UNIVERSITAT DE BARCELONA

Modern brains and bones: genomic analysis of derived Homo sapiens traits

Alejandro Gregorio Munoz Andirkó

Modern brains and bones: genomic analysis of derived *Homo sapiens* traits

Alejandro Gregorio Muñoz Andirkó

Thesis submitted to the

University of Barcelona

in partial fulfillment of the requirements of the degree of

Doctor per la UB

Cognitive Science and Language PhD program

Under the supervision and tutorization of

Dr. Cedric Boeckx

(ICREA/University of Barcelona)

UNIVERSITAT DE BARCELONA

July, 2021

This page intentionally left blank

Acknowledgements

I have the joy of having many people to thank.

First of all, I'd like to thank Cedric, for introducing me to the academic world, and for believing since the beginning that this is my place. I've learned much under his supervision, and he set me on a path to learn much more. I've rarely met someone so diligent; I always knew that I could ask him anything and he would at the very least point me in the right direction. He has revised countless manuscripts and plots, several times and always almost immediately, and always managed to point out how to improve the reasoning or the clarity of the ideas. Most of this thesis is as much mine as it is his.

I'd like to thank all my CBL peers, with whom I've shared years of work, but also friendship. In no particular order: Tom, Juan, Steffi, Pedro. My best memories of this period are with you, and you have been a constant source of inspiration and support. I admire each one of you, both as friends and as researchers - you embody a kind of never-ending curiosity and intellectual honesty that is very rare and precious. I also would like to extend my thanks to everyone who has passed by the office for shorter periods: Sara, Mireia, Lucia. You are all great.

I'd like to thank my father and mother, for always pushing me to make the best of myself and giving me the liberty to take the many turns that have led me to this point. Their economic support in very difficult circumstances was what made moving to Barcelona possible in the first place, including the excruciating months between the finish of my MA and the start of the PhD contrat. I should include also their support during the global pandemic and lockdowns. I want

also to thank also my grandfather: *nélkuled se lett volna lehetséges.*

There are other people who have made this possible and provided constant support: Marcelo, Eli and Mónica, both Manus and the rest of my family in Murcia, who played a big role in shaping who I am today.

I thank as well every person who has provided input, ideas, work and support at some point, and all the people who have kindly given me some of their time in conferences, courses and outside academic settings. Here's a definitively non-exhaustive list: Martin Kuhlwilm, Dan Dediu, Hannes Schroeder, Raül Buisán, Marcel Ruland, Iza Romanowska, Bill Thompson, Cory Stade, Marcin Kierczak, Lianne Scholtens, Alessandro Vitriolo and everyone else in the Giuseppe Testa lab, the genetics department of the UB, anonymous reviewers (kind and not) for crucial contributions, and all the strangers on the internet who develop and maintain, sometimes in their free time, the countless packages and libraries I've used in my work (I've tried to cite you as much as I could!). I'd like also to thank Mariona Taulé for all her help battling bureaucracy, and for always being kind to us. I also learned much during my stay on the CCiL master course - I want to thank all the people that made the same journey with me. Even if I haven't see most of you since then, I think of every one of you often. Lastly, I'd like to send a special thank you to Camille Fauth, from the Phonetics department of the University of Strasbourg – your class on the biology of language in 2014 during my Erasmus was the start of the path that led to this thesis.

But my greatest, deepest thank you is to Cari, for her endless patience, love, and support despite the circumstances imposed by this weird job that is academic research. You are more than I could have ever asked for, and you more than anyone know what has gone into writing this thesis. Thank you so much for all these years.

Abstract

Over the last years we have come to realize that the history of our species is not as simple as we once thought. We are now faced with complex demographic scenarios, events of admixture with human species and the realization that our differences with Neanderthals might be of gradient, not quality. Knowing all of this, we have to face the question of what did in fact change since the emergence of our species, to what extent, when did it happened in our evolution, and what consequences those changes bring for complex traits. This thesis tries to contribute answers to these questions, either through *in silico* analysis of open, publicly available genomic databases, or, when possible, *in vitro*, in collaboration with other research groups.

Chapter 1 outlines the questions that are treated in this thesis and summarizes the results and arguments of each chapter. First, I present an overview of the current understanding of human evolution since the split between the Neanderthal/Denisovan lineage and the *Homo sapiens* branch. Then, I show what kind of core questions arise from the state of the field, namely what can be done to explore the relationship between genotype and phenotype over evolutionary time in humans. I follow with a succinct summary of the main methods and results of each chapter, showing how they try to answer each of these questions, as well as a final note on challenges and future steps.

Chapter 2 explores the effects of *Homo sapiens* variants in gene expression in specific brain tissues, using the GTEx consortium eQTL database. We highlight the effect of regulation in human evolution, present contrasts with previous literature regarding directional gene regulation, and

presents genes that correlate with brain volume GWAS signals.

