M., Ringbauer, H., and Peter, B. M. (2021). An Extended Admixture Pulse Model Reveals the Limitations to Human–Neanderthal Introgression Dating. *Mol. Biol. Evol.* doi:10.1093/molbev/msab210
- Konopka, G., Friedrich, T., Davis-Turak, J., Winden, K., Oldham, M. C., Gao, F., et al. (2012). Human-Specific Transcriptional Networks in the Brain. *Neuron* 75, 601–617. doi:10.1016/j.neuron.2012.05.034
- Kuhlwillm, M., Han, S., Sousa, V. C., Excoffier, L., and Marques-Bonet, T. (2019). Ancient Admixture from an Extinct Ape Lineage into Bonobos. *Nat. Ecol. Evol.* 3, 957–965. doi:10.1038/s41559-019-0881-7
- Kuhlwillm, M. (2018). The Evolution of FOXP2 in the Light of Admixture. *Curr. Opin. Behav. Sci.* 21, 120–126. doi:10.1016/j.cobeha.2018.04.006
- Kyrousi, C., and Cappello, S. (2020). Using Brain Organoids to Study Human Neurodevelopment, Evolution and Disease. *WIREs Develop. Biol.* 9, e347. doi:10.1002/wdev.347
- Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F., and Monaco, A. P. (2001). A Forkhead-Domain Gene Is Mutated in a Severe Speech and Language Disorder. *Nature* 413, 519–523. doi:10.1038/35097076
- Lai, C. S. L., Gerrelli, D., Monaco, A. P., Fisher, S. E., and Copp, A. J. (2003). FOXP2 Expression during Brain Development Coincides with Adult Sites of Pathology in a Severe Speech and Language Disorder. *Brain* 126, 2455–2462. doi:10.1093/brain/awg247
- Li, M., Santpere, G., Kawasawa, Y. I., Evgrafov, O. V., Gulden, F. O., Pochareddy, S., et al. (2018). Integrative Functional Genomic Analysis of Human Brain Development and Neuropsychiatric Risks. *Science* 362. doi:10.1126/science.aat7615
- Long, J. E., Cobos, I., Potter, G. B., and Rubenstein, J. L. R. (2009). Dlx1&2 and Mash1 Transcription Factors Control MGE and CGE Patterning and Differentiation through Parallel and Overlapping Pathways. *Cereb. Cortex* 19, i96–i106. doi:10.1093/cercor/bhp045

- López-Bendito, G., Flames, N., Ma, L., Fouquet, C., Meglio, T. D., Chedotal, A., et al. (2007). Robo1 and Robo2 Cooperate to Control the Guidance of Major Axonal Tracts in the Mammalian Forebrain. *J. Neurosci.* 27, 3395–3407. doi:10.1523/JNEUROSCI.4605-06.2007
- Mafessoni, F., Grote, S., Filippo, C. d., Slon, V., Kolobova, K. A., Viola, B., et al. (2020). A High-Coverage Neandertal Genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci.* 117, 15132–15136. doi:10.1073/pnas.2004944117
- Martin, S. H., and Jiggins, C. D. (2017). Interpreting the Genomic Landscape of Introgression. *Curr. Opin. Genet. Develop.* 47, 69–74. doi:10.1016/j.gde.2017.08.007
- McCoy, R. C., Wakefield, J., and Akey, J. M. (2017). Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell* 168, 916–927. e12. doi:10.1016/j.cell.2017.01.038
- Mekki, Y., Guillemot, V., Lemaître, H., Carrión-Castillo, A., Forkel, S., Frouin, V., et al. (2022). The Genetic Architecture of Language Functional Connectivity. *NeuroImage* 249, 118795. doi:10.1016/j.neuroimage.2021.118795
- Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., et al. (2012). A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* 338, 222–226. doi:10.1126/science.1224344
- Mostajo-Radji, M. A., Schmitz, M. T., Montoya, S. T., and Pollen, A. A. (2020). Reverse Engineering Human Brain Evolution Using Organoid Models. *Brain Res.* 1729, 146582. doi:10.1016/j.brainres.2019.146582
- Muchnik, S. K., Lorente-Galdos, B., Santpere, G., and Sestan, N. (2019). Modeling the Evolution of Human Brain Development Using Organoids. *Cell* 179, 1250–1253. doi:10.1016/j.cell.2019.10.041
- Neubauer, S., Hublin, J.-J., and Gunz, P. (2018). The Evolution of Modern Human Brain Shape. *Sci. Adv.* 4, eaao5961. doi:10.1126/sciadv.aao5961
- Pääbo, S. (2014). The Human Condition—A Molecular Approach. *Cell* 157, 216–226. doi:10.1016/j.cell.2013.12.036
- Petr, M., Pääbo, S., Kelso, J., and Vernot, B. (2019). Limits of Long-Term Selection against Neandertal Introgression. *Proc. Natl. Acad. Sci.* 116, 1639–1644. doi:10.1073/pnas.1814338116
- Peyrégne, S., Boyle, M. J., Dannemann, M., and Prüfer, K. (2017). Detecting Ancient Positive Selection in Humans Using Extended Lineage Sorting. *Genome Res.* 27, 1563–1572. doi:10.1101/gr.219493.116
- Prüfer, K., Filippo, C. d., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., et al. (2017). A High-Coverage Neandertal Genome from Vindija Cave in Croatia. *Science* 358, 655–658. doi:10.1126/science.aao1887
- Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., et al. (2014). The Complete Genome Sequence of a Neandertal from the Altai Mountains. *Nature* 505, 43–49. doi:10.1038/nature12886
- R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rinker, D. C., Simonti, C. N., McArthur, E., Shaw, D., Hodges, E., and Capra, J. A. (2020). Neandertal Introgression Reintroduced Functional Ancestral Alleles Lost in Eurasian Populations. *Nat. Ecol. Evol.* 4, 1332–1341. doi:10.1038/s41559-020-1261-z
- Sadakata, T., Kakegawa, W., Shinoda, Y., Hosono, M., Katoh-Semba, R., Sekine, Y., et al. (2014). Axonal Localization of Ca²⁺-dependent Activator Protein for Secretion 2 Is Critical for Subcellular Locality of Brain-Derived Neurotrophic Factor and Neurotrophin-3 Release Affecting Proper Development of Postnatal Mouse Cerebellum. *PLOS ONE* 9, e99524. doi:10.1371/journal.pone.0099524
- Sadakata, T., Shinoda, Y., Ishizaki, Y., and Furuichi, T. (2017). Analysis of Gene Expression in Ca²⁺-dependent Activator Protein for Secretion 2 (Cadps2) Knockout Cerebellum Using GeneChip and KEGG Pathways. *Neurosci. Lett.* 639, 88–93. doi:10.1016/j.neulet.2016.12.068
- Sadakata, T., Washida, M., Iwayama, Y., Shoji, S., Sato, Y., Ohkura, T., et al. (2007). Autistic-like Phenotypes in Cadps2-Knockout Mice and Aberrant CADPS2 Splicing in Autistic Patients. *J. Clin. Invest.* 117, 931–943. doi:10.1172/JCI29031
- Sankararaman, S., Mallick, S., Patterson, N., and Reich, D. (2016). The Combined Landscape of Denisovan and Neandertal Ancestry in Present-Day Humans. *Curr. Biol.* 26, 1241–1247. doi:10.1016/j.cub.2016.03.037
- Skov, L., Coll Macià, M., Sveinbjörnsson, G., Mafessoni, F., Lucotte, E. A., Einarssdóttir, M. S., et al. (2020). The Nature of Neandertal Introgression Revealed by 27,566 Icelandic Genomes. *Nature* 582, 78–83. doi:10.1038/s41586-020-2225-9
- St Pourcain, B., Cents, R. A. M., Whitehouse, A. J. O., Haworth, C. M. A., Davis, O. S. P., O'Reilly, P. F., et al. (2014). Common Variation Near ROBO2 Is Associated with Expressive Vocabulary in Infancy. *Nat. Commun.* 5, 4831. doi:10.1038/ncomms5831
- Veller, C., Edelman, N. B., Muralidhar, P., and Nowak, M. A. (2021). *Recombination and Selection against Introgressed DNA*. bioRxiv doi:10.1101/846147
- Vernot, B., Tucci, S., Kelso, J., Schraiber, J. G., Wolf, A. B., Gittelman, R. M., et al. (2016). Excavating Neandertal and Denisovan DNA from the Genomes of Melanesian Individuals. *Science* 352, 235–239. doi:10.1126/science.aad9416
- Vilgalys, T. P., Fogel, A. S., Mututua, R. S., Warutere, J. K., Siodi, L., Anderson, J. A., et al. (2021). *Selection against Admixture and Gene Regulatory Divergence in a Long-Term Primate Field Study*. bioRxiv. doi:10.1101/2021.08.19.456711
- Wang, R., Chen, C.-C., Hara, E., Rivas, M. V., Rouillac, P. L., Howard, J. T., et al. (2015). Convergent Differential Regulation of SLIT-ROBO Axon Guidance Genes in the Brains of Vocal Learners. *J. Comp. Neurol.* 523, 892–906. doi:10.1002/cne.23719
- Wang, S., Rohwer, S., Zwaan, D. R. d., Toews, D. P. L., Lovette, I. J., Mackenzie, J., et al. (2020). Selection on a Small Genomic Region Underpins Differentiation in Multiple Color Traits between Two Warbler Species. *Evol. Lett.* 4, 502–515. doi:10.1002/evl3.198
- Weiss, C. V., Harshman, L., Inoue, F., Fraser, H. B., Petrov, D. A., Ahituv, N., et al. (2021). The Cis-Regulatory Effects of Modern Human-specific Variants. *eLife* 10, e63713. doi:10.7554/eLife.63713
- Wolf, A. B., and Akey, J. M. (2018). Outstanding Questions in the Study of Archaic Hominin Admixture. *PLOS Genet.* 14, e1007349. doi:10.1371/journal.pgen.1007349

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Buisan, Moriano, Andirkó and Boeckx. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Supplementary Material

for

“A brain region-specific expression profile for genes within large introgression deserts and under positive selection in *Homo sapiens*”

Raül Buisan^{1,*}, Juan Moriano^{1,2,*}, Alejandro Andirkó^{1,2}, and Cedric Boeckx^{1,2,3,**}

¹Universitat de Barcelona

²Universitat de Barcelona Institute of Complex Systems

³Catalan Institute for Research and Advanced Studies (ICREA)

*Contributed equally

**Correspondence: cedric.boeckx@ub.edu

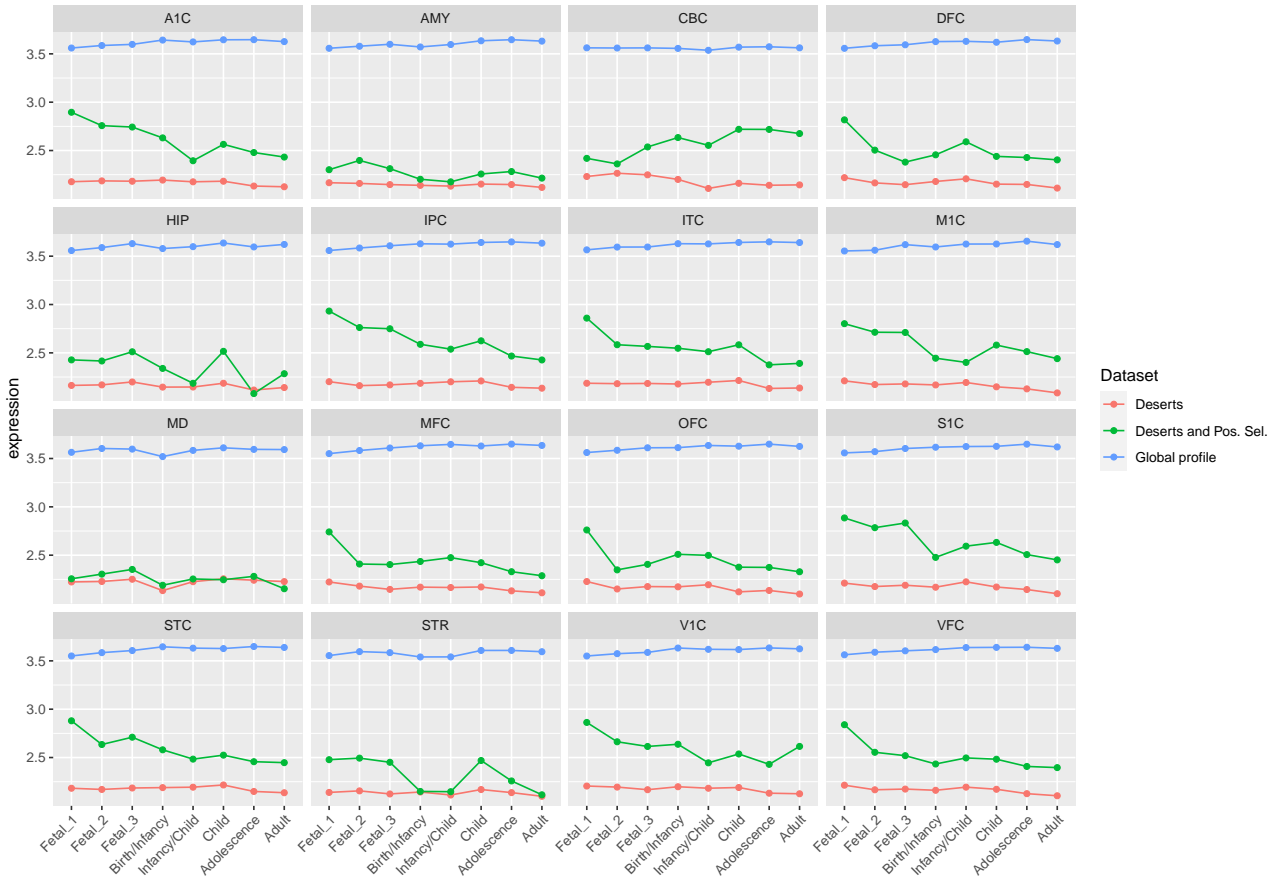


Figure 1: **Median expression profile of genes within large deserts, deserts/positively-selected regions and the global dataset, across structures and stages.** Genes within deserts/positively-selected regions have higher expression in comparison to the broader subset of genes within deserts, with a peak of expression at prenatal stages in neocortical areas. This peak is not observed in non-neocortical areas and, in the case of the cerebellar cortex, the genes (within deserts/positively-selected regions) show increasing expression at postnatal stages.

As explained in *Methods* section, the log-transformed expression values for the global profile was filtered by setting a threshold of median expression value > 2 , as in the original publication [1], since the inclusion of too many zeros makes the determination of trajectories unreliable. However, for genes from our regions of interest, no threshold was set in order to detect outliers of potential relevance.

A1C, primary auditory cortex; AMY, amygdala; CBC, cerebellar cortex; DFC, dorsolateral prefrontal cortex; HIP, hippocampus; IPC, posterior inferior parietal cortex; ITC, inferior temporal cortex; M1C, primary motor cortex; MD, mediodorsal nucleus of thalamus; MFC, medial prefrontal cortex; OFC, orbital prefrontal cortex; S1C, primary somatosensory cortex; STC, superior temporal cortex; STR, striatum; V1C, primary visual cortex; VFC, ventrolateral prefrontal cortex.

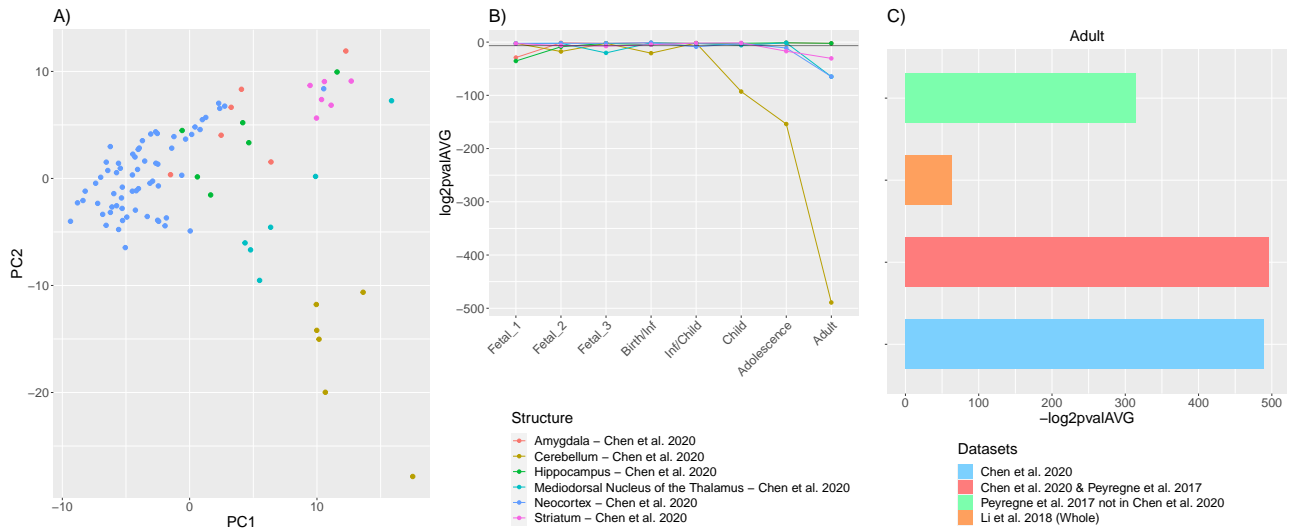


Figure 2: The cerebellum's transcriptomic profile significantly diverges at postnatal stages. For genes within large deserts of introgression, the cerebellum exhibits the most significant differences when evaluating pairwise distances between structures based on the statistically significant principal components. A) Distribution of structures at the adult stage using the first two principal components. B) P-values (log₂-transformed) obtained from pairwise comparisons among structures at each developmental stage (Wilcoxon rank sum test with Bonferroni correction). C) Contribution of genes within large deserts of introgression [7], deserts of introgression under putative positive selection [7, 35], regions under putative selection [35] not within large deserts, and the raw dataset used in this study [1] to the observed divergence at adult stage for the cerebellum, with the greatest value for genes within large deserts.

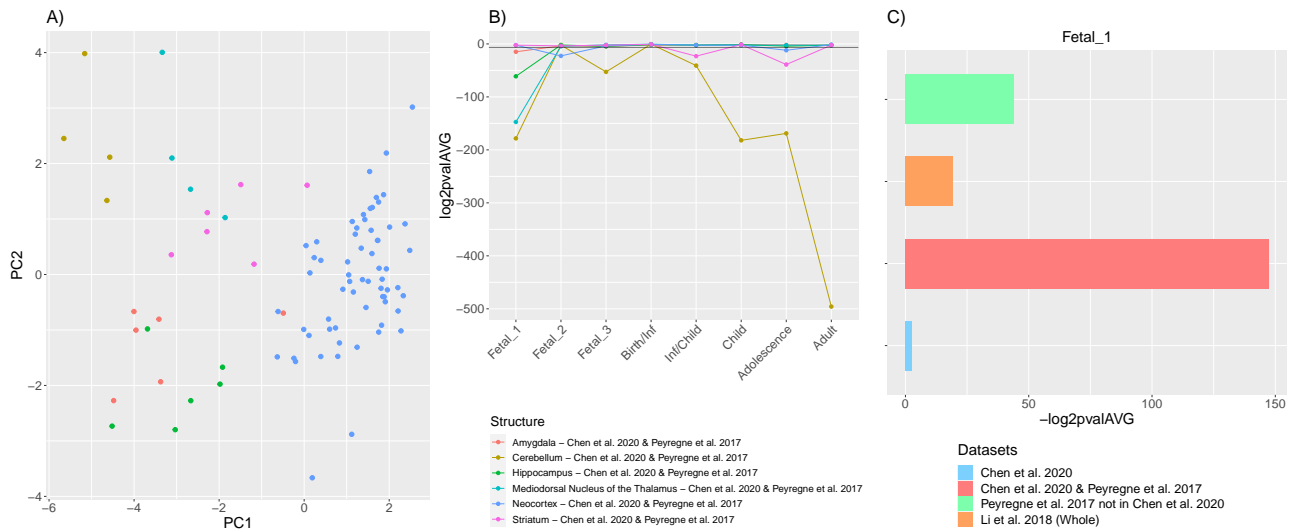


Figure 3: The transcriptomic profile of the mediodorsal nucleus of the thalamus significantly diverges at fetal stage 1 for genes within deserts under positive selection. A) Distribution of structures at the fetal stage 1 using the first two principal components. B) P-values (log₂-transformed) obtained from pairwise comparisons among structures at each developmental stage (Wilcoxon rank sum test with Bonferroni correction). C) Contribution of genes within deserts of introgression [7], deserts of introgression under putative positive selection [7, 35], regions under putative selection [35] not within large deserts, and the raw dataset used in this study [1], to the observed divergence at adolescence for the striatum. The greatest value is found for genes within deserts under putative positive selection.