Chapter 3 presents the result of our collaboration with the team of Dr. Giuseppe Testa, showing how neurodevelopmental disease modeling can inform our understanding of human evolution. We show how *BAZ1B*, a gene implicated in Williams-Beuren syndrome, regulates neural crest stem cell induction and migration. Our work suggests that the *BAZ1B* regulatory network has undergone changes in human evolution, shaping the facial morphology of our species and validating previously-standing hypothesis about the biology of craniofacial morphology in humans.

Chapter 4 applies an open large-scale database of allele age estimations (GEVA) to *Homo sapiens* specific variants in order to map species-specific genetic variation over time. We also apply gene expression predictions via machine learning and other time-sensitive genomic analysis, and present genes that have undergone changes in specific windows of human evolution.

Resum

En los últimos años nos hemos dado cuenta de que la historia de nuestra especie no es tan simple como una vez pensamos. Nos enfrentamos ahora a escenarios demographicos complejos, eventos de admixtura en el pasado con otras especies humanas y la realización de que nuestras diferencias con los Neandertales son de grado, no de cualidad. Sabiendo esto, tenemos que preguntarnos qué cambios tuvieron lugar desde la aparición de nuestra especie, hasta qué punto el fenotipo cambió, cuándo ocurrió en nuestra historia evolutiva y cuales son las consequences de esos cambios para rasgos complejos. Esta tesis intenta contribuir respuestas a estas preguntas a través de análisis *in silico* de bases de datos genomicas públicas, o, cuando es posible, *in vitro*.

El Capítulo 1 delinea las preguntas que trato en esta tesis y resume los argumentos y resultados de cada capítulo. Primero determino las claves de nuestro conocimiento actual en evolución humana desde la separación entre el lineaje Neanderthal/Denisovano y la rama *Homo sapiens*. Después, muestro qué tipo de preguntas centrales surgen del estado del campo; específicamente, qué puede hacerse para explorar la relación entre genotipo y fenotipo en tiempo evolutivo en humanos. A continuación presento un resumen de los métodos y resultados de cada capítulo, mostrando cómo intentan resolver cada una de estas preguntas, así como una última nota sobre pasos futuros y retos.

El Capítulo 2 explora el efecto de variantes *Homo sapiens* en expressión genética en tejidos cerebrales específicos, usando la base de datos de eQTL del consorcio GTEx. Los resultados resaltan el efecto de la regulación genómica en evolución humana, presenta contrastes con literatura ante-

rior respecto a la direccionalidad de regulación genómica, y presentan genes que se correlacionan con señales de volumen cerebral en GWAS.

El capítulo 3 presenta los resultados de nuestra colaboración con el grupo del Dr. Giuseppe Testa, mostrando cómo el modelaje de desórdener neurodevelomentales puede informar nuestra comprensión de la evolución humana. En este capítulo demostramos cómo *BAZ1B*, un gen implicado en el síndrome de Williams-Beuren, regula la inducción y migración de células madre de la cresta neuronal. Nuestro trabajo sugiere que la red reguladora de *BAZ1B* ha cambiado a lo largo de la evolución humana, cincelando la morfología facial de nuestra especie tal y como había sido predicho en la literatura anterior.

El capítulo 4 aplica una base de datos de estimaciones de edad de variantes (GEVA) a variantes specíficas de *Homo sapiens* para localizar en el tiempo variación específica de nuestra especie en el tiempo. También aplicamos predicciones de expresión genética a través de Machine Learning y otros análisis, y presentamos genes que han cambiado en ventanas específicas de nuestra historia evolutiva.

Contents

This page intentionally left blank

1 Introduction

1.1 The paradox of human uniqueness

There was a time when we thought *Homo sapiens* was unique; that explaining what we are as a species was as easy as pointing at the one factor that made us human, or at the one place were we emerged. This view is now unsustainable [1, 2] – and with this rebuttal comes the challenge of building a body of knowledge that fully acknowledges the complexity of our past.

Over the last decades we have discovered that our species and other closely related extinct humans, in particular Neanderthals, are not so different. We are no longer the only human species to create art [3], be it in form of engravings [4], ochre use [5] or body ornaments [6]. Neanderthals are also credited with creating complex technology previously thought to be unique of *Homo sapiens* [7, 8]. In the past, these findings used to be part of a single cohesive package, sometimes named 'behavioral modernity', but now this term is no longer useful to describe the gradual emergence of cultural and technological practices, no longer the product of a single species, time or place [9]. A new paradigm of what constitutes being human is emerging, and even the line that defines what is a distinct human species is more blurred than ever considering the accumulating evidence that there have been multiple events of introgression in the past between early *Homo sapiens* and Neanderthals and Denisovans [1, 10].

And yet, paradoxically, there are still many things that are not exactly equivalent in the biology of our species relative to the Neanderthal/Denisovan lineage. While many cultural practices seem to have been shared between species, claims to similarity do not necessarily entail total equivalence; that is, some aspects of *Homo sapiens* (and of Neanderthals and Denisovans) are still species-specific , as evidenced by the fossil record and genetic evidence. Plausibly, this means that any difference associated with cognition is likely just a matter of subtle degree, in the same way humans and other animals are not differentiated by any cognitive quality but rather by enhanced capabilities in each domain [11]: any potential differences between Neanderthals and extant humans might follow this pattern, once we clarify them. But we haven't yet: the matter of cognitive differences between human species remains a topic of heated discussion, specially when considering human-specific traits such as the faculty of language [12] or even cognitive pathologies such as Alzheimer's disease [13].