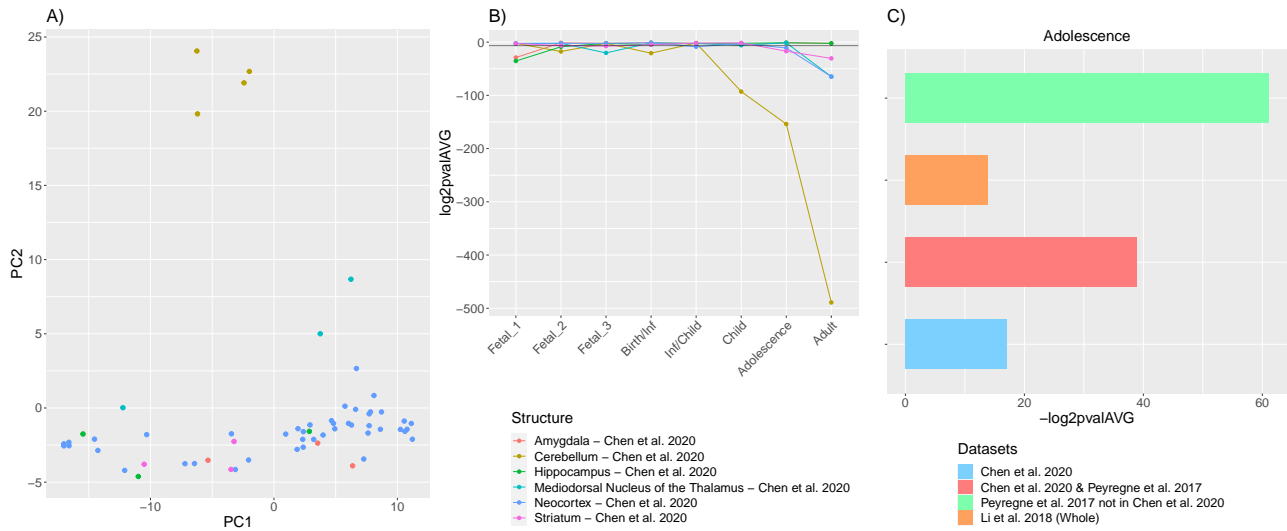


Figure 4: **The striatum's transcriptomic profile significantly diverges at adolescence for genes within large deserts.** When evaluating the transcriptome of genes within large deserts of introgression, the striatum reports significant differences postnatally at adolescence and adult stages. A) Distribution of structures at the adolescence using the first two principal components. B) P-values (log₂-transformed) obtained from pairwise comparisons among structures at each developmental stage (Wilcoxon rank sum test with Bonferroni correction). C) Contribution of genes within large deserts of introgression [7], deserts of introgression under putative positive selection [7, 35], regions under putative selection [35] not within large deserts, and the raw dataset used in this study [1], to the observed divergence at adolescence for the striatum. The greatest value is found for genes under putative positive selection not within deserts of introgression, an effect also shown in Supplementary Figure 9.

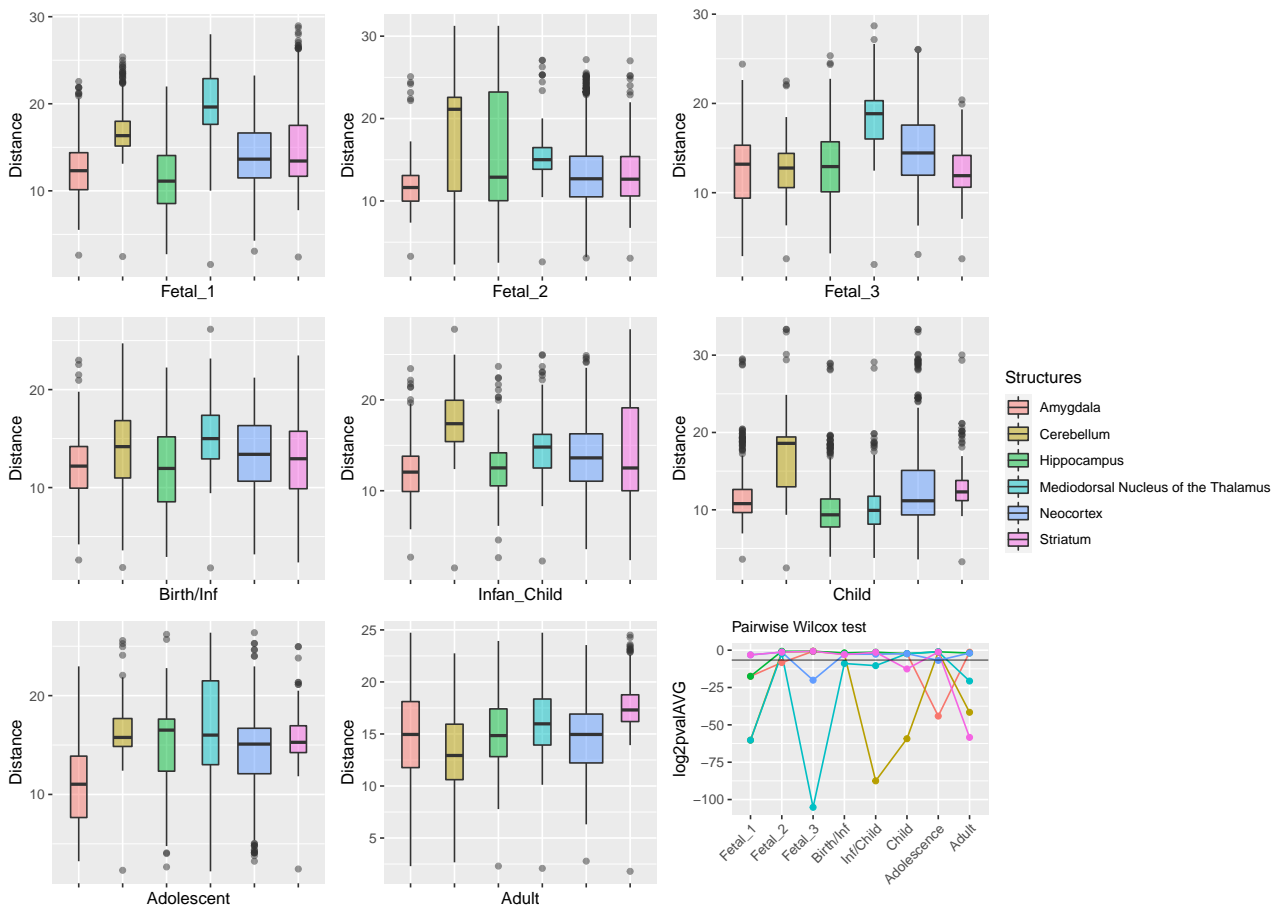


Figure 5: **Evaluation of transcriptomic divergence considering genes within deserts for chromosome 1.** These and the following Supplementary Figures 6, 7 and 8 show the expression of genes within large deserts for each of the chromosomes harboring deserts of introgression, as reported in [7]. For genes within chromosome 1 desert ($n = 132$), the sharpest difference is observed for the mediodorsal nucleus of the thalamus at fetal stage 3. Other structures that show marked statistically significant differences are the cerebellum (fetal stage 1, infancy/childhood, childhood, adulthood), the amygdala (adolescence) and the striatum (adulthood). All p-values can be found at the online repository <https://github.com/jjaa-mp/desertsHomo>.

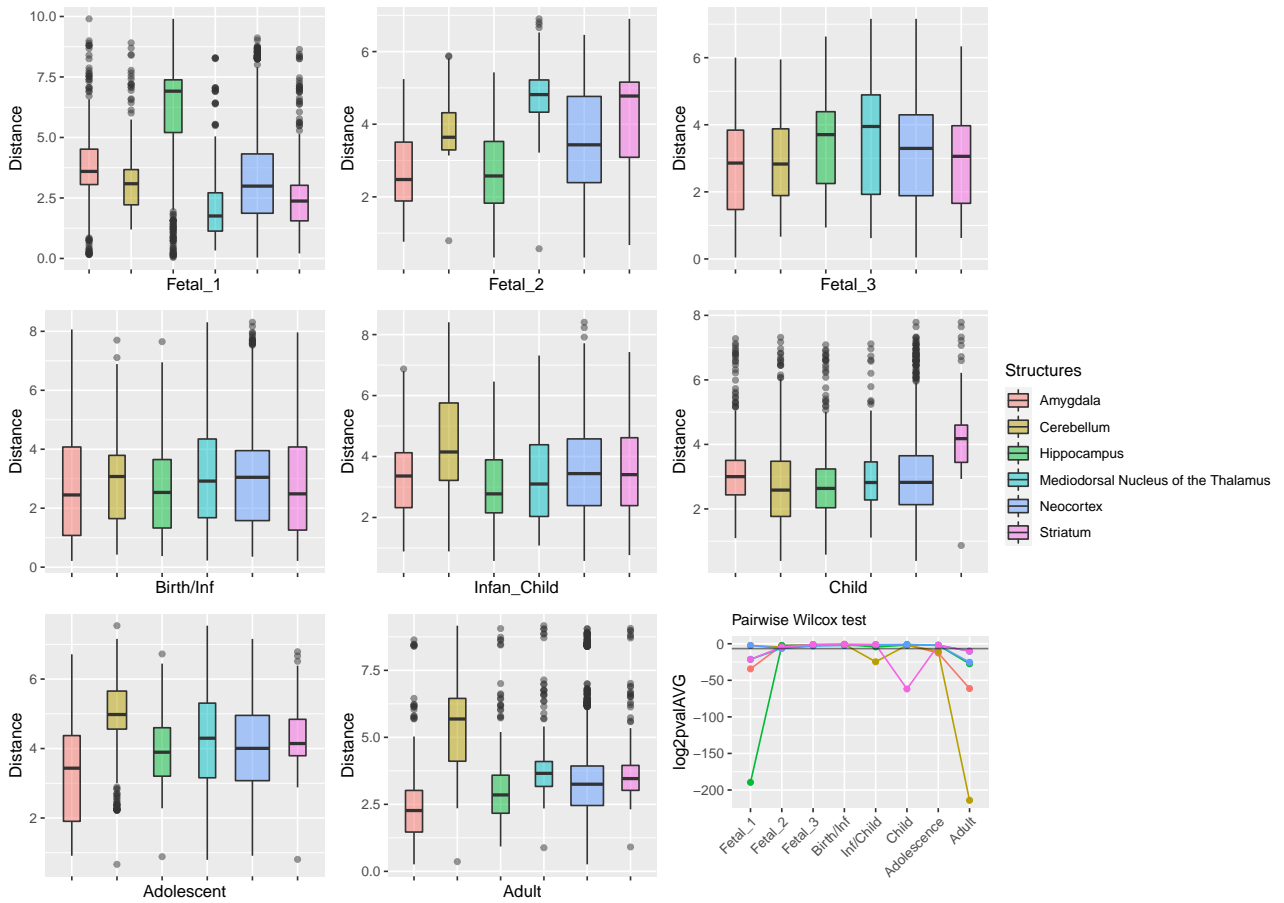


Figure 6: **Evaluation of transcriptomic divergence considering genes within deserts for chromosome 3.** For genes within chromosome 3 desert ($n = 15$), the hippocampus at fetal stage 1 and the cerebellum at adulthood are the most transcriptomically divergent structures, followed by the striatum (childhood) and the amygdala (adulthood). All p-values can be found at the online repository <https://github.com/jjaa-mp/desertsHomo>

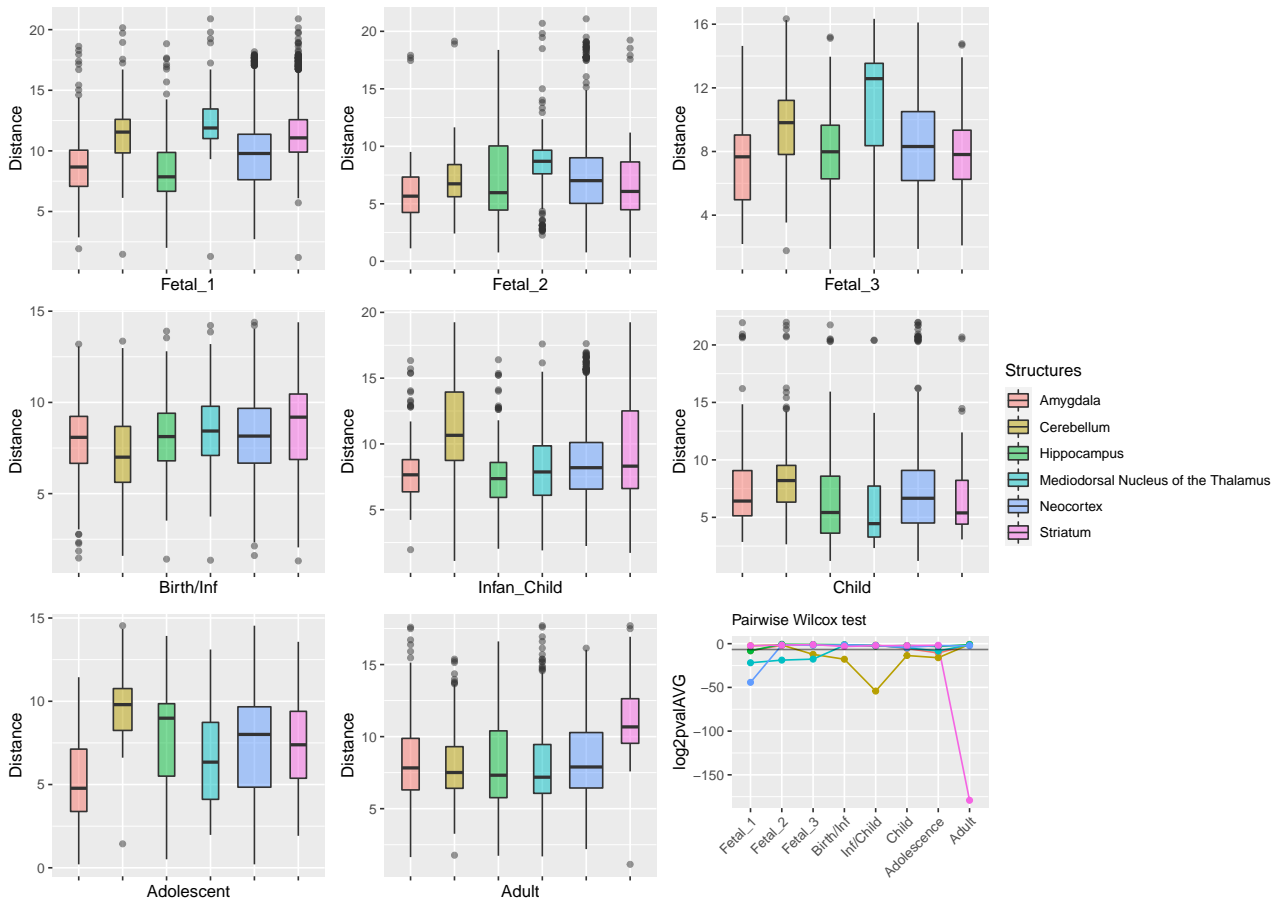


Figure 7: **Evaluation of transcriptomic divergence considering genes within deserts for chromosome 7.** In chromosome 7 desert ($n = 62$), the striatum is found as the most significantly different transcriptome at adulthood. The mediodorsal nucleus of the thalamus (fetal stages) and the cerebellum (from fetal stage 3 and birth to adolescence) are also found as statistically divergent structures. All p-values can be found at the online repository <https://github.com/jjaa-mp/desertsHomo>

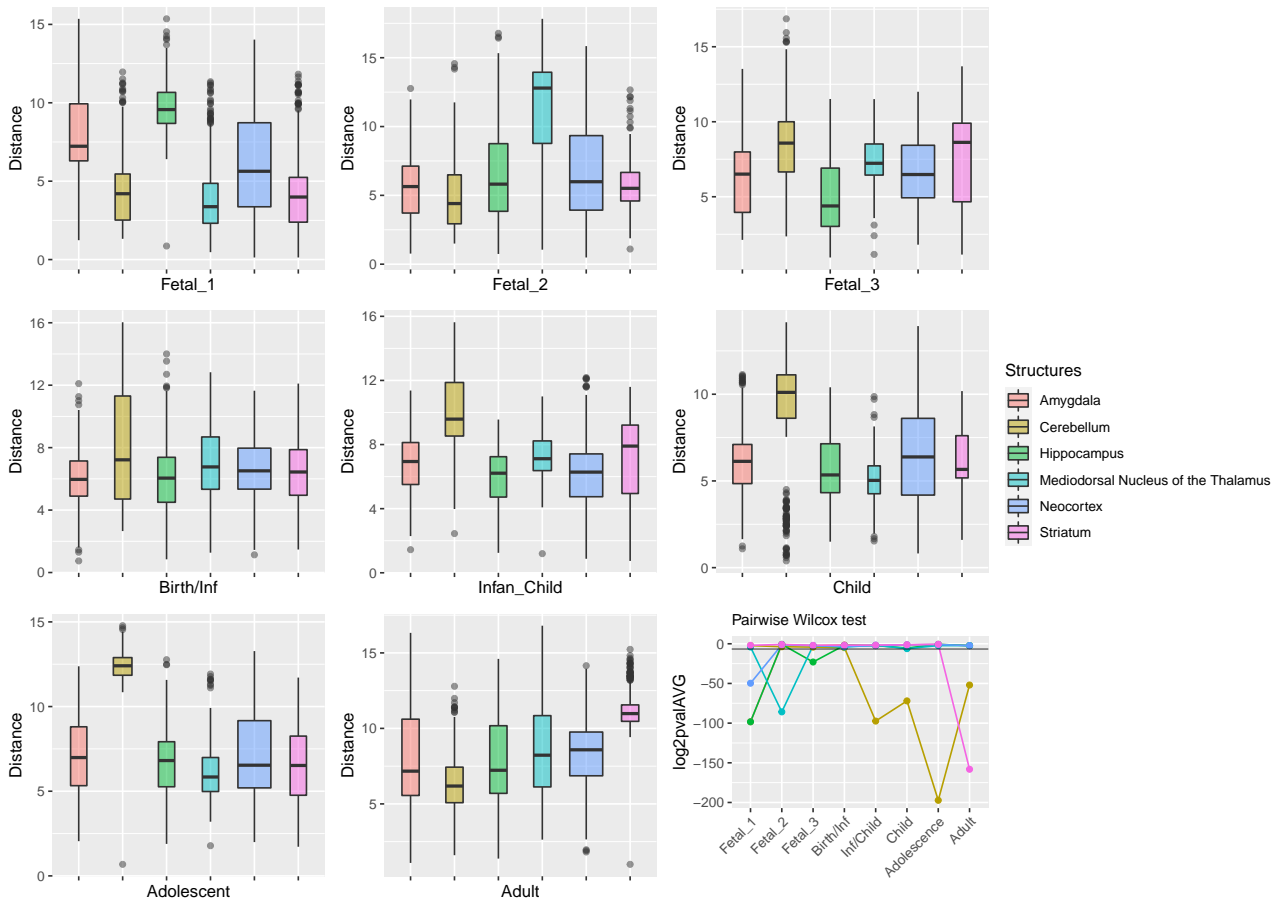


Figure 8: **Evaluation of transcriptomic divergence considering genes within deserts for chromosome 8.** Regarding genes within chromosome 8 desert ($n = 46$), the cerebellar profile, at successive stages postnatally, stand out as the most divergent transcriptome. The striatum at adulthood and several structures prenatally (neocortex, hippocampus, mediodorsal nucleus of the thalamus) were also statistically different when considering their transcriptomic profiles. All p-values can be found at the online repository <https://github.com/jjaa-mp/desertsHomo>