To start tracing back such *sapiens* or human-specific traits in time we have to first be aware of the currently settled differences between *Homo sapiens* and the Neanderthal and Denisovan lineage. While there have been other derived traits that differentiate these two human lineages (such as the immune system response [14, 15]), for the purposes of this thesis I will focus on two aspects of human biology: differences in brain tissue, for their obvious implications in cognition, and the genetics of bone morphology, for reasons detailed in [16] and related to the complex interplay between brain morphology and braincase recently explored in human evolution research ([17–19]). Note that given the history of racist misinterpretations of the idea of an interaction between facial morphology and cognition, I include a cautionary note on this topic at the end of Section 1.3.2.

In terms of bone anatomy, the differences between extant humans and other close branches of human species can be roughly summarized in differences in the morphology of the craniofacial complex, bone mineral density and rib cage shape. *Homo sapiens* have overall less protruding faces, the product of developmental trajectories of bone deposition and remodelling in the nasal and maxillary area and the supraorbital torus [20–22]. These changes are already present in an early *Homo sapiens* fossil, found in Jebel Irhoud and dated around 300 thousand years ago [23], though decoupled from other derived *sapiens* traits of the skull such as a overall rounder cranium. Neanderthals also had overall higher bone density and anatomical differences in ribcage width (likely product of differences in the respiratory system [24]) and the lumbo-pelvic complex [24, 25]. While the fossil record of Denisovans is notoriously scarce, a reconstruction from Denisovan DNA methylation maps suggests that most of the traits that characterize Neanderthals relative to *Homo sapiens* are also present in Denisovans [26]. The craniofacial bone complex is affected by multiple factors, including dietary modifications during the Neolithic [27], the interaction of bone resorption and deposition patterns with facial muscles, and Neural Crest cells early ontogeny [28], among others. It's worth noting that parts of the face have diverse degrees of derivation in *Homo sapiens*, and sometimes the history of facial bone remodelling is convoluted [21].

There are also differences in brain ontogeny and morphology between the *Homo sapiens* and Neanderthal branches, as evidenced prominently by endocast simulations using the fossil record. These kind of geometric morphometric analysis show that the skulls of *Homo sapiens* and Neanderthals have species-specific developmental trajectories [18, 19, 29]. *Homo sapiens* are characterized by a globular, rounder braincase, as opposed to the typically elongated endocasts of Neanderthals and early *sapiens* individuals such as the Jebel Irhoud fossil, a fossil that displays a modern-like facial morphology [23] as discussed above. A globular shape of the brain likely emerges from differential development trajectories of the cerebellum, the lateral parietotemporal and occipital areas [17–19, 29]. The effects of introgressed variants from Neanderthals in modern populations seem to confirm these differential brain trajectories, as Neanderthal alleles affect globularity [17]; introgressed variants are also downregulated in the cerebellum in extant humans [30].

The occipital, parietal, and temporal lobes, as well as the cerebellum, are related to cognitive abilities such as memory, attention, learning and speech [16, 17, 19, 31], though to what extent cognitive domains affected by subtle changes in species-specific brain tissue morphology is currently unknown. Additionally, other subcortical tissues that are not reflected in endocasts might have undergone species-specific variation that remains undetected. Such questions can be answered from the finer-grained perspective of paleogenomic analysis, aided by the recovered DNA of extinct humans (such as the high-coverage sequences of various Neanderthals and a Denisovan [32–34]).

Paleogenomic analysis supports the idea that there were differences in brain tissue, from the *Homo sapiens*-specific globularization process to findings at the cell level. For example, a catalog of species-specific genetic differences [35] showed that there are *Homo sapiens*-specific variants in genes associated to brain growth and cell division, and that some of these genes have fallen within genomic regions under positive selection since the split with the Neanderthal/Denisovan lineage [36]. There are also regions in the *sapiens* genetic pool that, despite multiple and semi-continuous admixture events across species [10], have resisted the introduction of genetic variation from extinct humans [37–40]. These areas seem to have endured purging selection against Neanderthal introgression [41], though other reasons such as population bottlenecks or drift have been proposed with various degrees of confidence [10].

Again, the question of how much genes in deserts of introgression contribute to cognitive abilities or tissue development is ambiguous. Fueling the dispute, among these genes there are some that are known to play an important role in language and neural development, including *FOXP2* or genes of the *ROBO* family [10, 42]. Variation in *FOXP2* was though to have been positively selected in *Homo sapiens* in early studies. Such claims have been refuted [43], but there are still other genomic regions in the deserts of introgression that have been putatively selected [42].

At heart, the classic question 'what makes us human?' encounters a paradox in these findings. The evidence from archaeological remains, genomics and other disciplines in-between shows that there are derived traits in specific human species, but we have yet to solve when, how and which traits changed over our evolutionary history, specially considering the similarities between our species and other extinct humans. These basic questions are crucial if we want to clarify the complex demographic events in the *Homo sapiens* pre-Out-of-Africa deep past [2, 23].

Early attempts using paleogenomic evidence to understand how do *Homo sapiens* differ from other human species focused on the few fixed missense mutations that separate us from Neanderthals and Denisovans [44]. While this perspective still finds its way into cutting-edge experimental settings, such as brain organoids edited by CRISPR-Cas9 technology [45], most reconstructions of the evolution of cognitive traits have instead acknowledged the effects of tissue and cell-specific expression and genetic regulation [22, 29, 46–50]. Only through an accurate picture of the effects of species-specific genetic variation can we effectively start mapping cognitive skills in an evolutionary scale – i.e., to clarify trait phylogenies, including the history of the emergence of our language, requires to solve first to what degree we are similar to other human species.

1.2 A triple problem

It's here, at the intersection between human evolution, Genetics and the Cognitive Sciences, where we are faced with a triple problem. First, mapping genotype and phenotype in modern humans in order to infer cognitive traits from molecular changes is a formidable quest on its own. Second, we have to be able to integrate these methods with ancient DNA sequences, where only predictive methods can provide information that is not obtainable (such as gene expression data), all the while taking into account that only a handful of high-quality sequences are available. And, third, we have to be able to reconcile any information we derive from this form of 'molecular archaeology' with the traditional sources of information in human evolution: the fossil and technological record.

The intersection between these three problems generates a number of methodological questions, namely: How do we map variation in ancient DNA to inferred phenotypical changes? How do we integrate the new information obtained through genomic methods with the patchy, complex nature of the *Homo sapiens* deep history? Are there things that can be inferred from genetic sequences that are not possible to detect by other means, such as endocasts? For example, have any subcortical regions been less represented due to the lack of endocast imprinting, considering how tied some of them are to the cerebellum [51]? How can evolutionary data inform the biology of modern-day traits (and vice versa)? Can modern pathologies inform evolutionary information (as proposed by [13], among others)?

The spirit of the work presented in this thesis follows, at its core, these questions. There is no way to answer any of these without involving an overarching body of disciplines, ranging from bioinformatic analysis that profit from variation data in current humans (in the line of [30] or [17]) to, when possible, the integration with *in vitro* approaches. The methods in this thesis fluctuate between my work from a computational perspective and the integration of the methods here presented with the work of collaborators in laboratory settings (such as in [49, 52]).

Harnessing the diversity of our species as reflected in genomic resources for evolutionary data can involve, for example, studying the variants that are specific to our species relative to Neanderthals and Denisovans. This approach has remained so far a less explored path compared to the study of introgressed data (though not totally ignored, as evidenced by [22, 35]). In a context where ideally the functionality of variation is validate through often economic and effort-wise expensive *in vitro* methods, we might also find that we have to prioritize specific genomic targets of study. Luckily, there are ways to prioritize which *Homo sapiens* genomic regions are of special interest, such as studying the effects of variation under positive selection and in deserts of introgression, those genes known to be at the center of large regulatory networks, or focusing on high-frequency *sapiens*-specific variants that are shared across most modern humans. This labor of prioritization has been a constant over the work included in this thesis [52–54], including those projects were I've played a more secondary role [42, 49].

1.3 Thesis outline

Each of the chapters of this thesis addresses a specific question sketched in the previous sections. I provide here a brief summary of each one of them, which are presented as either published articles (chapter 3) or in their current form (chapters 2 and 4, currently under peer-review). I also discuss briefly other articles that touch on the topics explored here and where I am a coauthor [42, 49, 55].

1.3.1 The effect of regulatory derived variants on brain regions

Chapter 2 deals with genetic regulation as a driving force in the evolution of *Homo sapiens*. There have been multiple recent attempts at understanding gene regulation in the context of evolution, as it likely plays a key role in human speciation (see [22, 26, 30, 46–48, 56, 57]). In particular, eQTL mapping was considered in at least one of these studies [22], but the method was discarded because it requires polymorphism data and thus is not suitable to understand fixed variants separating *sapiens* from other human species, the original intent of the authors. A similar approach using Allele-Specific Expression [30] found that Neanderthal alleles are downregulated in the cerebellum and striatum; however, the mirror question of whether *Homo sapiens*-specific variants had a particular effect in gene expression in the brain was left unanswered. While the effect of totally fixed variants can't be derived from eQTL mapping, it is possible to obtain the effects of alleles in very high frequencies across modern human populations.

Using the GTEx database, which includes 15 region-specific brain-related samples, we explored the effects of genetic regulation by these almost-fixed variants (at an arbitrarily chosen allele frequency of $\geq$ 90% across human metapopulations [35]). Our hypotheses were that: i) as predicted before in the literature [32, 36, 56], gene regulation detected by eQTL mapping would fall significantly more often within genomic windows under positive selection; ii) that we would find upregulation of high-frequency *H. sapiens*-specific variants in the cerebellum, mirroring the effects of the introgressed alleles found in [30], as well as on other derived areas of the brain (such as the parietal lobes). Additionally, while we didn't have specific predictions for this, eQTL mapping has the advantage of capturing gene expression variation in other potentially underlooked brain regions, most prominently subcortical regions that are not capture by endocast studies.