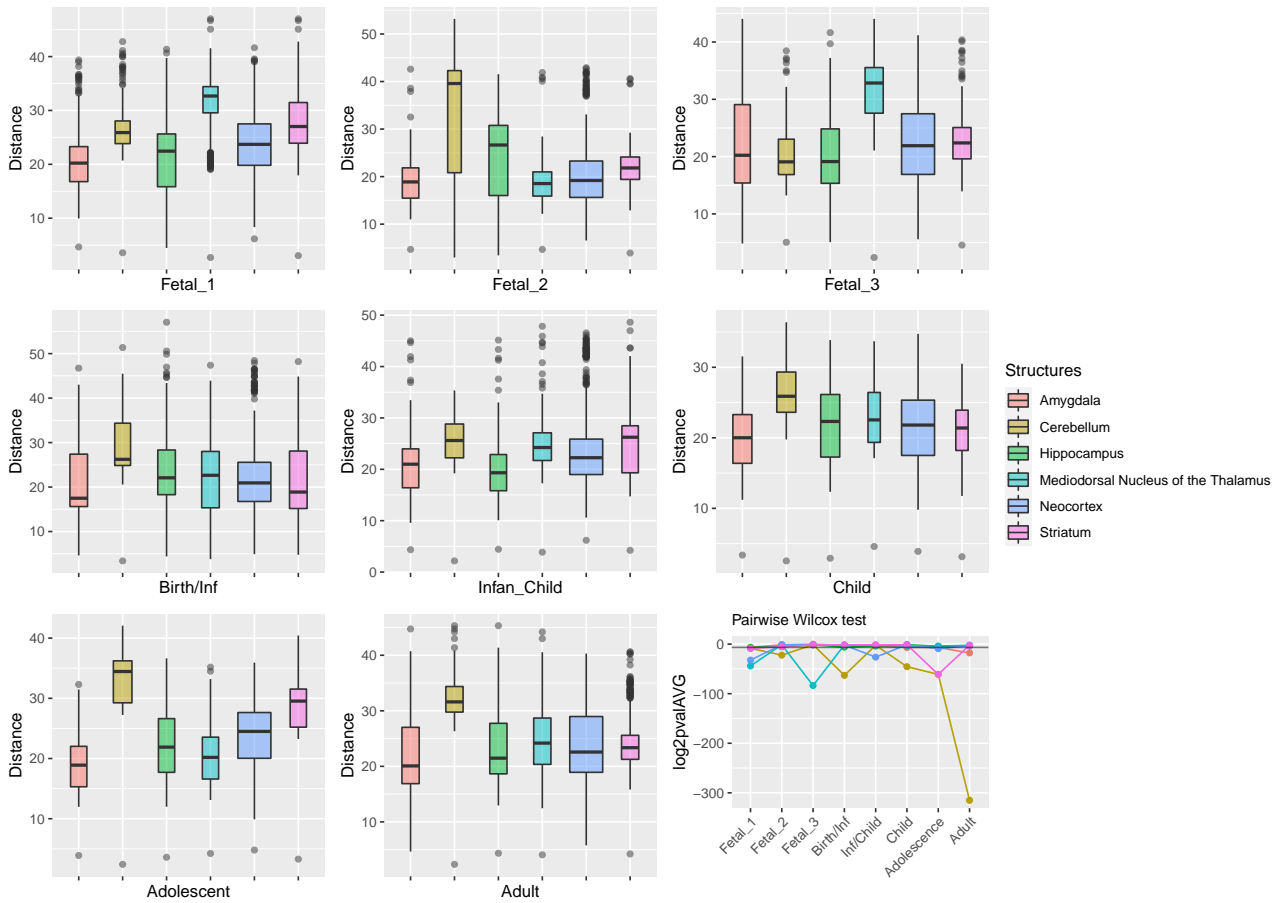


Figure 9: **Evaluation of transcriptomic divergence considering genes under positive selection not within large deserts of introgression.** Similarly to what is found in Supplementary Figure4 and 5, the cerebellum stand out postnatally (birth $p = 1.07 \times 10^{-19}$, childhood $p = 1.95 \times 10^{-14}$, adolescence $p = 4.11 \times 10^{-19}$ and adulthood $p = 1.42 \times 10^{-95}$). Other significant results are found for the thalamus (fetal stages 1 and 3), striatum (adolescence) or neocortex (fetal stage 1).

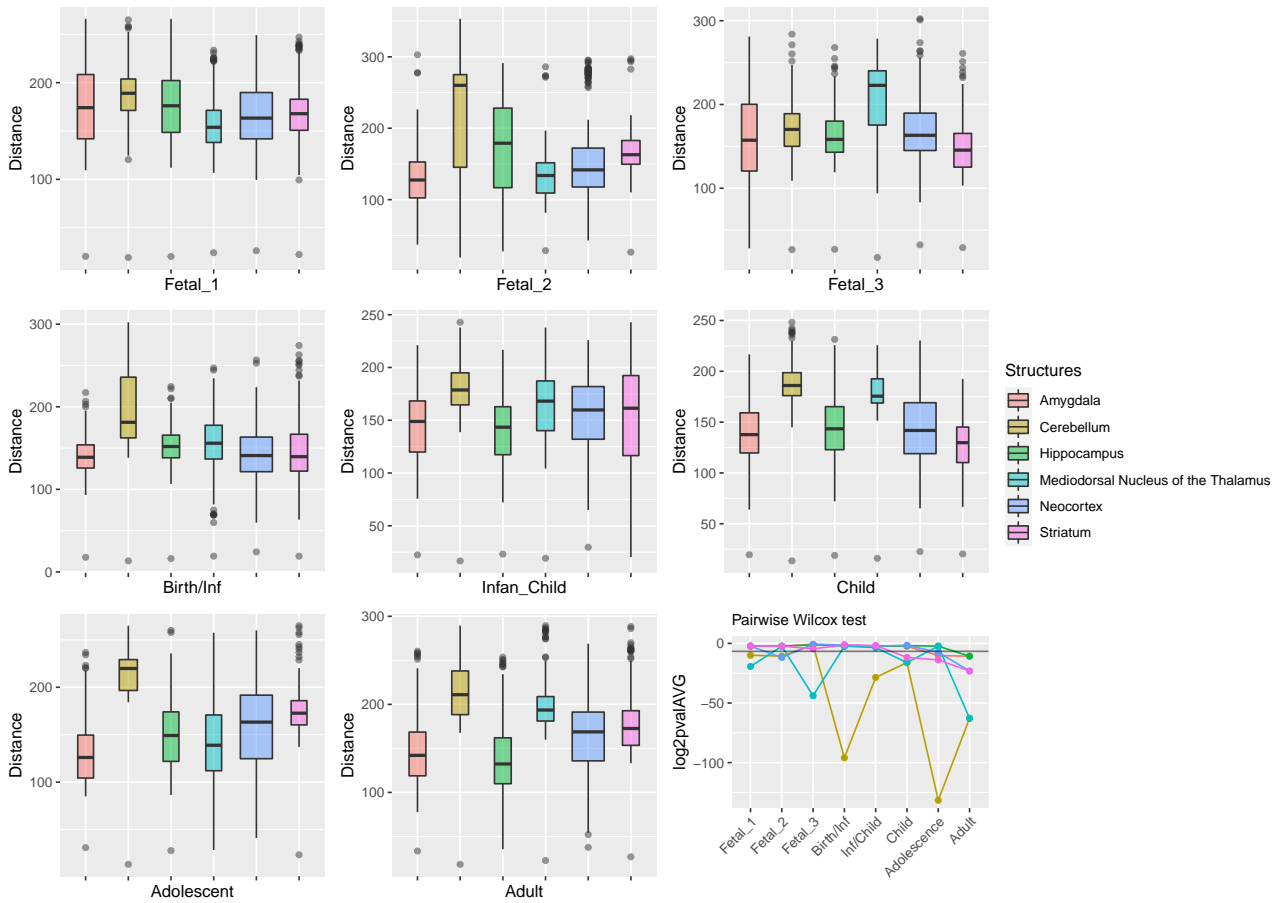


Figure 10: **Global profile of genes across stages and structures.** We processed the mRNA-seq data from [1] filtering out genes with a median across stages and structures less than 2, resulting in a total of 9358 genes. As described in section 4 of the main text, log-transformed, RPKM normalized counts were used to calculate pairwise Euclidean distances using statistically significant principal components for each stage. The most significant differences (pairwise Wilcoxon test with Bonferroni correction) are found for the cerebellum at postnatal stages (birth $p = 1.29 \times 10^{-29}$; infancy $p = 2.53 \times 10^{-9}$; adolescence $p = 2.44 \times 10^{-30}$; adulthood $p = 1.21 \times 10^{-19}$), and for the thalamus at fetal stage 3 ($p = 6.12 \times 10^{-14}$) and adulthood ($p = 1.21 \times 10^{-19}$). All p-values can be found in Supplementary Material. The horizontal black line in the line plot denotes $p = 0.01$.

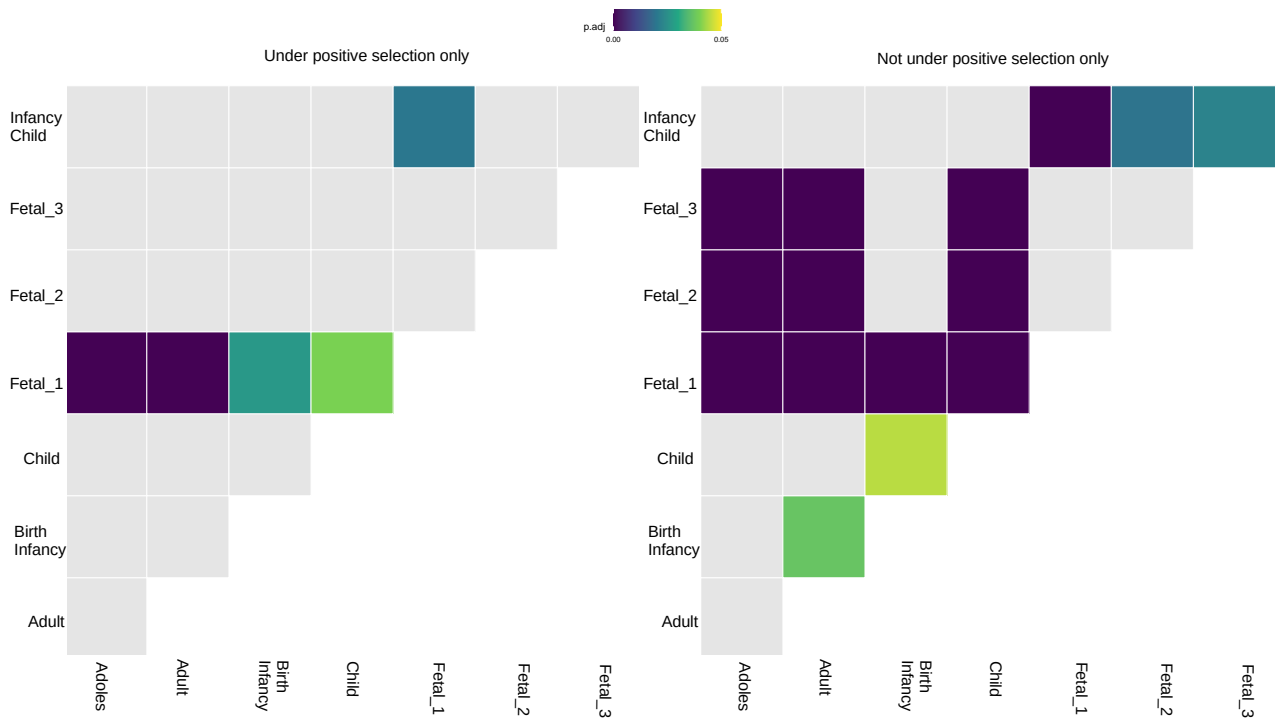


Figure 11: **Visualization of a pairwise post-hoc Tukey test** comparing the mean expression of genes in deserts of introgression under the effect of positive selection (left) and in genes not affected by positive selection (right) in each of the developmental stages included in [1].

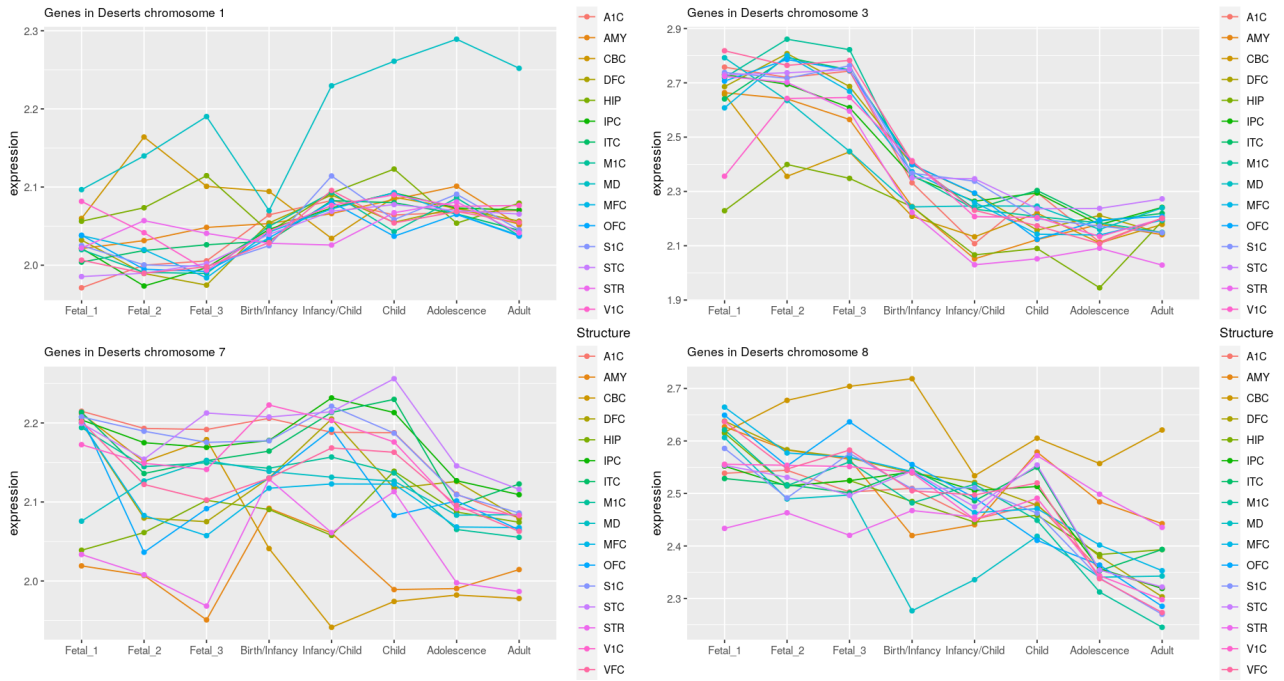


Figure 12: **Decomposition of median expression profile of genes within large deserts per chromosome.** Genes within chromosomes 3 ($n = 15$) and 8 ($n = 46$) report higher mean expression values than those within chromosomes 1 ($n = 132$) and 7 ($n = 62$), markedly at prenatal stages (as discussed in *Methods* section 4, we kept for these analyses possible outliers, including therefore those with low expression values, which are more abundant in datasets with a higher n . This reinforces the strategies to characterize gene-specific expression dynamics undertaken in this study). Some structures show specific profiles, as the for mediadorsal nucleus of thalamus genes within chromosome 1, or the cerebellum in chromosomes 7 and 8. This view complements Supplementary Figure 13, dedicated to genes under positive selection within introgression deserts.

A1C, primary auditory cortex; AMY, amygdala; CBC, cerebellar cortex; DFC, dorsolateral prefrontal cortex; HIP, hippocampus; IPC, posterior inferior parietal cortex; ITC, inferior temporal cortex; M1C, primary motor cortex; MD, mediadorsal nucleus of thalamus; MFC, medial prefrontal cortex; OFC, orbital prefrontal cortex; S1C, primary somatosensory cortex; STC, superior temporal cortex; STR, striatum; V1C, primary visual cortex; VFC, ventrolateral prefrontal cortex.

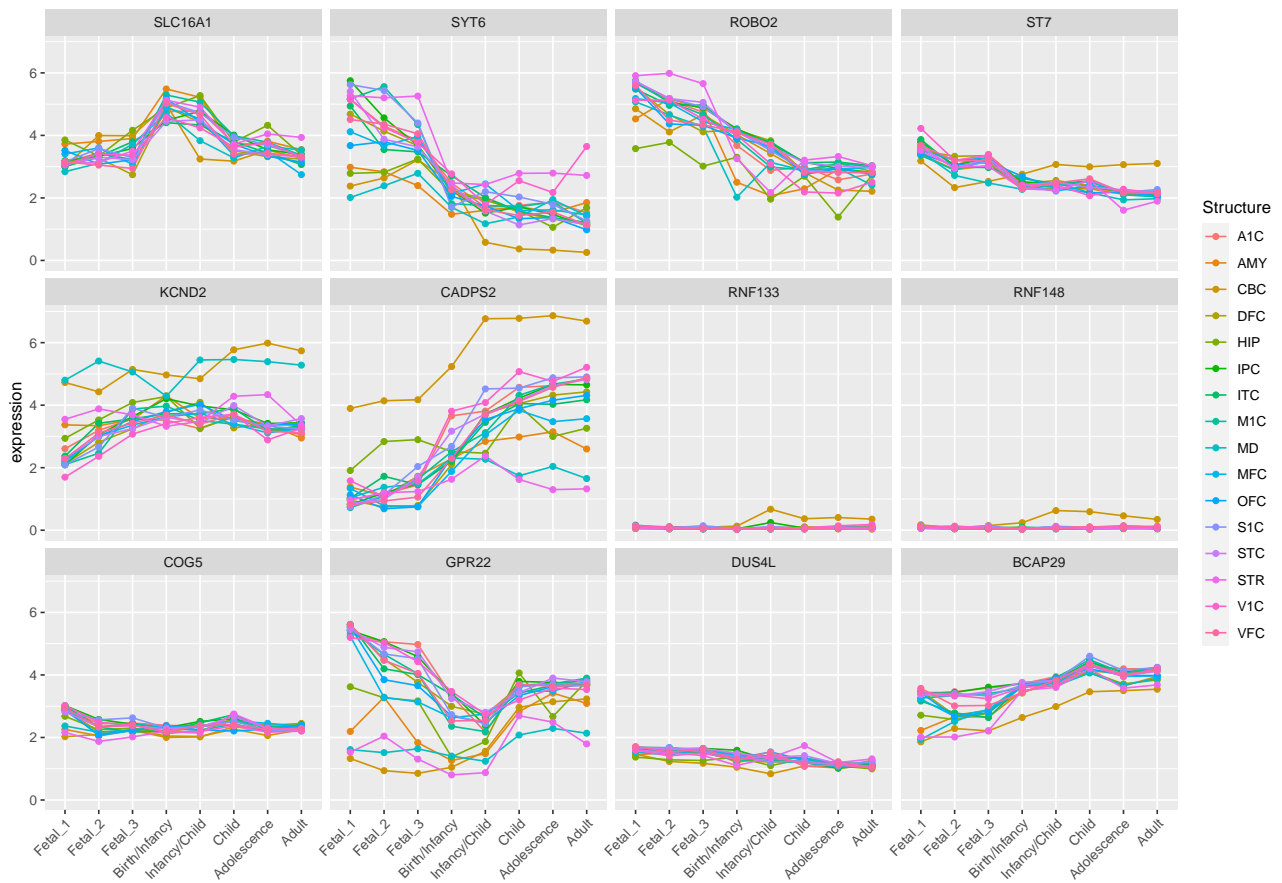


Figure 13: **Expression profile of genes within large deserts and positively-selected overlapping regions.** Twelve protein-coding genes were found at the intersection between large deserts of introgression and putative positively-selected regions. The expression of *CADPS2* and *KCND2* largely recapitulate the unique profile shown before (Supplementary Figure S1) for the cerebellar cortex: Increasing expression from prenatal to postnatal stages, reaching the highest median expression value from childhood to adulthood in comparison to all other structures.

A1C, primary auditory cortex; AMY, amygdala; CBC, cerebellar cortex; DFC, dorsolateral prefrontal cortex; HIP, hippocampus; IPC, posterior inferior parietal cortex; ITC, inferior temporal cortex; M1C, primary motor cortex; MD, mediodorsal nucleus of thalamus; MFC, medial prefrontal cortex; OFC, orbital prefrontal cortex; S1C, primary somatosensory cortex; STC, superior temporal cortex; STR, striatum; V1C, primary visual cortex; VFC, ventrolateral prefrontal cortex.