A permutation test with two independent positive selection studies [36, 58] did find that almost-fixed, current human alleles are more likely to lie in genomic regions under putative positive selection than in random regions of the genome, confirming our suspicions. We also found that 5'UTR variants are overrepresented in derived eQTL compared to a control. Regarding directional regulation, we did not find a mirror effect of the results in [30] when controlling for linkage disequilibrium, though we did see a statistically significant tendency to upregulation in both derived alleles and control sets when linkage disequilibrium was not accounted for. As for specific tissues, we did find that cerebellum and pituitary accumulate more variants than expected by chance (compared to a control set), partially confirming another of our initial hypotheses. Finally, a two sample Mendelian Randomization analysis and colocalization test between eQTL top hits and 10 brain volume GWAS found signal correlation in genes involved in neurodevelopment.

The article overall reinforces the idea that genetic regulation is core to our understanding of human evolution. We also discuss several methodological limitations. First, eQTL mapping from bulk-tissue RNA sequencing has a very low probability of finding causal variants [59]. Second, and assuming that the top derived eQTL per gene is causal, colocalization did not find a causal relationship with GWAS, despite the Wald ratio mendelian randomization correlations.

In essence, this might mean that either derived variants are acting in conjunction with previously-existing variants to shape the regulatory landscape that lead to the *Homo sapiens* phenotype, or that we are detecting instances of pleiotropy, as expected from the high phenotypic heterogeneity of the brain [60] and as is usual in bulk-tissue RNA sequencing, which does not distinguish between cell types.

1.3.2 Williams-Beuren syndrome, *BAZ1B* and the human face

Chapter 3 focuses on the role of the regulatory network of a gene, *BAZ1B*, in shaping the human face. This work is the fruit of a collaboration with the group of Dr. Giuseppe Testa, and the chapter reflects our joint publication in *Science Advances*.

BAZ1B is a chromatin regulator that lies in a genomic region (7q11.23) that, when deleted, causes a neurodevelopmental disorder called Williams-Beuren (WB) syndrome. This same region can also be duplicated, causing an autism-like syndrome (7dupASD, or WB region duplication syndrome) [61, 62].

WB syndrome causes a particular craniofacial profile that mirrors some aspect of the evolutionary trajectory of the *Homo sapiens* face [63]. WB also has prominent effects in social cognition: patients with WB syndrome are typically non-aggressive, prone to conversation (their verbal competencies are relatively well conserved) and overall overtly trusting of others, including complete strangers [61]. Neural Crest cells migration and induction, phenomena that underlie the characteristic overtly retracted facial bone morphology of WB patients[61], are regulated by BAZ1B [52]. Neural Crest cells are at the center of a previously proposed hypothesis in human evolution, one that stresses the parallels between human evolution and domestication, another biological phenomenon associated with neural crest cell ontogeny [63].

As humans have undergone both changes in their social cognition [64] and a retraction of the face [21], [63] proposed that there might be parallels between the evolution of social cognition, craniofacial development and neurochristopathies (i.e., disruptions in Neural Crest cell development) in *Homo sapiens*. Some neurodevelopmental disorders, such as WB syndrome, are considered neurochristopathies [52]; their underlying genetic makeup would thus serve as an experimental point of entry for the proposed parallels with domestication-like processes. WB syndrome is unique among neurochistopathies for the particular cognitive profile it is associated with, and its defined genetic causes. On top of that, WB was already associated with domestication-related genes. Some of the genes identified to be associated to domestication in domesticated foxes and dogs are paralogs of those in the genetic window of WB syndrome [65, 66].

Under this hypothesis as formulated by [63], neural crest cells should be associated to the evolution of facial morphology of *Homo sapiens* relative to Neanderthal, Denisovans and other species such as *Homo erectus* preceding the last common ancestor between *H. sapiens* and other human species [64]. Our article on BAZ1B was the first one to provide evidence that this was the case for one of the genes involved in Williams-Beuren syndrome, as predicted in [63, 64]. While the effect of WB-associated genes in the evolution of cognitive traits such as decreased reactive agression (as proposed in [63]) still require empirical testing and separation from other mechanistic explanations [67, 68], it's the first time that the so called 'human self-domestication' hypothesis has been explored *in vitro* (though previous attempts had tried to do so *in silico* [69]). Our work on *BAZ1B* is an example of how neurodevelopmental disorder modelling can be useful to inform our understanding of specific complex traits over evolution.

While the theory originally proposed by [63] was called the theory of 'human self-domestication' [63], the concept is not without problems: the neural crest mechanism for domestication is not universal across domesticated species, and what we conceptualize as "domestication syndrome", or the host of phenotypes accompanied by domestication is i) not consistently producing the same traits in different species, and ii) probably overlapping with other complex factors (as discussed in detail in [67, 70]). It's also worth noting is that the term 'domestication' is prone to racism-motivated oversimplification due to the cultural charge of domestication as a directed biological process [67]. Despite the title of the article [52], I want to note that I fully acknowledge that 'human self-domestication' is at the moment a muddy concept that will probably need reconceptualization. We have to provide a more detailed clarification of the role of Neural Crest cells in domestication-like processes, and, when applied to human evolution, integrate this understanding with their specific role in human evolution (in consonance with [67]'s assessment of the term). Neural crest cells play a role in many other biological phenomena, such as normal variation resulting in craniofacial divergence across many species [28]; any further empirical validation of the human domestication-like hypothesis would require pinpointing a chronology of these during phylogeny and ontogeny in other species, specially in contrast with other factors that might shape craniofacial morphology such as dietary changes [67].

To avoid confusion, I also emphasise that this particular article does not pretend to establish links between craniofacial genetics and any psychological traits in the sense that past pseudo-scientific disciplines (prominently, Gall's phrenology) did, prominently in racial terms. It should be clear to this day that cranial and facial bone genetics have a high degree of interplay, but that these do not entail psychological correlates in *Homo sapiens*. This was recently revisited in [71], an exhaustive empirical study on the genetics of craniofacial morphology that found no co-inheritance patterns with an exhaustive list of cognitive traits.

Thus, whether cognitive abilities are associated with craniofacial genetics at the species level in humans (that is, between Denisovans, Neanderthals and *sapiens*) is a different question by virtue of the focus on the evolutionary scale. Particularly, the questions asked in Chapter 3 are more akin to the interrogation of differences between domestic and wild canines or *Pan paniscus* and *Pan Troglodytes* [64]. Due to the kind of connotations that the term 'domestication' brings and the aforementioned reasons, to the day of writing of this thesis I favour the much more neutral 'mild neurocristopathy' term for this hypothesis, which has the advantage of focusing on the specific biological mechanism underlying the domestication-like process – even if it is at the expense of separating it from other prominent mechanistic explanations of domestication-like processes, such as the glutamatergic signalling hypothesis [68, 69].

Not related to the 'mild neurocristopathy' hypothesis, but in consonance with the search of the genetics underlying bone morphology, I've also participated in a project led by Juan Moriano and Nuria Martínez-Gil [49] which aimed to test the effects of species-specific genetic variation in genes associated with bone growth. This work was motivated by the differences in bone structure and growth in Neanderthals and *Homo sapiens* reviewed at the beginning of this introduction [20–

22, 24, 25]. The results can be found the form of preprint in [49].

We show that *SOST* and *RUNX2* 3′UTRs cause differential regulation of bone growth in *Homo sapiens* and Neanderthals. *RUNX2* is a gene proposed to be under positive selection in *Homo sapiens* in early studies [72], and that is known to affect craniofacial formation [73] and specifically fontanelle closing developmental timing (a derived trait in our species [72]), while *SOST* is associated with sclerosteosis and van Buchem disease [74]. These results help explain a part of the evolutionary history of our craniofacial structure, complement our work with Dr. Giuseppe Testa and his team, and provide an evolutionarily-aware account of pathologies that affect bone ontogeny in modern populations.

1.3.3 A chronology of *Homo sapiens* variants

Chapter 4 presents our work in establishing a temporal distribution of *Homo sapiens*-specific alleles. As discussed before, the evolution of *Homo sapiens* has not been linear, but rather reticular [1, 2], in that early human populations before the Out-of-Africa event likely constituted different subpopulations with a diversity of traits and complex relationships between them. To unravel this complexity, we need to be able to locate the apparition of alleles in time.

We made use of a large-scale database of allele age estimates, GEVA [75], that uses a non-parametric coalescence method to infer the age of more than 45 million variants in the current human genetic pool. We discovered that there is a two mode distribution to *Homo sapiens*-specific alleles age estimates, roughly coinciding with a period of population dispersal and expansion around 100 thousand years ago (kya) and the period of split between our species and other human lineages [53]. We note also that the distribution itself goes back far more than the estimated time of divergence, specially when not taking into account frequency, for reasons discussed in the chapter. We applied this data to genomic regions of evolutionary interest, such as windows under putative positive selection, introgressed alleles from Neanderthals and Denisovans and deserts of introgression [36–38, 40].

We then divided alleles in temporal slices relevant to milestones in human evolution, such as the period of pre-Out-of-Africa *Homo sapiens* history between 300 and 500kya, for functional analysis. We applied time-sensitive gene ontology analysis and a machine learning gene expression predictor [76] to highlight those genes with a higher directional and absolute regulation in specific time windows. We find genes associated with cerebellar Purkinje neurons or the glutamate and dopaminergic system in different times of evolutionary history [53]. As a case example, we also assigned temporal estimates to variants lying in genes associated to BAZ1B expression, as determined previously in our work in Chapter 3. We found that the network of genes that shape the human face has arisen over a long evolutionary time, and that some of the alleles that shape our face were present before the emergence of our species. These alleles might have gained function over the course of evolution, or else contributed to morphological traits that are not *Homo sapiens*-specific. This last possiblity was already formulated on the basis of the fossil record in [21] and builds upon previous results on the evolutionary trajectory of the human face ([52], presented here as Chapter 3).