References

- [1] M. Li, *et al.*, “Integrative functional genomic analysis of human brain development and neuropsychiatric risks,” *Science*, vol. 362, Dec. 2018.
- [2] S. Grote, *et al.*, “ABAEnrichment: an R package to test for gene set expression enrichment in the adult and developing human brain,” *Bioinformatics*, vol. 32, pp. 3201–3203, Oct. 2016.
- [3] M. Kuhlwilm *et al.*, “A catalog of single nucleotide changes distinguishing modern humans from archaic hominins,” *Scientific Reports*, vol. 9, p. 8463, June 2019.
- [4] A. Andirkó *et al.*, “Modern human alleles differentially regulate gene expression across brain regions: implications for brain evolution,” *bioRxiv*, p. 771816, Nov. 2020. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Appendices

Appendix A

Temporal mapping of derived high-frequency gene variants supports the mosaic nature of the evolution of *Homo sapiens*

Published as:

Andirkó, A., Moriano, J., Vitriolo, A., Kuhlwilm, M., Testa, G., & Boeckx, C. 2022. Fine-grained temporal mapping of derived high-frequency variants supports the mosaic nature of the evolution of *Homo sapiens*. *Scientific Reports*

doi:[10.1038/s41598-022-13589-0](https://doi.org/10.1038/s41598-022-13589-0)



OPEN

Temporal mapping of derived high-frequency gene variants supports the mosaic nature of the evolution of *Homo sapiens*

Alejandro Andirkó^{1,2,9}, Juan Moriano^{1,2,9}, Alessandro Vitriolo^{3,4,5}, Martin Kuhlilm^{6,7}, Giusepe Testa^{3,4,5} & Cedric Boeckx^{1,2,8}✉

Large-scale estimations of the time of emergence of variants are essential to examine hypotheses concerning human evolution with precision. Using an open repository of genetic variant age estimations, we offer here a temporal evaluation of various evolutionarily relevant datasets, such as *Homo sapiens*-specific variants, high-frequency variants found in genetic windows under positive selection, introgressed variants from extinct human species, as well as putative regulatory variants specific to various brain regions. We find a recurrent bimodal distribution of high-frequency variants, but also evidence for specific enrichments of gene categories in distinct time windows, pointing to different periods of phenotypic changes, resulting in a mosaic. With a temporal classification of genetic mutations in hand, we then applied a machine learning tool to predict what genes have changed more in certain time windows, and which tissues these genes may have impacted more. Overall, we provide a fine-grained temporal mapping of derived variants in *Homo sapiens* that helps to illuminate the intricate evolutionary history of our species.

The past decade has seen a significant shift in our understanding of the evolution of our lineage. We now recognize that anatomical features used as diagnostic for our species (globular neurocranium, small, retracted face, presence of a chin, narrow trunk, to cite only a few of the most salient traits associated with “anatomical modernity”) did not emerge as a package, from a single geographical location, but rather emerged gradually, in a mosaic-like fashion across the entire African continent and quite possibly beyond^{1–3}. Likewise, behavioral characteristics once thought to be exclusive of *Homo sapiens* (funerary rituals, parietal art, ‘symbolic’ artefacts, etc.) have recently been attested in some form in closely related (extinct) clades, casting doubt on a simple definition of ‘cognitive/behavioral’ modernity⁴. We have also come to appreciate the extent of repeated (multidirectional) gene flow between *Homo sapiens* and Neanderthals and Denisovans, raising interesting questions about speciation^{5–8}. Last, but not least, it is now well established that our species has a long history. Robust genetic analyses⁹ indicate a divergence time between us and other hominins for whom genomes are available of roughly 700kya, leaving perhaps as many as 500ky between then and the earliest fossils displaying a near-complete suite of modern traits (Omo Kibish 1, Herto 1 and 2)¹⁰. Such a long period of time is likely to contain enough opportunities for multiple rounds of evolutionary modifications. Taken together, these findings render completely implausible simplistic narratives about the ‘modern human condition’ that seek to identify a specific geographical location or genetic mutation that would ‘define’ us¹¹.

Genomic analysis of ancient human remains in Africa reveal deep population splits and complex admixture patterns among populations^{12–14}. At the same time, reanalysis of fossils in Africa¹⁵ points to the extended presence of multiple hominins on this continent, together with real possibilities of admixture^{16,17}. Lastly, our deeper understanding of other hominins points to derived characteristics in these lineages that make some of our species’ traits more ancestral (less ‘modern’) than previously believed¹⁸.

¹Universitat de Barcelona, Barcelona, Spain. ²Universitat de Barcelona Institute of Complex Systems (UBICS), Barcelona, Spain. ³University of Milan, Milan, Italy. ⁴European Institute of Oncology (IEO), Milan, Italy. ⁵Human Technopole, Milan, Italy. ⁶University of Vienna, Vienna, Austria. ⁷Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Vienna, Austria. ⁸Catalan Institute for Research and Advanced Studies (ICREA), Catalonia, Spain. ⁹These authors contributed equally: Alejandro Andirkó and Juan Moriano. ✉email: cedric.boeckx@ub.edu

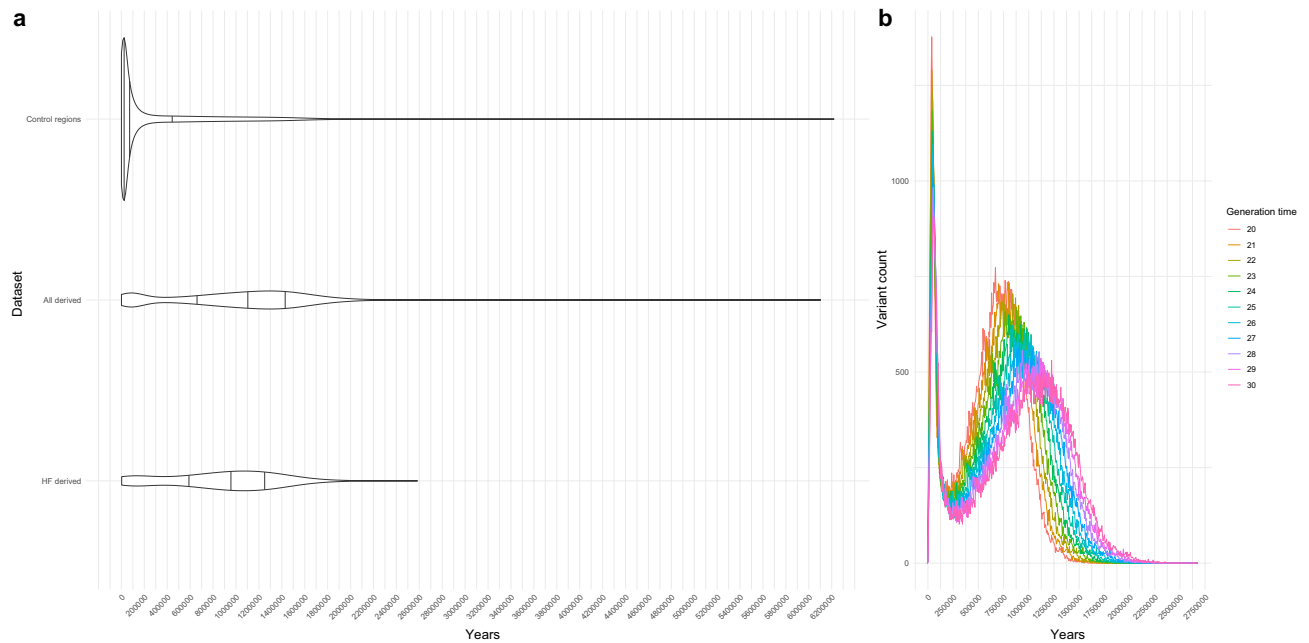


Figure 1. (a) Density of distribution of derived *Homo sapiens* alleles over time in an aggregated control set ($n = 1000$) of random variants across the genome and two sets of derived ones: all derived variants, and those found at high-frequency. Horizontal lines mark distribution quantiles 0.25, 0.5 and 0.75. (b) Line plot showing the bimodal distribution of high-frequency variants using different generation times (in the text, we used 29 years, following⁶²).

In the context of this significant rewriting of our deep history, we decided to explore the temporal structure of an extended catalog of single nucleotide changes found at high frequency ($HF \geq 90\%$) across major modern populations we previously generated on the basis of 3 high-coverage “archaic” genomes¹⁹, that is, Neanderthal/Denisovan individuals, used as outgroups. This catalog aims to offer a richer picture of molecular events setting us apart from our closest extinct relatives. In order to probe the temporal nature of this data, we took advantage of the Genealogical Estimation of Variant Age (GEVA) tool²⁰. GEVA is a coalescence-based method that provides age estimates for over 45 million human variants. GEVA is non-parametric, making no assumptions about demographic history, tree shapes, or selection (for additional details on GEVA, see “Methods”). Our overall objective here is to use the temporal resolution afforded by GEVA to estimate the age of emergence of polymorphic sites, and gain further insights into the complex evolutionary trajectory of our species.

Our analysis reveals a bimodal temporal distribution of modern human derived high-frequency variants and provides insights into milestones of *Homo sapiens* evolution through the investigation of the molecular correlates and the predicted impact of variants across evolutionary-relevant periods. Our chronological atlas allows us to provide a time window estimate of introgression events and evaluate the age of variants associated with signals of positive selection, tissue-specific changes, and specifically an estimate of the age of emergence of (enhancer) regulatory variants associated with different brain regions. Our enrichment analysis uncovers GO-terms unique to specific temporal windows, such as facial and behavioral-related terms for a period (between 300 and 500 k years) preceding the dating of human fossils like that of Jebel Irhoud. Our machine learning-based analyses predicting differential gene expression regulation of mapped variants (through²¹) reveals a trend towards downregulation in brain-related tissues and allowed us to identify variant-associated genes whose differential regulation may specifically affect brain structures such as the cerebellum.

Results

The distribution of derived alleles over time follows a bimodal distribution (Fig. 1a,b; see also Fig. S2 for a more elaborated version), with a global maximum around 40 kya (for complete allele counts, see “Methods”). The two modes of the distribution of HF variants likely correspond to two periods of significance in the evolutionary history of *Homo sapiens*. The more recent peak of HF variants arguably corresponds to the period of population dispersal and replacement following the last major out of Africa event^{22,23}, while the older distribution contains the period associated with the divergence between *Homo sapiens* and other *Homo* species^{9,24}.

In order to divide the data into smaller temporal clusters for downstream analysis we considered a k -means clustering analysis (at $k = 3$ and $k = 4$, Fig. S1). This clustering method yields a division clear enough to distinguish between “early” and “late” *Homo sapiens* “specimens”¹⁰, with a protracted period overlapping with the split with other *Homo* species. (The availability of ancient DNA from other hominins would yield a better resolution of that period.) However, we reasoned that such a k -means division is not precise enough to represent key milestones used to test specific time-sensitive hypotheses. For this reason, we adopted a literature-based approach, establishing different cutoffs adapted to the need of each analysis below. Our basic division consisted of three

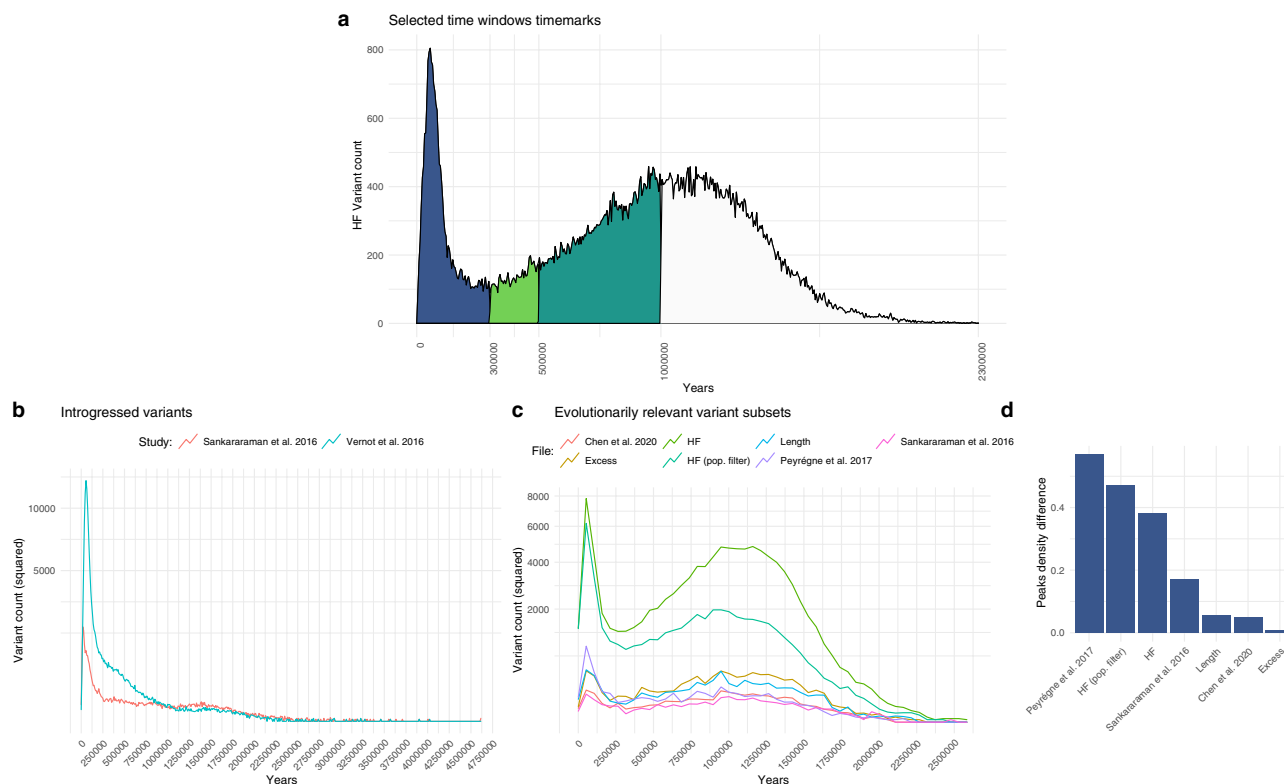


Figure 2. (a) Selected temporal windows used in our study to further interrogate the nature and distribution of HF variants. (b) Distribution of introgressed alleles over time, as identified by^{27,30}. (c) Plots of HF variants in datasets relevant to human evolution, including regions under positive selection²⁹, regions depleted of archaic introgression^{27,28} and genes showing an excess of HF variants ('excess' and 'length'¹⁹). Variant counts in (a,c,d) are squared to aid visualization. (d) Kernel density difference between the highest point in the distributions of (d) (leftmost peak) and the second, older highest density peak, normalized, in percentage units.

periods (see Fig. 2a): a recent period from the present to 300 thousand years ago (kya), the local minimum, roughly corresponding to the period considered until recently to mark the emergence of *Homo sapiens*¹²; a later period from 300 to 500 kya, the period right before the dating of fossils associated with earlier members of our species such as the Jebel Irhoud fossil²⁵ and, incidentally, the critical juncture between the first and second temporal windows when comparing the two *k*-means clustering analyses we performed (Fig. S1); and a third, older period, from 500 kya to 1 million years ago, corresponding to the time of the most recent common ancestor with the Neanderthal and Denisovan lineages^{24,26}.

We note that the distribution goes as far back as 2.5 million years ago (see Fig. 1a) in the case of HF variants, and even further back in the case of the derived variants with no HF cutoff. This could be due to our temporal prediction model choice (GEVA clock model, of which GEVA offers three options, as detailed in "Methods"), as changes over time in human recombination rates might affect the timing of older variants²⁰, or to the fact that we do not have genomes for older *Homo* species. Some of these very old variants may have been inherited from them and lost further down Neanderthal/Denisovan lineages.

Variant subset distributions. In an attempt to see if specific subsets of variants clustered in different ways over the inferred time axis, we selected a series of evolutionary relevant sets of data publicly available, such as genome regions depleted of "archaic" introgression (so-called 'deserts of introgression')^{27,28}, and regions under putative positive selection²⁹, and mapped the HF variants from¹⁹ falling within those regions. We also examined genes that accumulate more HF variants than expected given their length and in comparison to the number of mutations these genes accumulate on the Neanderthal/Denisovan lineages ('length' and 'excess' lists from¹⁹—see "Methods"). Finally, we also examined the temporal distribution of introgressed alleles^{27,30}. A bimodal distribution is clearly visible in all the subsets except the introgression datasets (Fig. 2b). Introgressed variants peak locally in the more recent period (0–100 kya). The distribution roughly fades after 250 kya, in consonance with the possible timing of introgression events^{6,16,28,31}. As a case study, we focused on those introgressed variants associated with phenotypes highlighted in Table 1 of³². As shown in Fig. S3, half of the variants cluster around the highest peak, but other variants may have been introduced in earlier instances of gene flow. We caution, though, that multiple (likely) factors, such as gene flow from Eurasians into Africa, or effects of positive selection affecting frequency, influence the distribution of age estimates and make it hard to draw any firm conclusions. We also note that the two introgressed variant counts, derived from the data of^{27,30}, follow a significantly different distribution over time ($p < 2.2 \times 10^{-16}$, Kolmogorov–Smirnov test) (Fig. 2c).

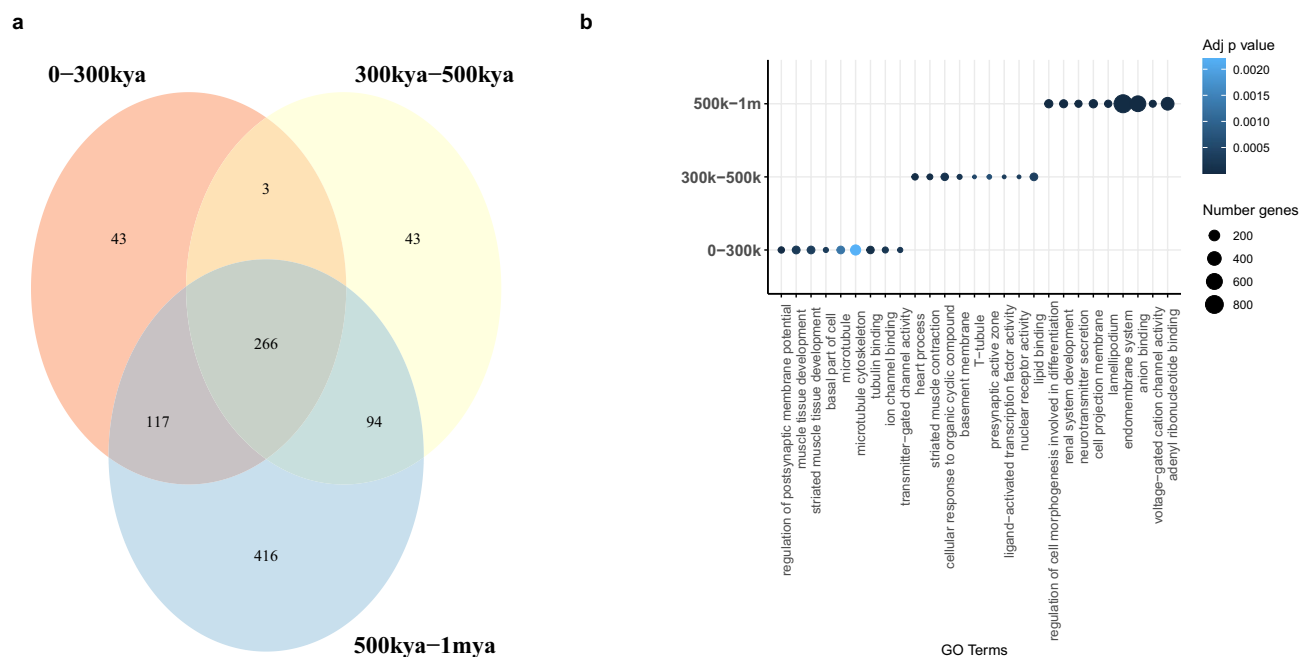


Figure 3. (a) Venn diagram of GO terms associated with genes shared across time windows. (b) Top GO terms per time window.

Finally, we examined the distribution of putatively introgressed variants across populations, focusing on low-frequency variants whose distributions vary when we look at African vs. non-African populations (Fig. S4). As expected, those variants that are more common in non-African populations are found in higher proportions in both of the Neanderthal genomes studied here, with a slightly higher proportion for the Vindija genome, which is in fact assumed to be closer to the main source population of introgression³³. We detect a smaller contribution of Denisovan variants overall, which is expected on several grounds: given the likely more frequent interactions between modern humans and Neanderthals, the Denisovan individual whose genome we relied on is likely part of a more pronounced “outgroup”. Gene flow from modern humans into Neanderthals also likely contributed to this pattern.

In the case of the regions under putative positive selection, we find that the distribution of variant counts has a local peak in the most recent period (0–100 kya) that is absent from the deserts of introgression datasets, pointing to an earlier origin of alleles found in these latter regions. Also, as shown in Fig. 2d, the distribution of variant counts in these regions under selection shows the greatest difference between the two peaks of the bimodal distribution. Still, we should stress that our focus here is on HF variants, and that of course, not all HF variants falling in selective sweep regions were actual targets of selection. Figure S5 illustrates this point for two genes that have figured prominently in early discussions of selective sweeps since⁵: *RUNX2* and *GLI3*. While recent HF variants are associated with positive selection signals (indicated in purple), older variants exhibit such associations as well. Indeed some of these targets may fall below the 90% cutoff chosen in¹⁹. In addition, we are aware that variants enter the genome at one stage and are likely selected for at a (much) later stage^{34,35}. As such our study differs from the chronological atlas of natural selection in our species presented in³⁶ (as well as from other studies focusing on more recent periods of our evolutionary history, such as³⁷). This may explain some important discrepancies between the overall temporal profile of genes highlighted in³⁶ and the distribution of HF variants for these genes in our data (Fig. S6).

Having said this, our analysis recaptures earlier observations about prominent selected variants, located around the most recent peak, concerning genes such as *CADPS2*³⁸ (Fig. S7). This study also identifies a set of old variants, well before 300kya, associated with genes belonging to putative positively-selected regions before the deepest divergence of *Homo sapiens* populations³⁹, such as *LPHN3*, *FBXW7*, and *COG5* (Fig. S8).

Finally, focusing on the brain as the organ that may help explain key features of the rich behavioral repertoire associated with *Homo sapiens*, we estimated the age of putative regulatory variants linked to the prefrontal (PFC), temporal (TC), and cerebellar cortices (CBC), using the large scale characterization of regulatory elements of the human brain provided by the PsychENCODE Consortium⁴⁰. We did the same for the modern human HF missense mutations¹⁹. A comparative plot reveals a similar pattern between the three structures, with no obvious differences in variant distribution (see Fig. S9). The cerebellum contains a slightly higher number of variants assigned to the more recent peak when the proportion to total mapped variants is computed. This may relate to the more recent modifications reported for this brain region⁴¹, which contributed to the globularized shape of our brain(case). We also note that the difference of dated variants between the two local maxima is more pronounced in the case of the cerebellum than in the case of the two cortical tissues, whereas this difference is more reduced in the case of missense variants (Fig. S9). We caution, though, that the overall number of missense variants is considerably lower in comparison to the other three datasets.

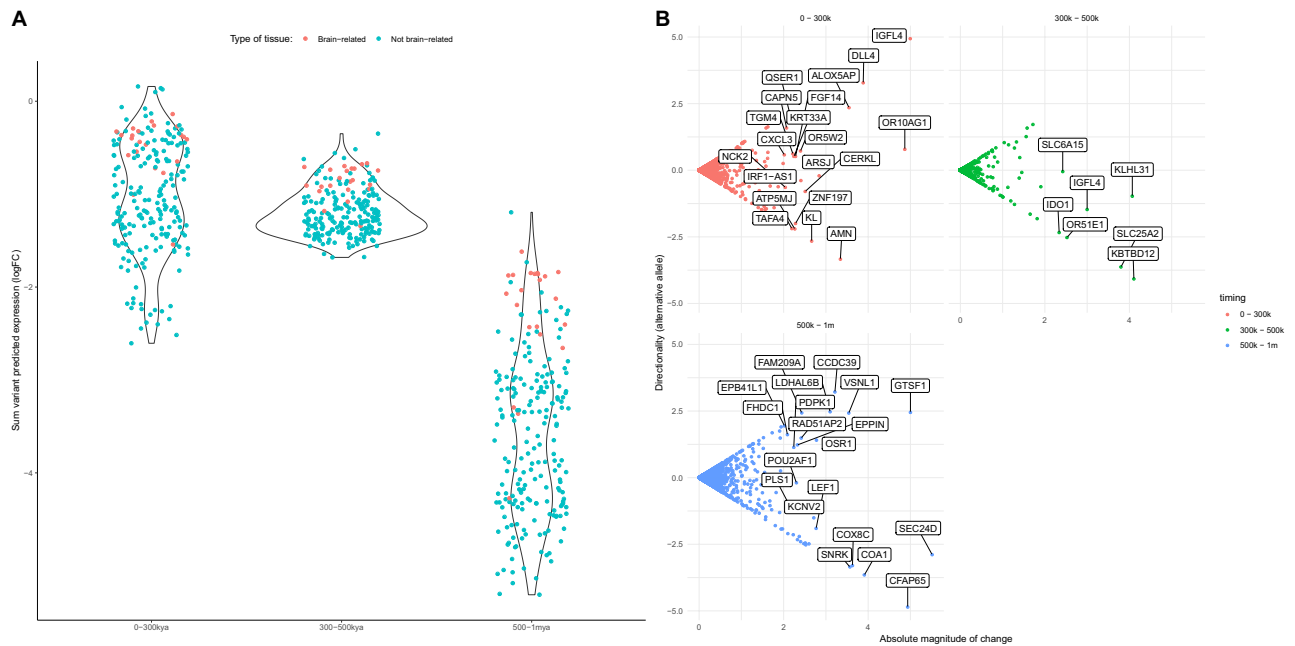


Figure 4. (A) Sum of all directional mutation effects within 1 kb to the TSS per time window in 22 brain and brain-related tissues (red) and the rest of tissues included by the ExPecto trained model as a control group (blue). Significant differences exist across time periods when non-brain and brain-related tissues are compared (Kruskal–Wallis test; $p = 2.2e - 16$). (B) Genes with a high sum of all directional mutation effects, and cumulative directionality of expression values in brain tissues per time window.

Gene Ontology analysis across temporal windows. In order to interpret functionally the distribution of HF variants in time, we performed enrichment analyses accessing curated databases via the *gProfiler2* R package⁴². For the three time windows analyzed (corresponding to the recent peak: 0–300 kya; divergence time and earlier peak: 500 kya–1 mya; and time slot between them: 300 kya–500 kya), we identified unique and shared gene ontology terms (see Fig. 3a,b; “Methods”). Notably, when we compared the most recent period against the two earlier windows together (from 300 kya to 1 mya), we found bone, cartilage, and visual system-related terms only in the earlier periods (hypergeometric test; adj. $p < 0.01$; Table S1). Further differences are observed when thresholding by an adjusted $p < 0.05$. In particular, terms related to behavior (startle response), facial shape (narrow mouth) and hormone systems only appear in the middle (300–500 k) period (Table S2; Fig. S10). Unique gene ontology terms may point to specific environmental conditions causing the organism to react in specific ways. A summary of terms shared across the three time windows can be seen in Fig. S11.

Gene expression predictions. To evaluate the expression profiles associated to our HF variant dataset (from¹⁹), we made use of ExPecto²¹, a sequence-based tool to predict gene expression *in silico* (see description in “Methods”). We found a skewness towards more extreme negative values (downregulation) in brain-related tissues, which is not observed when analyzing all tissues jointly (as shown in quantile-quantile plots in Fig. S12). A series of Kruskal–Wallis test shows that, when either all or just brain-related tissues are considered, statistically significant differences in predicted gene expression values are found across the three time periods studied here ($p = 2.2e-16$ and $p = 4.95e-12$, respectively). Overall, the latest period (500 k–1 mya) reports the strongest predicted effect toward downregulation (see Fig. 4A). Especially for brain-related terms, some structures show the highest sum of variant predicted expression (top downregulation): such as the Adrenal Gland, the Pituitary, Astrocytes, or Neural Progenitor Cells (see Fig. S13). Among these structures, the presence of the cerebellum in a period preceding the last major Out-of-Africa event is noteworthy (consistent with⁴¹).

The authors of the article describing the ExPecto tool²¹ suggest that genes with a high sum of absolute variant effects in specific time windows tend to be tissue or condition-specific. We explored our data to see if the genes with higher absolute variant effect were also phenotypically relevant (Fig. 4B). Among these we find genes such as *DLL4*, a Notch ligand implicated in arterial formation⁴³; *FGF14*, which regulates the intrinsic excitability of cerebellar Purkinje neurons⁴⁴; *SLC6A15*, a gene that modulates stress vulnerability through the glutamate system⁴⁵; and *OPRM1*, a modulator of the dopamine system that harbors a HF derived loss of stop codon variant in the genetic pool of modern humans but not in that of extinct human species¹⁹.

We also crosschecked if any of the variants in our high-frequency dataset with a high predicted expression value (RPKM variant-specific values at $\log > 0.01$) were found in GWASs related to brain volume. The Big40 UKBiobank GWAS meta-analysis⁴⁶ shows that some of these variants are indeed GWAS top hits and can be assigned a date (see Table 1). Of note are phenotypes associated with the posterior Corpus Callosum (Splenum), precuneus, and cerebellar volume. In addition, in a large genome-wide association meta-analysis of brain magnetic resonance imaging data from 51,665 individuals seeking to identify specific genetic loci that

Location	rsid	Nearest gene(s)	GWAS trait	Age (GEVA)
20:49070644	rs75994450	PTPN1	Fractional anisotropy measurement, Splenium (Corpus Callosum)	36,735.46
14:59669037	rs75255901	DAAMI	Functional connectivity (rfMRI)	39,543.24
1:22498451	rs2807369	WNT4	Volume of gray matter in Cerebellum (left)	50,060.96
2:63144695	rs17432559	EHBP1	Volume of Corpus Callosum (Posterior)	52,290.48
12:2231744	rs75557252	CACNA1C	Functional connectivity (rfMRI)	93,924.62
10:92873811	rs17105731	PCGF5	Volume of inferiortemporal gyrus (right)	255,792.5
17:59312894	rs73326893	BCAS3	Functional connectivity (rfMRI)	418,742.6
22:27195261	rs72617274	CRYBA4	Functional connectivity (rfMRI)	445,477.7
2:230367803	rs56049535	DNER	Functional connectivity (rfMRI)	523,629.8
16:3687973	rs78315731	DNASE1	Volume of Pars triangularis (left)	698,856.5

Table 1. Big40 Brain volume GWAS⁴⁶ top hits with high predicted gene expression in ExPecto ($\log > 0.01$, RPKM), along with dating as provided by GEVA. ‘Functional connectivity’ is a measure of temporal activity synchronization between brain parcels at rest (originally defined in⁵¹).

influence human cortical structure⁴⁷, one variant (rs75255901) in Table 1, linked to *DAAMI*, has been identified as a putative causal variant affecting the precuneus. All these brain structures have been independently argued to have undergone recent evolution in our lineage^{41,48–50}, and their associated variants are dated amongst the most recent ones in the table.

Discussion

Deploying GEVA to probe the temporal structure of the extended catalog of HF variants distinguishing modern humans from their closest extinct relatives ultimately aims to contribute to the goals of the emerging attempts to construct a molecular archaeology⁵² and as detailed a map as possible of the evolutionary history of our species⁵³. Like any other archaeology dataset, ours is necessarily fragmentary. In particular, fully fixed mutations, which have featured prominently in early attempts to identify candidates with important functional consequences⁵², fell outside the scope of this study, as GEVA can only determine the age of polymorphic mutations in the present-day human population. By contrast, the mapping of HF variants was reasonably good, and allowed us to provide complementary evidence for claims regarding important stages in the evolution of our lineage. This in and of itself reinforces the rationale of paying close attention to an extended catalog of HF variants, as argued in¹⁹.

While we wait for more genomes from more diverse regions of the planet and from a wider range of time points, we find our results encouraging: even in the absence of genomes from the deep past of our species in Africa, we were able to provide evidence for different epochs and classes of variants that define these. But whereas different clusters can be identified, the emerging picture is very much mosaic-like in its character, in consonance with recent work^{1,3}. In no way do we find evidence for earlier evolutionary narratives that relied on one or a handful of key mutations.

Our analysis shows a bimodal distribution of the age of modern human-derived high-frequency variants (in consonance with the findings of⁵⁴ on a more limited set of variants). The two peaks likely reflect, on the one hand, the point of divergence between *Homo sapiens* and other *Homo* species and, on the other, the period of population dispersal and replacement following the last major out of Africa event.

Our work also highlights the importance of a temporal window right before 300 ky that may well correspond to a significant behavioral shift in our lineage, such as increased ecological resource variability⁵⁵, and evidence of long-distance stone transport and pigment use⁵⁶. Other aspects of our cognitive and anatomical make up emerged much more recently, in the last 150 k years, and for these our analysis points to the relevance of gene expression regulation differences in recent human evolution, in line with^{57–59}.

Lastly, our attempt to date the emergence of mutations in our genomes points to multiple episodes of introgression, whose history is likely to turn out to be quite complex.

Methods

Homo sapiens variant catalog. We made use of a publicly available dataset¹⁹ that takes advantage of the Neanderthal and Denisovan genomes to compile a genome-wide catalog of *Homo sapiens*-specific variation. The original complete dataset is available at <https://doi.org/10.6084/m9.figshare.8184038>. As described in the original article, this catalog includes “archaic”-specific variants and all loci showing variation within modern populations. The 1000 genomes project and ExAc data were used to derive frequencies and the human genome version *hg19* as reference. As indicated in the original publication¹⁹, quality filters in the “archaic” genomes were applied (specifically: sites with less 5-fold coverage and more than 105-fold coverage for the Altai individual, or 75-fold coverage for the rest of “archaic” individuals were filtered out). In ambiguous cases, variant ancestry was determined using multiple genome alignments⁶⁰ and the macaque reference sequence (*rheMac3*)⁶¹.

In addition to the full data, the authors offered a subset of the data that includes derived variants at a $\geq 90\%$ global frequency cutoff. Since such a cutoff allows some variants to reach less than 90% in certain populations, as long as the total is $\geq 90\%$, we also considered including a metapopulation-wide variant $\geq 90\%$ frequency cutoff dataset to this study (Fig. S2). All files (including the original full and high-frequency sets and the modified,

stricter high-frequency one) are provided in the accompanying code. Controls in 1 were obtained through a probabilistic permutation approach with sets of random variants (100 sets, 50,000 variants each).

GEVA. The Genealogical Estimation of Variant Age (GEVA) tool²⁰ uses a hidden Markov model approach to infer the location of ancestral haplotypes relative to a given variant. It then infers time to the most recent ancestor in multiple pairwise comparisons by coalescent-based clock models. The resulting pairwise information is combined in a posterior probability measure of variant age. We extracted dating information for the alleles of our dataset from the bulk summary information of GEVA age predictions. The GEVA tool provides several clock models and measures for variant age. We chose the mean age measure from the joint clock model, that combines recombination and mutation estimates. While the GEVA dataset provides data for the 1000 genomes project and the Simons Genome Diversity Project, we chose to extract only those variants that were present in both datasets. Ensuring a variant is present in both databases implicitly increases genealogical estimates (as detailed in Supplementary document 3 of²⁰), although it decreases the amount of sites that can be looked at. We give estimated dates after assuming 29 years per generation, as suggested in⁶². While other measures can be chosen, this value should not affect the nature of the variant age distribution nor our conclusions.

Out of a total of 4,437,804 for our total set of variants, 2,294,023 were mapped in the GEVA dataset (51% of the original total). For the HF subsets, the mapping improves: 101,417 (74% of total) and 48,424 (69%) variants were mapped for the original high-frequency subset and the stricter, meta-population cutoff version, respectively.

ExPecto. In order to predict gene expression we made use of the *ExPecto* tool²¹. *ExPecto* is a deep convolutional network framework that predicts tissue-specific gene expression directly from genetic sequences. *ExPecto* is trained on histone mark, transcription factor and DNA accessibility profiles, allowing ab initio prediction that does not rely on variant information training. Sequence-based approaches, such as the one used by *ExPecto*, allow to predict the expression of high-frequency and rare alleles without the biases that other frameworks based on variant information might introduce. We introduced the high-frequency dated variants as input for *ExPecto* expression prediction, using the default tissue training models trained on the GTEx, Roadmap genomics and ENCODE tissue expression profiles.

gProfiler2. Enrichment analysis was performed using *gProfiler2* package⁴² (hypergeometric test; multiple comparison correction, 'gSCS' method; p values 0.01 and 0.05). Dated variants were subdivided in three time windows (0–300 kya, 300–500 kya and 500 kya–1 mya) and variant-associated genes (retrieved from¹⁹) were used as input (all annotated genes for *H. sapiens* in the Ensembl database were used as background). Following²¹, variation potential directionality scores were calculated as the sum of all variant effects in a range of 1 kb from the TSS. Summary GO figures presented in Fig. S11 were prepared with *GO Figure*⁶³.

For enrichment analysis, the Hallmark curated annotated sets⁶⁴ were also consulted, but the dated set of HF variants as a whole did not return any specific enrichment.

Code availability

All the analysis here presented can be reproduced following the scripts in the following Github repository: <https://github.com/AGMAndirko/Temporal-mapping>.

Received: 12 January 2022; Accepted: 25 May 2022

Published online: 15 June 2022

References

- Scerri, E. M. L. *et al.* Did our species evolve in subdivided populations across Africa, and why does it matter?. *Trends Ecol. Evol.* **33**, 582–594. <https://doi.org/10.1016/j.tree.2018.05.005> (2018).
- Groucutt, H. S. *et al.* Multiple hominin dispersals into Southwest Asia over the past 400,000 years. *Nature* **597**, 376–380. <https://doi.org/10.1038/s41586-021-03863-y> (2021).
- Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L. & Skoglund, P. Origins of modern human ancestry. *Nature* **590**, 229–237. <https://doi.org/10.1038/s41586-021-03244-5> (2021).
- Sykes, R. W. *Kindred: 300,000 Years of Neanderthal Life and Afterlife* OCLC: 1126396038 (Bloomsbury Publishing, 2020).
- Green, R. E. *et al.* A draft sequence of the neanderthal genome. *Science* **328**, 710–722. <https://doi.org/10.1126/science.1188021> (2010).
- Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433. <https://doi.org/10.1038/nature16544> (2016).
- Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. & Akey, J. M. Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell* **173**, 53–61.e9. <https://doi.org/10.1016/j.cell.2018.02.031> (2018).
- Gokcumen, O. Archaic hominin introgression into modern human genomes. *Am. J. Phys. Anthropol.* **171**, 60–73. <https://doi.org/10.1002/ajpa.23951> (2020).
- Posth, C. *et al.* Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nat. Commun.* **8**, 16046. <https://doi.org/10.1038/ncomms16046> (2017).
- Stringer, C. The origin and evolution of *Homo sapiens*. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20150237. <https://doi.org/10.1098/rstb.2015.0237> (2016).
- de Boer, B., Thompson, B., Ravnani, A. & Boeckx, C. Evolutionary dynamics do not motivate a single-mutant theory of human language. *Sci. Rep.* **10**, 451. <https://doi.org/10.1038/s41598-019-57235-8> (2020).
- Schlebusch, C. M. *et al.* Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655. <https://doi.org/10.1126/science.aao6266> (2017).
- Prendergast, M. E. *et al.* Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science* <https://doi.org/10.1126/science.aaw6275> (2019).
- Lipson, M. *et al.* Ancient DNA and deep population structure in sub-Saharan African foragers. *Nature* <https://doi.org/10.1038/s41586-022-04430-9> (2022).