Our work in Chapter 4 serves as an example of how open, large-scale databases (such as GEVA) can be harnessed to test evolutionary questions. We also reinforce the idea that any study aiming to understand the function of a variant in human evolution has to be mindful of this kind of temporal estimates. Once more, this evidences the necessary collaboration of different subdisciplines to obtain an integral view of human evolution.

There are two separate projects that complement chapter 4 but were not included in the body of this thesis.

First, there is [57], led by Raül Buisán and Juan Moriano, where we looked at the effects of genes in deserts of introgression in gene expression over development in brain regions. The results can be read at [42]. While Chapter 4 deals with mapping age estimates to genetic variation, we introduce in this project an additional developmental perspective that is key if we want to pinpoint how and when did Neanderthal/Denisovan and *Homo sapiens* brain growth trajectories diverge [17, 18].

To this end, we used a public transcriptomic database that includes human brain developmental tags [77]. We aimed at bringing a finer perspective on the developmental aspect of genetic variation in brain tissues, specially in the context of genes that have resisted Neanderthal and Denisovan introgression. Introducing a developmental perspective was aimed at pinpointing if previously suggested differences in ontogeny across human species could be traced back to the effects of genes in deserts of introgression [29]. Overall, we show that in the set of genes under positive selection within introgression deserts the cerebellum and the striatum show the most divergent transcriptomic profiles at prenatal stages [42].

Second, there is [55], led by Marcel Ruland. We have discussed above that the work in Chapter 4 reflects a large population expansion of *Homo sapiens*. The effects of this expansion in cultural traditions are thoroughly discussed in the archaeological and cultural evolution literature. [55] presents an agent-based model simulation of factors having affected language complexity in the past. The model in this work integrates these factors in proxies for cognitive capacity, demographic expansion and hostility. The results show that some of these factors follow each other closely, but that it's possible to discern the order and causal relationship between them, mirroring our current knowledge of human evolution. In general, this work advances an interdisciplinary approach towards the evolution of language as a complex phenomenon – one that integrates an evolutionary narrative that is not always taken into account in classic linguistic circles [78].

1.4 Challenges and future directions

There are inherent challenges to the kind of computational approaches to human evolution I've taken in most this thesis. First, we have a low number of high-coverage genomes from extinct humans. This might change in the future, but at the moment of writing of this thesis it has conditioned the certainty with which we can call a variant *Homo sapiens*-specific, as the full picture of extinct human genetic diversity is obscured. The consequences of this limitation to the conclusions of each project are discussed in the corresponding discussion sections.

Second, there are known problems of underrepresentation of African diversity in current genomes. Despite being the continent with more genetic variation, we have relatively less information from current populations of the African continent compared to other continents, particularly in terms of GWAS studies and DNA sequences. Capturing the full picture of the African genetic pool is paramount to understand both current variation in *Homo sapiens*, the genetics of complex traits, and demographic inferences of the deep past [2].

Having stated these limitations, there are highlights to extract from the body of work here presented, in the two lines of research covered: bone development and brain evolution.

In terms of bone development, my coauthors and I have built upon previous literature on genes that were considered to play a prominent role in bone development [49, 52]. We have used for this purpose an approach mindful of natural variation in *Homo sapiens*, particularly making use of genes associated with disorders affecting extant humans. Researchers can build upon our results on *BAZ1B* [52], *RUNX2* and *SOST* [49], such as in the understanding of the role of neural crest cells in other human traits or species. We have also opened a road for researchers to explore the role that hypothesized mild neurochistopathies might have in human evolution, providing the first partial validation of the hypothesis [63]. Methodologically, we also explore a really interdisciplinary use of neurodevelopmental disorders in human evolution. Our chapter on the timing of *sapiens*-specific alleles (4) also includes a timescale of alleles affecting BAZ1B regulation, effectively mapping both a functional role and the history of this gene in *Homo sapiens*.

As for brain evolution, we reinforce the role for *sapiens*-specific genetic regulation in Chapter 2. Various of my efforts, sometimes along with collaborators, (Chapter 2, 4 and [42], not included here) highlight some of the genes and mechanisms by which the cerebellum is a key divergent structure, consistently with previous literature [17, 18]. We also discuss the role of other candidate brain regions that might have changed in *Homo sapiens* evolution: [42] shows divergent profiles of the thalamus and striatum during ontogeny, while Chapter 2 highlights regulatory changes in the pituitary. At a finer perspective, in Chapter 4 we also provided a time-sensitive account of variation that highlights genes related to the glutamate and dopaminergic systems.