15. Grün, R. *et al.* Dating the skull from Broken Hill, Zambia, and its position in human evolution. *Nature* **580**, 372–375. <https://doi.org/10.1038/s41586-020-2165-4> (2020).
16. Hubisz, M. J., Williams, A. L. & Siepel, A. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS Genet.* **16**, e1008895. <https://doi.org/10.1371/journal.pgen.1008895> (2020).
17. Durvasula, A. & Sankararaman, S. Recovering signals of ghost archaic introgression in African populations. *Sci. Adv.* **6**, eaax5097. <https://doi.org/10.1126/sciadv.aax5097> (2020).
18. Lacruz, R. S. *et al.* The evolutionary history of the human face. *Nat. Ecol. Evol.* **3**, 726–736. <https://doi.org/10.1038/s41559-019-0865-7> (2019).
19. Kuhlwilm, M. & Boeckx, C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci. Rep.* **9**, 8463. <https://doi.org/10.1038/s41598-019-44877-x> (2019).
20. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586. <https://doi.org/10.1371/journal.pbio.3000586> (2020).
21. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179. <https://doi.org/10.1038/s41588-018-0160-6> (2018).
22. Groucutt, H. S. *et al.* Rethinking the dispersal of Homo sapiens out of Africa. *Evol. Anthropol.* **24**, 149–164. <https://doi.org/10.1002/evan.21455> (2015).
23. Prüfer, K. *et al.* A genome sequence from a modern human skull over 45,000 years old from Zlatý kůň in Czechia. *Nat. Ecol. Evol.* **5**, 820–825. <https://doi.org/10.1038/s41559-021-01443-x> (2021).
24. Gómez-Robles, A. Dental evolutionary rates and its implications for the Neanderthal-modern human divergence. *Sci. Adv.* **5**, eaaw1268. <https://doi.org/10.1126/sciadv.aaw1268> (2019).
25. Hublin, J.-J. *et al.* New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* **546**, 289–292. <https://doi.org/10.1038/nature22336> (2017).
26. BermúdezdeCastro, J. M. *et al.* A hominid from the lower Pleistocene of Atapuerca, Spain. *Science (New York, N.Y.)* **276**, 1392–1395. <https://doi.org/10.1126/science.276.5317.1392> (1997).
27. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol.* **26**, 1241–1247. <https://doi.org/10.1016/j.cub.2016.03.037> (2016).
28. Chen, L., Wolf, A. B., Fu, W., Li, L. & Akey, J. M. Identifying and interpreting apparent Neanderthal ancestry in African individuals. *Cell* **180**, 677–687.e16. <https://doi.org/10.1016/j.cell.2020.01.012> (2020).
29. Peyrégne, S., Boyle, M. J., Dannemann, M. & Prüfer, K. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res.* **27**, 1563–1572. <https://doi.org/10.1101/gr.219493.116> (2017).
30. Vernot, B. *et al.* Excavating Neanderthal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239. <https://doi.org/10.1126/science.aad9416> (2016).
31. Petr, M. *et al.* The evolutionary history of Neanderthal and Denisovan Y chromosomes. *Science* **369**, 1653–1656. <https://doi.org/10.1126/science.abb6460> (2020).
32. McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-Introgressed sequences on the landscape of human gene expression. *Cell* **168**, 916–927.e12. <https://doi.org/10.1016/j.cell.2017.01.038> (2017).
33. Taskent, O., Lin, Y. L., Patramanis, I., Pavlidis, P. & Gokcumen, O. Analysis of haplotypic variation and deletion polymorphisms point to multiple archaic introgression events, including from Altai Neanderthal Lineage. *Genetics* **215**, 497–509. <https://doi.org/10.1534/genetics.120.303167> (2020).
34. Zhang, X. *et al.* The history and evolution of the Denisovan-EPAS1 haplotype in Tibetans. *bioRxiv*. <https://doi.org/10.1101/2020.10.01.323113> (2020).
35. Yair, S., Lee, K. M. & Coop, G. The timing of human adaptation from Neanderthal introgression. *bioRxiv*. <https://doi.org/10.1101/2020.10.04.325183> (2020).
36. Zhou, H. *et al.* A chronological atlas of natural selection in the human genome during the past half-million years. *bioRxiv*. <https://doi.org/10.1101/018929> (2015).
37. Tilot, A. K. *et al.* The evolutionary history of common genetic variants influencing human cortical surface area. *Cereb. Cortex* <https://doi.org/10.1093/cercor/bhaa327> (2020).
38. Racimo, F. Testing for ancient selection using cross-population allele frequency differentiation. *Genetics* **202**, 733–750. <https://doi.org/10.1534/genetics.115.178095> (2016).
39. Schlebusch, C. M. *et al.* Khoe-San genomes reveal unique variation and confirm the deepest population divergence in homo sapiens. *Mol. Biol. Evol.* **37**, 2944–2954. <https://doi.org/10.1093/molbev/msaa140> (2020).
40. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464. <https://doi.org/10.1126/science.aat8464> (2018).
41. Neubauer, S., Hublin, J.-J. & Gunz, P. The evolution of modern human brain shape. *Sci. Adv.* **4**, eaao5961. <https://doi.org/10.1126/sciadv.aao5961> (2018).
42. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200. <https://doi.org/10.1093/nar/gkm226> (2007).
43. Pitulescu, M. E. *et al.* Dll4 and Notch signalling couples sprouting angiogenesis and artery formation. *Nat. Cell Biol.* **19**, 915–927. <https://doi.org/10.1038/ncb3555> (2017).
44. Bosch, M. K. *et al.* Intracellular FGF14 (iFGF14) Is required for spontaneous and evoked firing in cerebellar purkinje neurons and for motor coordination and balance. *J. Neurosci.* **35**, 6752–6769. <https://doi.org/10.1523/JNEUROSCI.2663-14.2015> (2015).
45. Santarelli, S. *et al.* SLC6A15, a novel stress vulnerability candidate, modulates anxiety and depressive-like behavior: Involvement of the glutamatergic system. *Stress (Amsterdam, Netherlands)* **19**, 83–90. <https://doi.org/10.3109/10253890.2015.1105211> (2016).
46. Smith, S. M. *et al.* Enhanced brain imaging genetics in UK Biobank. *bioRxiv*. <https://doi.org/10.1101/2020.07.27.223545> (2020).
47. Grasby, K. L. *et al.* The genetic architecture of the human cerebral cortex. *Science* <https://doi.org/10.1126/science.aay6690> (2020).
48. Theofanopoulou, C. Brain asymmetry in the white matter making and globularity. *Front. Psychol.* <https://doi.org/10.3389/fpsyg.2015.01355> (2015).
49. Bruner, E. Human Paleoneurology and the Evolution of the Parietal. *Cortex* <https://doi.org/10.1159/000488889> (2018).
50. Lombard, M. & Höglberg, A. Four-field co-evolutionary model for human cognition: Variation in the middle stone age/middle palaeolithic. *J. Archaeol. Method Theory* <https://doi.org/10.1007/s10816-020-09502-6> (2021).
51. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216. <https://doi.org/10.1038/s41586-018-0571-7> (2018).
52. Pääbo, S. The human condition—a molecular approach. *Cell* **157**, 216–226. <https://doi.org/10.1016/j.cell.2013.12.036> (2014).
53. Wohns, A. W. *et al.* A unified genealogy of modern and ancient genomes. *Science* **375**, 2eabi82eabi8264. <https://doi.org/10.1126/science.abi8264> (2021).
54. Schaefer, N. K., Shapiro, B. & Green, R. E. An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Sci. Adv.* **7**, eabc0776. <https://doi.org/10.1126/sciadv.abc0776> (2021).
55. Potts, R. *et al.* Increased ecological resource variability during a critical transition in hominin evolution. *Sci. Adv.* **6**, eabc8975. <https://doi.org/10.1126/sciadv.abc8975> (2020).
56. Brooks, A. S. *et al.* Long-distance stone transport and pigment use in the earliest Middle Stone Age. *Science* **360**, 90–94. <https://doi.org/10.1126/science.aao2646> (2018).

57. Moriano, J. & Boeckx, C. Modern human changes in regulatory regions implicated in cortical development. *BMC Genom.* **21**, 304. <https://doi.org/10.1186/s12864-020-6706-x> (2020).
58. Weiss, C. V. *et al.* The cis-regulatory effects of modern human-specific variants. *bioRxiv* <https://doi.org/10.1101/2020.10.07.330761> (2020).
59. Yan, S. M. & McCoy, R. C. Archaic hominin genomics provides a window into gene expression evolution. *Curr. Opin. Genet. Dev.* **62**, 44–49. <https://doi.org/10.1016/j.gde.2020.05.014> (2020).
60. Paten, B. *et al.* Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843. <https://doi.org/10.1101/gr.076521.108> (2008).
61. Yan, G. *et al.* Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat. Biotechnol.* **29**, 1019–1023. <https://doi.org/10.1038/nbt.1992> (2011).
62. Fenner, J. N. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. *Am. J. Phys. Anthropol.* **128**, 415–423. <https://doi.org/10.1002/ajpa.20188> (2005).
63. Reijnders, M. J. & Waterhouse, R. M. Summary visualisations of gene ontology terms with GO-Figure!. *bioRxiv* <https://doi.org/10.1101/2020.12.02.408534> (2020).
64. Liberzon, A. *et al.* The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425. <https://doi.org/10.1016/j.cels.2015.12.004> (2015).

Author contributions

Conceptualization: C.B., A.A. and J.M.; methodology: C.B., A.A. and J.M.; data curation: A.A. and J.M.; software: A.A. and J.M.; formal analysis: A.A. and J.M.; visualization: C.B., A.A., J.M., A.V., M.K. and G.T.; investigation: C.B., A.A., J.M., A.V., M.K. and G.T.; writing—original draft preparation: C.B., A.A. and J.M.; writing—review and editing: C.B., A.A., J.M., A.V., M.K. and G.T.; supervision: C.B.; funding acquisition: C.B.

Funding

CB acknowledges support from the Spanish Ministry of Science and Innovation (Grant PID2019-107042GB-I00), MEXT/JSPS Grant-in-Aid for Scientific Research on Innovative Areas #4903 (Evolinguistics: JP17H06379), Generalitat de Catalunya (2017-SGR-341), and the support of a 2020 Leonardo Grant for Researchers and Cultural Creators, BBVA Foundation. AA acknowledges financial support from the Spanish Ministry of Economy and Competitiveness and the European Social Fund (BES-2017-080366). JM acknowledges financial support from the Departament d'Empresa i Coneixement, Generalitat de Catalunya (FI-SDUR 2020). MK was supported by “la Caixa” Foundation (ID 100010434), fellowship code LCF/BQ/PR19/11700002, and by the Vienna Science and Technology Fund (WWTF) and the City of Vienna through project VRG20-001. Funding bodies take no responsibility for the opinions, statements and contents of this project, which are entirely the responsibility of its authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-13589-0>.

Correspondence and requests for materials should be addressed to C.B.

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

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

Supplementary Figures

for

“Temporal mapping of derived high-frequency gene variants supports
the mosaic nature of the evolution of *Homo sapiens*”

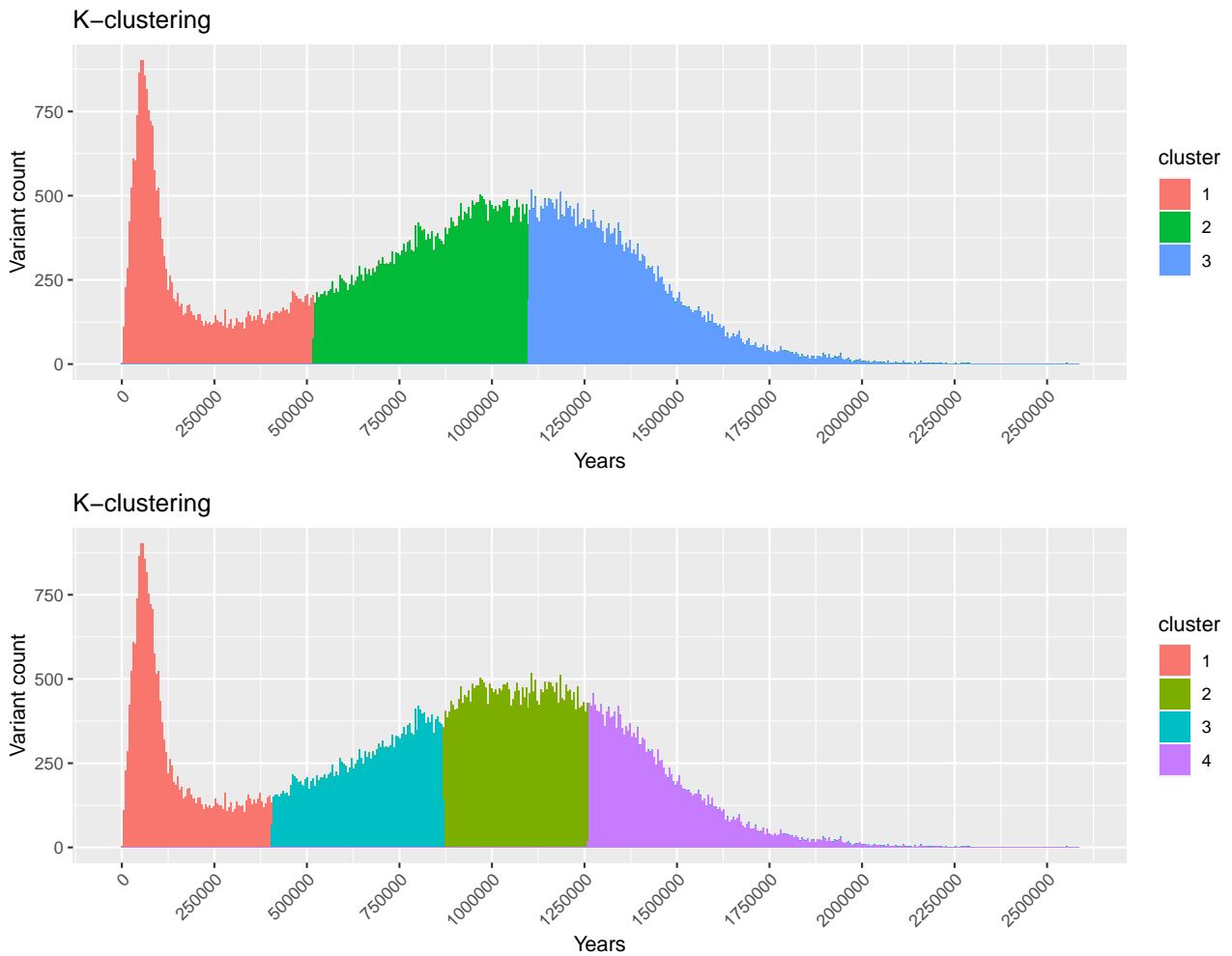


Figure 1: K -means clustering analysis of HF variant temporal distribution, for both $k = 3$ (top) and $k = 4$ (bottom).

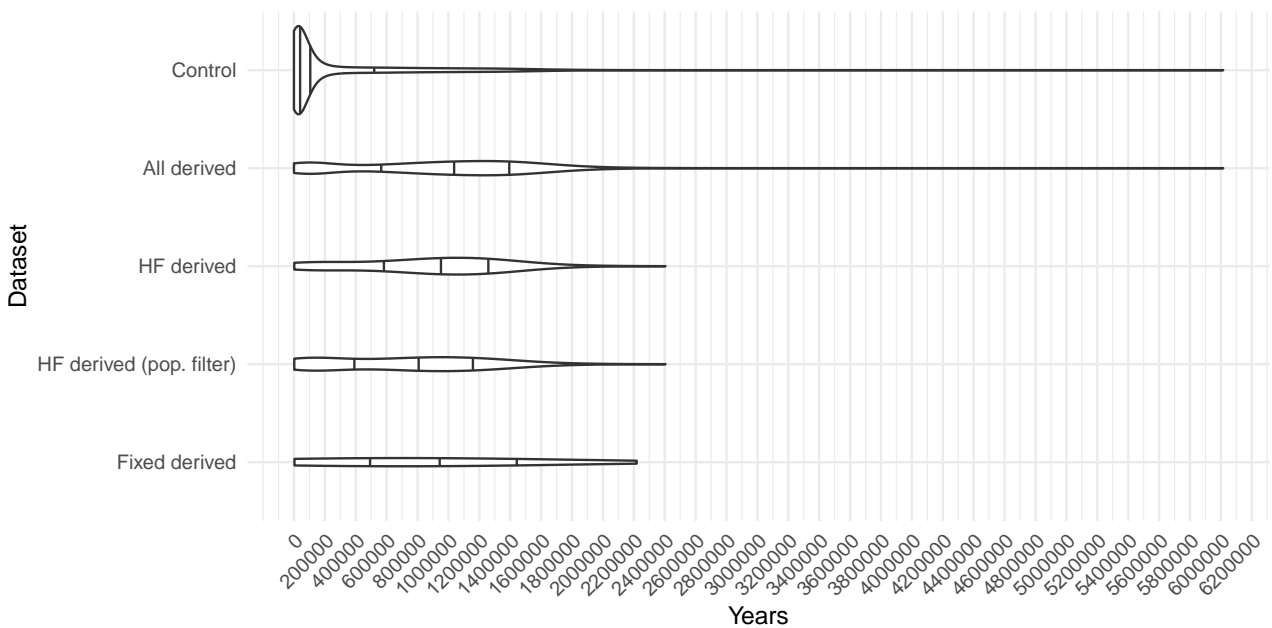


Figure 2: Density distribution of derived *Homo sapiens* alleles for different subsets used in this study (related to Figure 1). From top to bottom, control set of random variants; all derived variants in the *Homo sapiens* lineage; those variants at high-frequency (HF); HF variants with an added meta-population filter (see sec. 4); and variants that reached fixation. A direct comparison of all variants, HF and HF with the added filter can be found in Figure S2. Horizontal lines mark distribution quantiles 0.25, 0.5 and 0.75.

Top McCoy et al. (2017) Neanderthal-introgressed variants linked to phenotypes

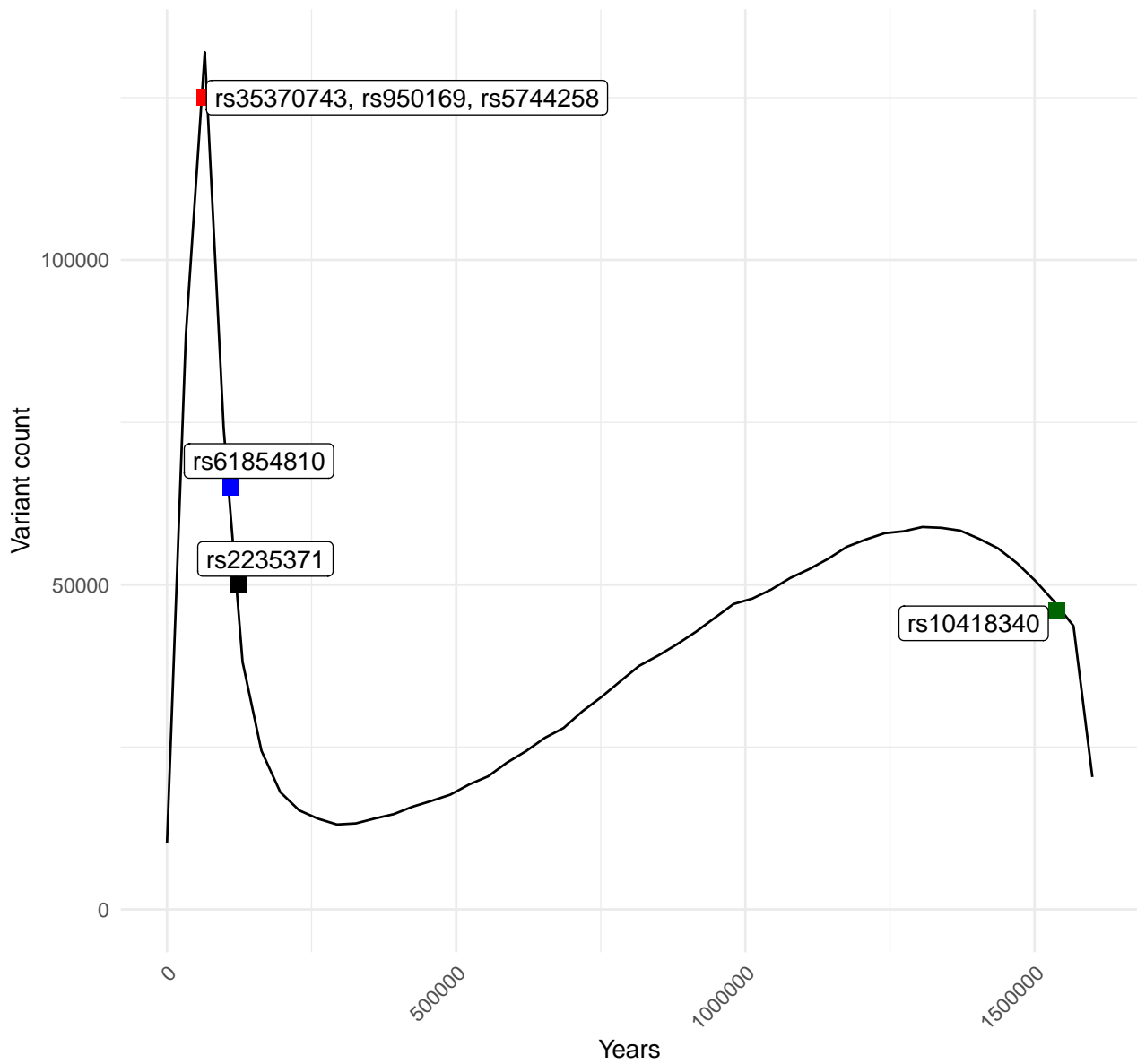


Figure 3: Temporal distribution of introgressed variants linked to phenotypes, as highlighted in Table 1 of [1], compared to the distribution of all derived variants over time.

Populations filters

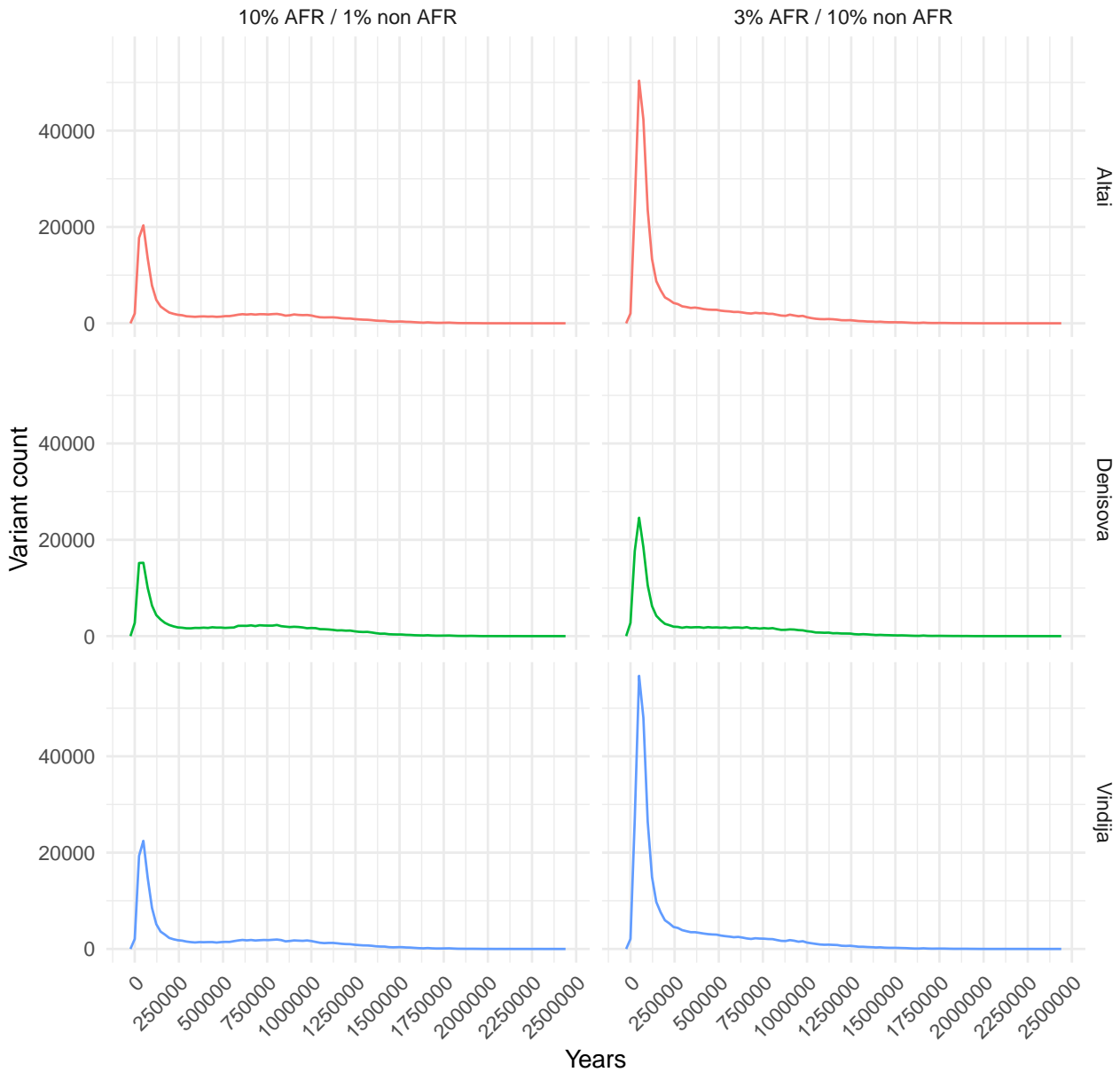


Figure 4: Temporal distribution of variants shared with each of the extinct human genomes after applying specific population frequency filters. These filter include a 10% minor allele frequency cutoff in the African metapopulation (AFR), coupled with a 1% cutoff in the rest of metapopulations, designed to detect potential introgressed alleles brought into the African genetic pool by back-to-Africa migration events. The second filter applied is a 3% cutoff in AFR populations and a 10% threshold in non-African populations, designed to detect the contribution of each extinct human sample to the introgressed variant genetic pool, accounting for a third of that pool to be introduced in AFR populations by back-to-Africa migrations.

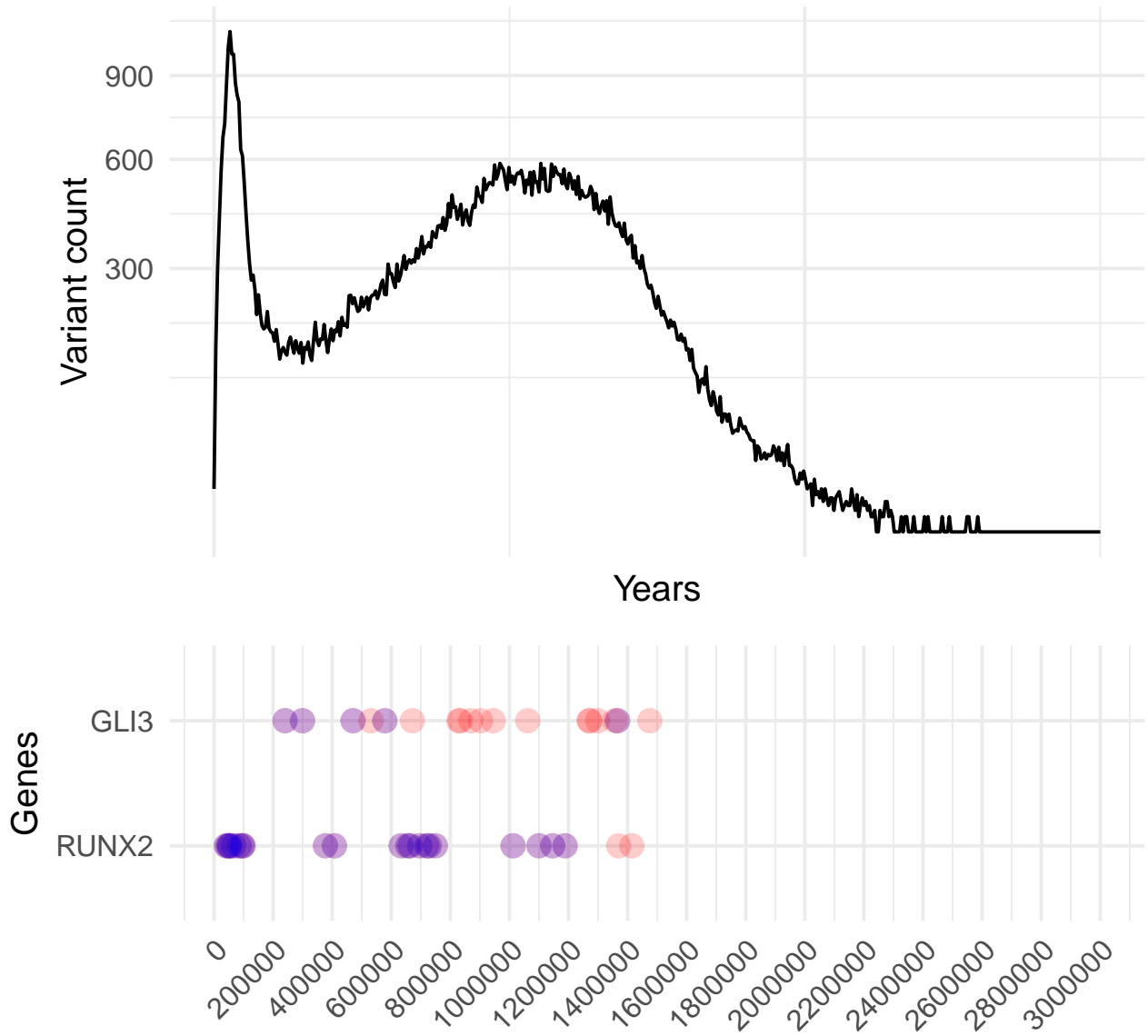


Figure 5: Temporal distribution of HF variants in two genes highlighted in early discussions of selective sweeps: *GLI3* (sweep region from [2]) and *RUNX2* (sweep region from [3]). Variants in purple fall within sweep regions.

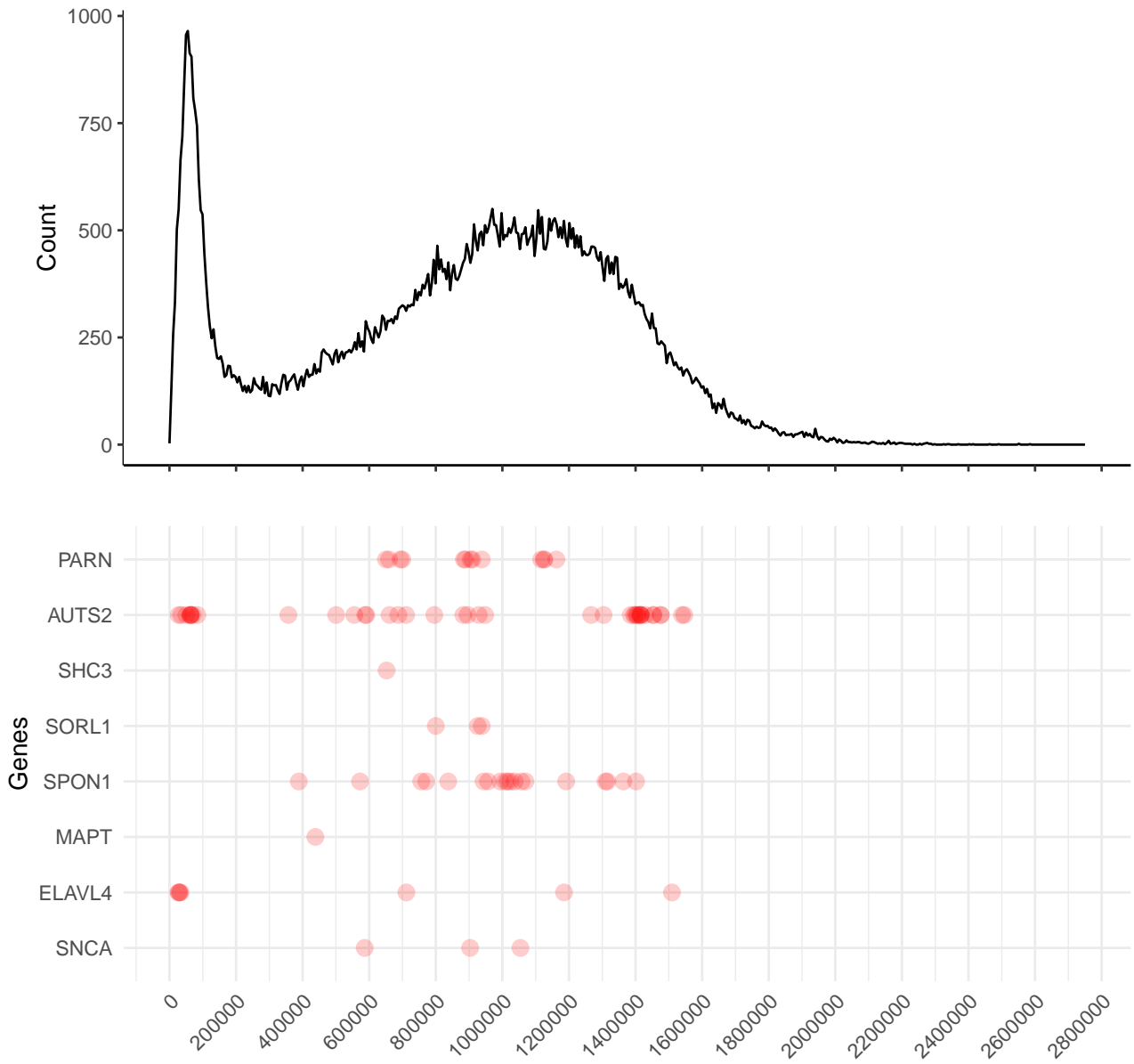


Figure 6: Temporal distribution of variants associated with genes highlighted in [4].

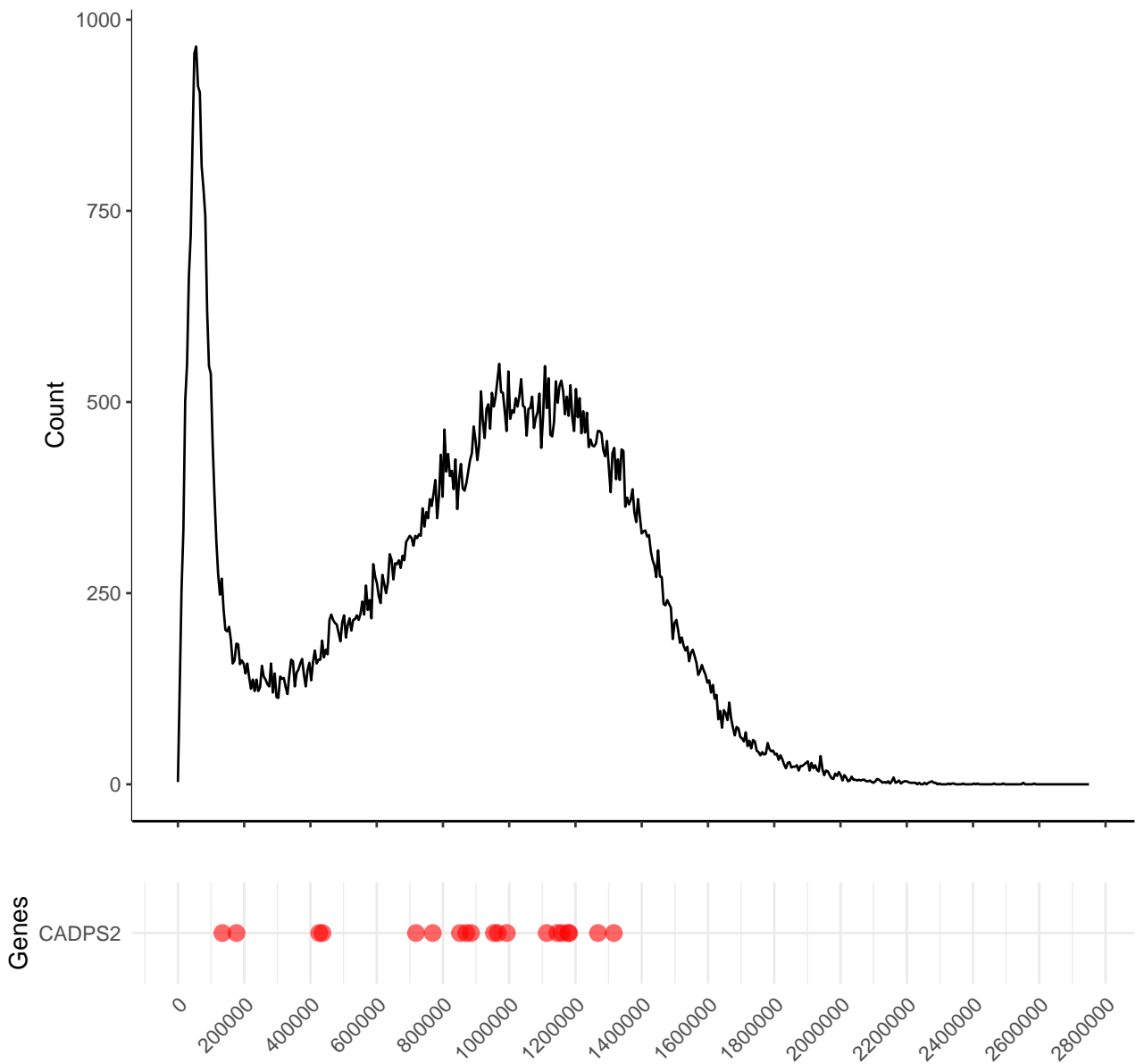


Figure 7: Temporal distribution of variants associated with *CADPS2*. The most recent variants around 200kya in particular capture the reasons this gene was highlighted in [5]: “*CADPS2* was identified in [3] as a candidate for selection The gene has been suggested to be specifically important in the evolution of all modern humans, as it was not found to be selected earlier in great apes or later in particular modern human populations”.

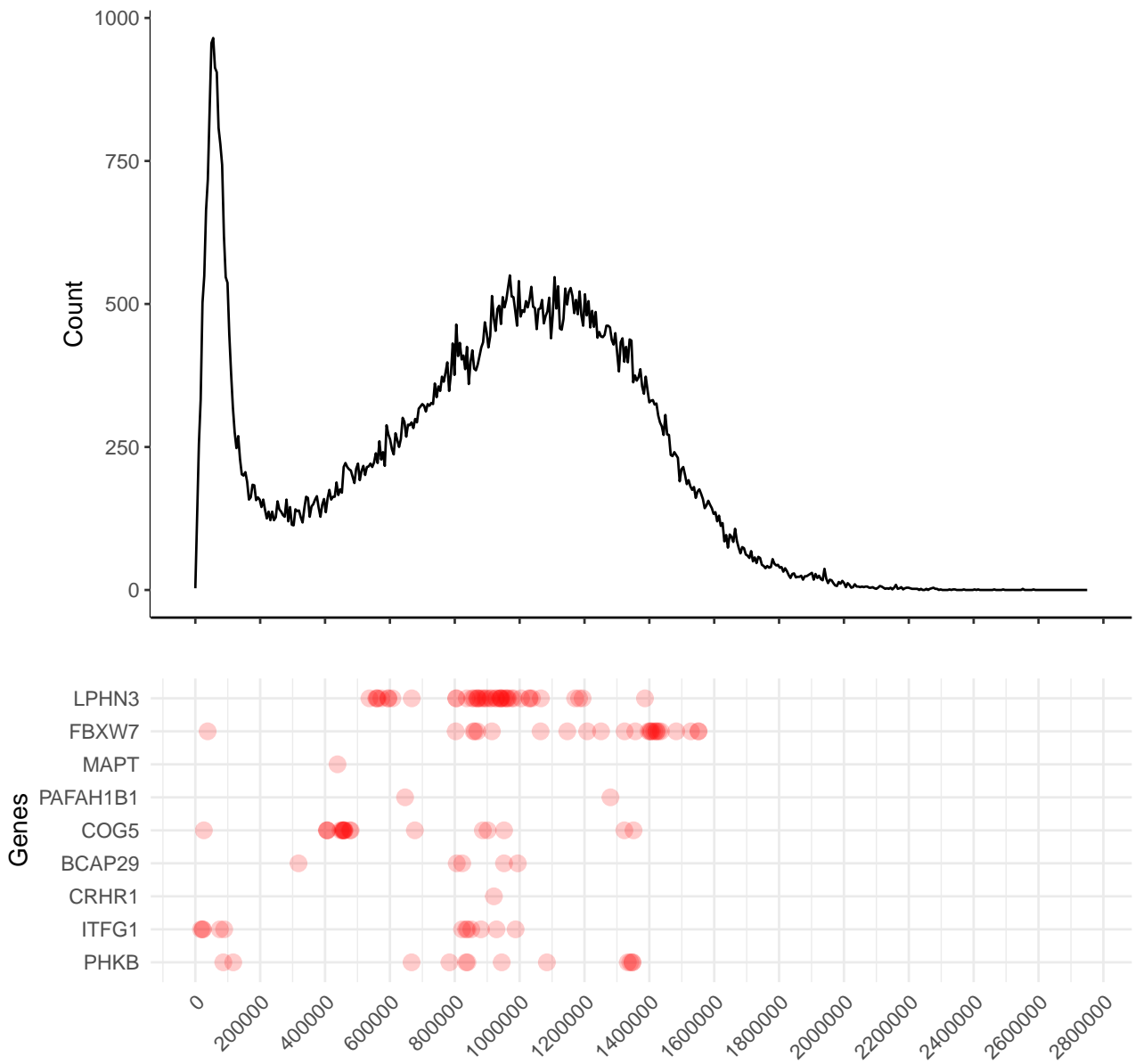


Figure 8: Temporal distribution of variants in genes found in putative positively-selected genetic windows before early *Homo sapiens* population divergence, as per [6]. Genes belonging to putative positively selected regions were retrieved from Supplementary Data, section 12 of [6].

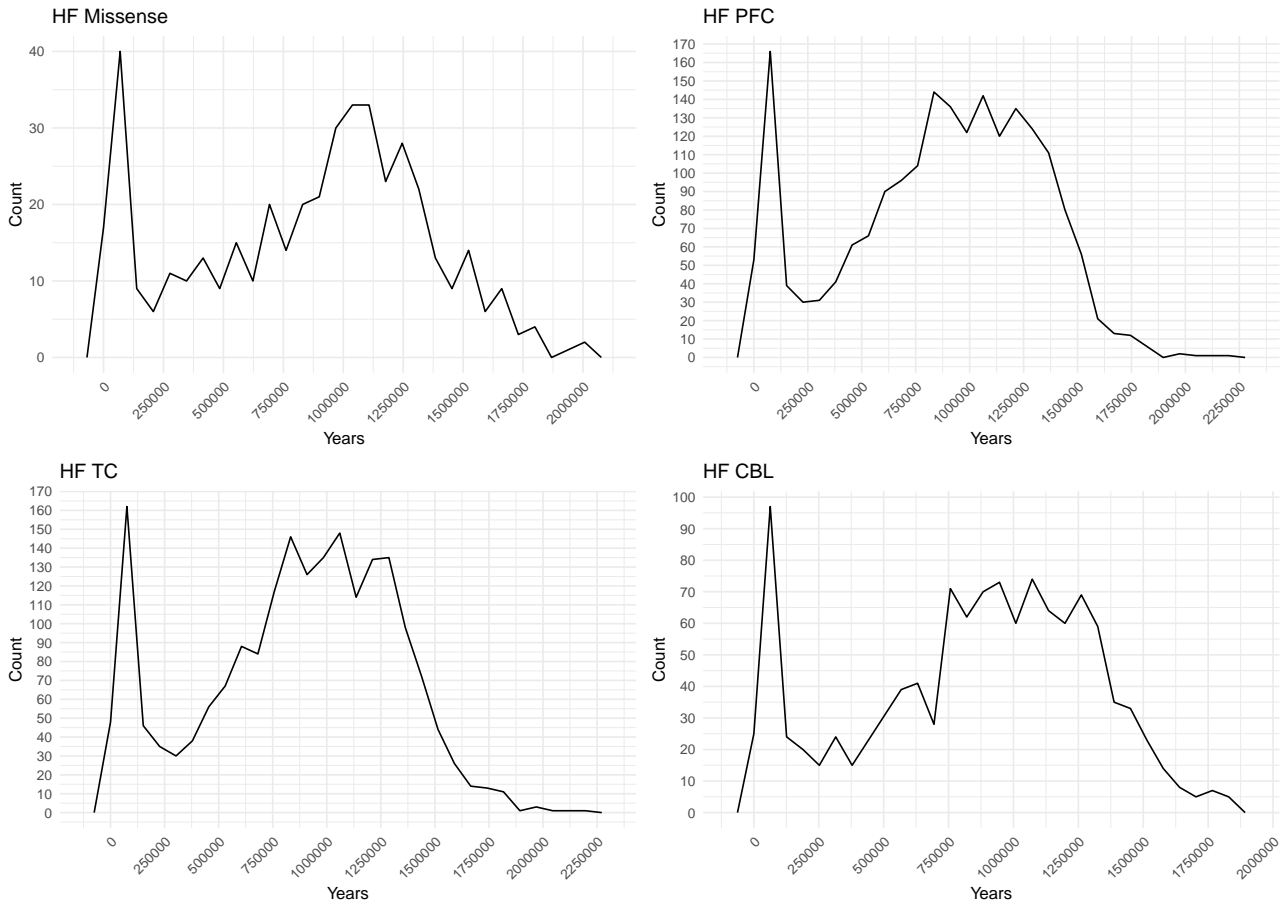


Figure 9: Temporal distribution of high-frequency missense and regulatory variants. Missense variants derived from [7]; enhancer annotations for the prefrontal, temporal and cerebellar cortices were retrieved from [8]. The difference between the two total maximum counts in the left to the right peak is more pronounced in the cerebellum and prefrontal cortices (23 and 22 more variants mapped to the left maximum peak, respectively). This same difference for missense variants is reduced to only 7 more variants mapped to the left maximum peak. For the temporal cortex, this difference amounts to 14 mapped variants.

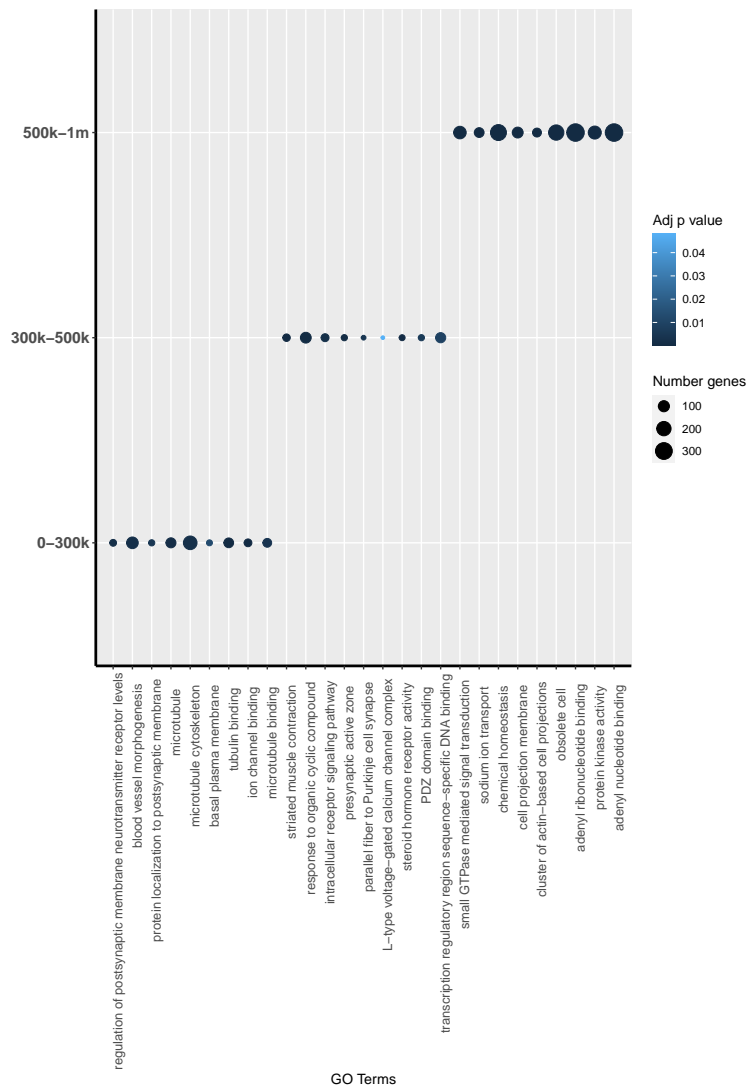
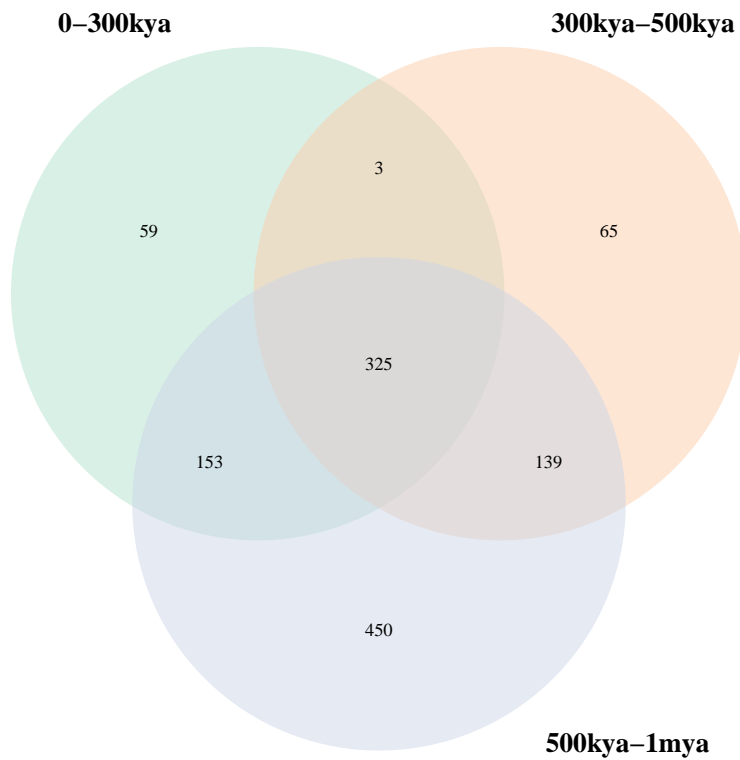
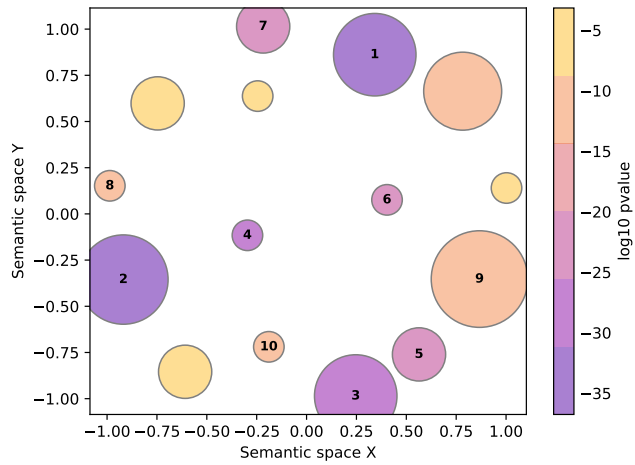
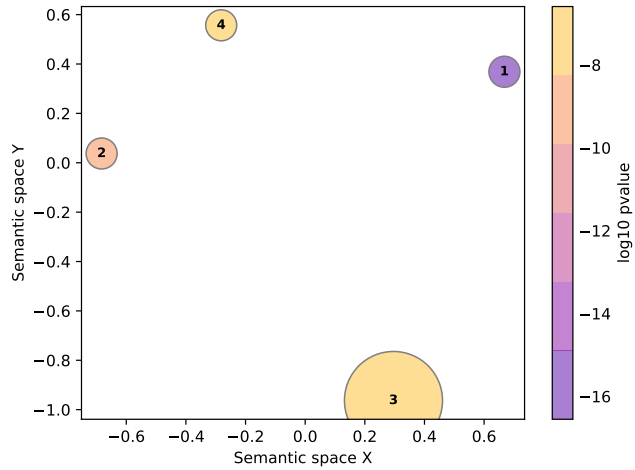


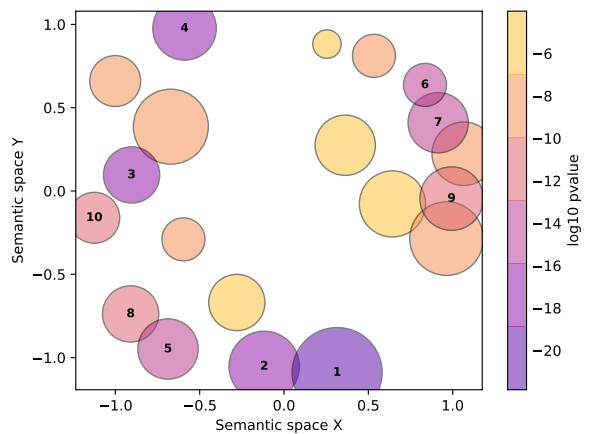
Figure 10: GO terms results when thresholding by an adjusted p -value of 0.05. Venn diagram (top) shows number of unique and shared GO terms across periods. Dot plot (bottom) highlights the top 3 GO terms by significance for each period.



- | | |
|---------------------------|---|
| 1. cell junction | 6. postsynapse |
| 2. cell projection | 7. postsynaptic density |
| 3. plasma membrane region | 8. somatodendritic compartment |
| 4. cell periphery | 9. intrinsic component of synaptic mem... |
| 5. plasma membrane | 10. presynapse |



- | | |
|---------------------------------|-------------------------|
| 1. cytoskeletal protein binding | 3. ion channel activity |
| 2. calcium ion binding | 4. ion binding |



- | | |
|---|---|
| 1. anatomical structure morphogenesis | 6. regulation of signaling |
| 2. trans-synaptic signaling | 7. regulation of cell communication |
| 3. cell adhesion | 8. cell junction organization |
| 4. multicellular organismal process | 9. regulation of multicellular organis... |
| 5. plasma membrane bounded cell projec... | 10. neuron projection guidance |

Figure 11: GO term reduction of shared terms across time windows (center of Venn diagram in Fig. 3A).

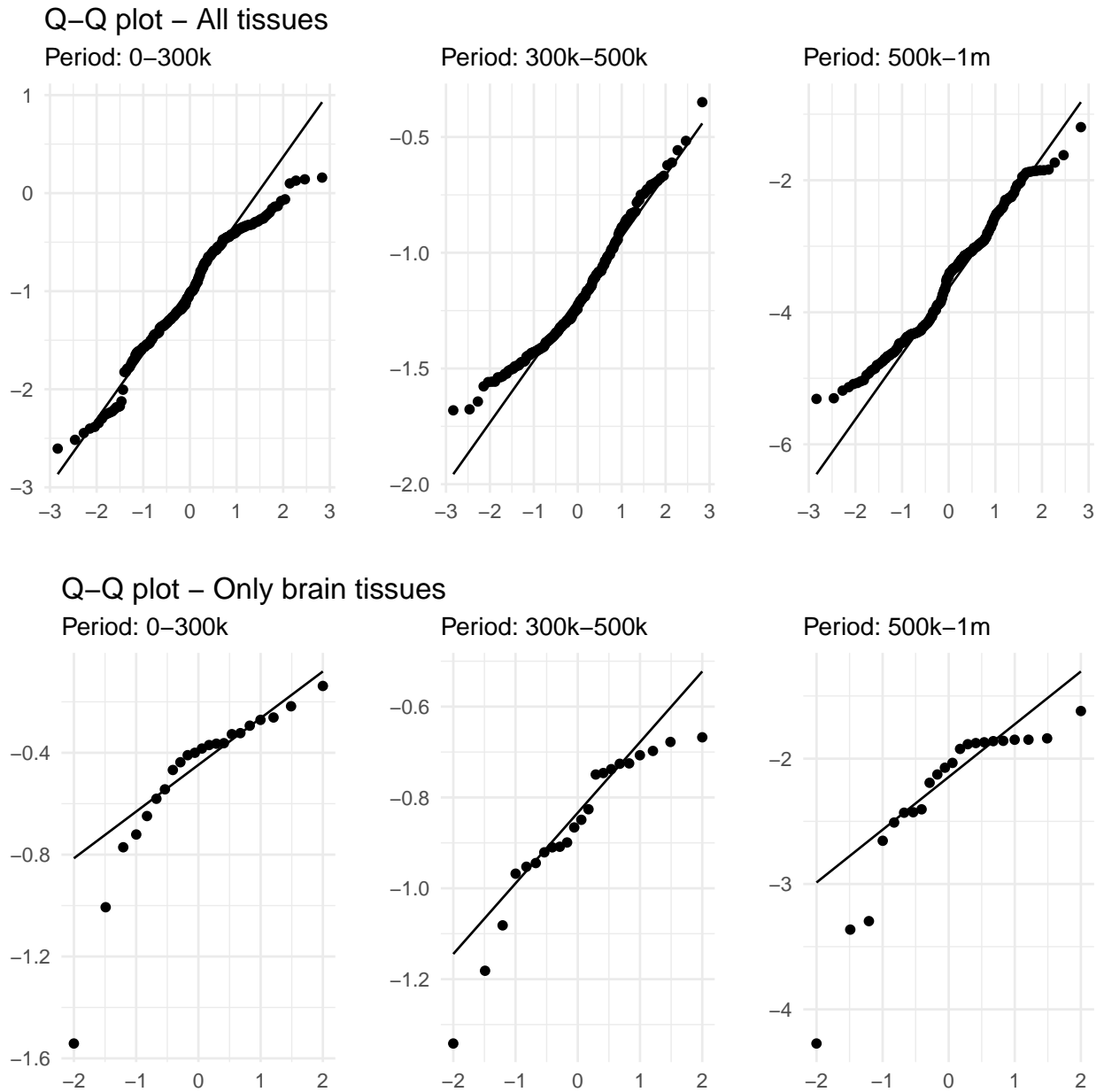


Figure 12: Quantile-quantile plots of predicted expression values of high-frequency variants, divided in three time periods (0-300kya, 300k-500kya and 500k-1mya). Applied to both all tissues and only brain-related tissues.

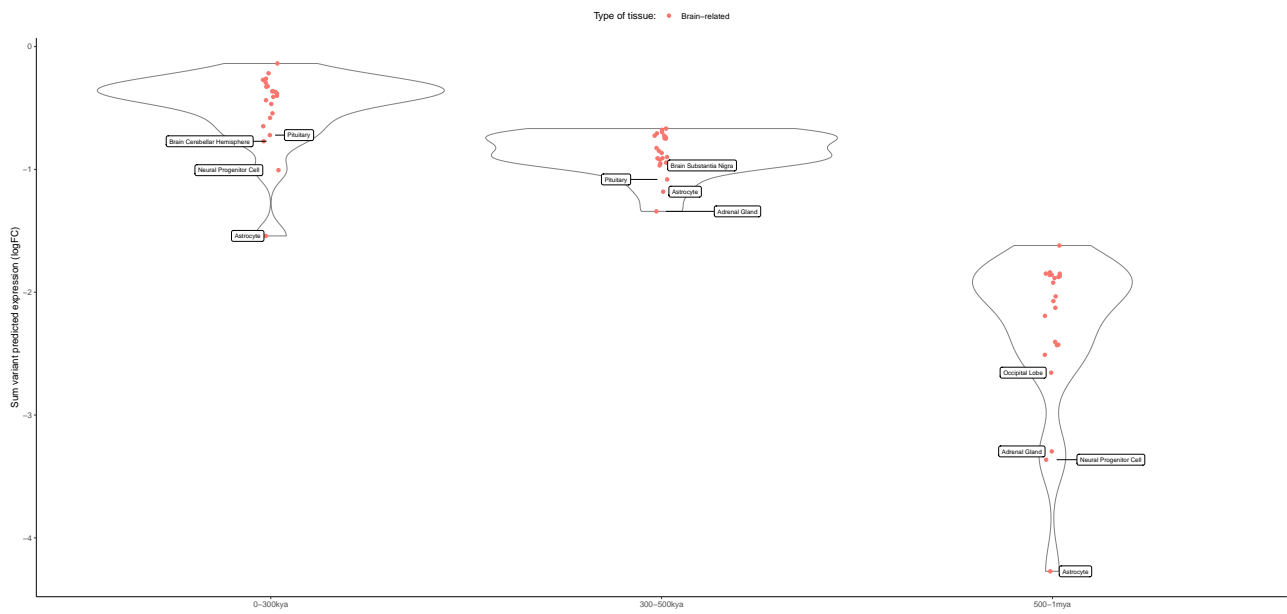


Figure 13: Violin plots per time window for 22 brain and brain-related tissues, showing the top-4 structures with strongest predicted downregulation.

References

- [1] McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell* **168**, 916–927.e12, DOI: 10.1016/j.cell.2017.01.038 (2017).
- [2] Peyrégne, S., Boyle, M. J., Dannemann, M. & Prüfer, K. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Research* **27**, 1563–1572, DOI: 10.1101/gr.219493.116 (2017).
- [3] Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710–722, DOI: 10.1126/science.1188021 (2010).
- [4] Zhou, H. *et al.* A Chronological Atlas of Natural Selection in the Human Genome during the Past Half-million Years. *bioRxiv* 018929, DOI: 10.1101/018929 (2015).
- [5] Racimo, F. Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. *Genetics* **202**, 733–750, DOI: 10.1534/genetics.115.178095 (2016).
- [6] Schlebusch, C. M. *et al.* Khoe-San Genomes Reveal Unique Variation and Confirm the Deepest Population Divergence in Homo sapiens. *Molecular Biology and Evolution* **37**, 2944–2954, DOI: 10.1093/molbev/msaa140 (2020).
- [7] Kuhlwilm, M. & Boeckx, C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Scientific Reports* **9**, 8463, DOI: 10.1038/s41598-019-44877-x (2019).
- [8] Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464, DOI: 10.1126/science.aat8464 (2018).

List of publications

List of papers published during the PhD (first authorship in bold):

- [1] **Moriano, J.**, Leonardi, O. Vitriolo, A., Testa, G., & Boeckx, C. 2023. A multi-layered integrative analysis reveals a cholesterol metabolic program in outer radial glia with implications for human brain evolution. *bioRxiv*
doi:[10.1101/2023.06.23.546307](https://doi.org/10.1101/2023.06.23.546307)
- [2] **Andirkó, A., Moriano, J.**, Vitriolo, A., Kuhlilm, M., Testa, G., & Boeckx, C. 2022. Fine-grained temporal mapping of derived high-frequency variants supports the mosaic nature of the evolution of *Homo sapiens*. *Scientific Reports*
doi:[10.1038/s41598-022-13589-0](https://doi.org/10.1038/s41598-022-13589-0)
- [3] **Buisán, R., Moriano, J.**, Andirkó, A., & Boeckx, C. 2022. A brain region-specific expression profile for genes within large introgression deserts and under positive selection in *Homo sapiens*. *Frontiers in Cell and Developmental Biology*
doi:[10.3389/fcell.2022.824740](https://doi.org/10.3389/fcell.2022.824740)
- [4] **Moriano, J., Martínez-Gil, N.**, Andirkó, A, Balcells, S., Grinberg, D. & Boeckx, C. 2021. Human-derived alleles in *SOST* and *RUNX2* 3'UTRs cause differential regulation in a bone cell-line model. *bioRxiv*
doi:[10.1101/2021.04.21.440797](https://doi.org/10.1101/2021.04.21.440797)
- [5] **Moriano, J.** & Boeckx, C. 2020. Modern human changes in regulatory regions implicated in cortical development. *BMC Genomics*
doi:[10.1186/s12864-020-6706-x](https://doi.org/10.1186/s12864-020-6706-x).