Integration with single-cell RNA-seq, at a finer resolution than bulk tissue sequencing, will be key in future efforts in this direction (as evidenced by [35, 57]). Integration with *in vitro* methods remains also important to functionally test the effects of specific variants. Much hope has been put on introducing variants from extinct human in brain organoids with gene editing techniques such as CRISPR-Cas9 (as in [45]), another potential testing ground, though researchers should be mindful of the problems of inferring species-specific differences from the functionality of a single variant outside its genomic context and typical developmental environment.

Additionally, and as discussed above, most of the work here presented has used the high-coverage aDNA sequences [32–34, 79], but over the last years fragmentary and low-coverage genomes of various places, ages, and species have emerged. Most of the work of this thesis has made use of [35] as a source of inspiration, but any future work that wants to infer past phenotypes should strive to integrate new samples, both high and low coverage, ideally in a scalable way to account for newly-emerging information.

While most of the questions formulated in this introduction remain largely unanswered, I hope that the work presented in this thesis provides at least a partial answer to some of the questions that arise from understanding how did exactly *Homo sapiens* came to be.

Bibliography

[1] Anders Bergström et al. "Origins of modern human ancestry". en. In: *Nature* 590.7845 (Feb. 2021), pp. 229–237. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03244-5.

[2] Eleanor M. L. Scerri, Lounès Chikhi, and Mark G. Thomas. "Beyond multiregional and simple out-of-Africa models of human evolution". en. In: *Nature Ecology & Evolution* 3.10 (Oct. 2019), pp. 1370–1372. ISSN: 2397-334X. DOI: 10.1038/s41559-019-0992-1.

[3] João Zilhão. "The Emergence of Ornaments and Art: An Archaeological Perspective on the Origins of "Behavioral Modernity"". en. In: *Journal of Archaeological Research* 15.1 (Mar. 2007), pp. 1–54. ISSN: 1573-7756. DOI: 10.1007/s10814-006-9008-1.

[4] Dirk L. Hoffmann et al. "Symbolic use of marine shells and mineral pigments by Iberian Neandertals 115,000 years ago". en. In: *Science Advances* 4.2 (Feb. 2018), eaar5255. ISSN: 2375-2548. DOI: 10.1126/sciadv.aar5255.

[5] Wil Roebroeks et al. "Use of red ochre by early Neandertals". en. In: *Proceedings of the National Academy of Sciences* 109.6 (Feb. 2012), pp. 1889–1894. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1112261109.

[6] Davorka Radovčić et al. "Evidence for Neandertal Jewelry: Modified White-Tailed Eagle Claws at Krapina". en. In: *PLOS ONE* 10.3 (Mar. 2015), e0119802. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0119802.

[7] John F. Hoffecker. “The complexity of Neanderthal technology”. en. In: *Proceedings of the National Academy of Sciences* 115.9 (Feb. 2018), pp. 1959–1961. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1800461115.

[8] James Blinkhorn et al. “Nubian Levallois technology associated with southernmost Neanderthals”. en. In: *Scientific Reports* 11.1 (Feb. 2021), p. 2869. ISSN: 2045-2322. DOI: 10.1038/s41598-021-82257-6.

[9] Marc Kissel and Agustín Fuentes. “‘Behavioral modernity’ as a process, not an event, in the human niche”. en. In: *Time and Mind* 11.2 (Apr. 2018), pp. 163–183. ISSN: 1751-696X, 1751-6978. DOI: 10.1080/1751696X.2018.1469230.

[10] Aaron B. Wolf and Joshua M. Akey. “Outstanding questions in the study of archaic hominin admixture”. eng. In: *PLoS genetics* 14.5 (May 2018), e1007349. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007349.

[11] Kevin Laland and Amanda Seed. “Understanding Human Cognitive Uniqueness”. In: *Annual Review of Psychology* 72.1 (Jan. 2021), pp. 689–716. ISSN: 0066-4308. DOI: 10.1146/annurev-psych-062220-051256.

[12] Dan Dediu and Stephen C Levinson. “Neanderthal language revisited: not only us”. en. In: *Current Opinion in Behavioral Sciences* 21 (June 2018), pp. 49–55. ISSN: 23521546. DOI: 10.1016/j.cobeha.2018.01.001.

[13] Enric Bufill, Rafael Blesa, and Jordi Augustí. “Alzheimer’s disease: an evolutionary approach”. eng. In: *Journal of anthropological sciences = Rivista di antropologia: JASS* 91 (2013), pp. 135–157. ISSN: 2037-0644. DOI: 10.4436/jass.91001.

[14] Hélène Quach et al. “Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations”. English. In: *Cell* 167.3 (Oct. 2016), 643–656.e17. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.09.024.

[15] Yohann Nédélec et al. "Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens". English. In: *Cell* 167.3 (Oct. 2016), 657–669.e21. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2016.09.025.

[16] Cedric Boeckx. "The language-ready head: Evolutionary considerations". en. In: *Psychonomic Bulletin & Review* 24.1 (Feb. 2017), pp. 194–199. ISSN: 1531-5320. DOI: 10.3758/s13423-016-1087-5.

[17] Philipp Gunz et al. "Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity". eng. In: *Current biology: CB* 29.1 (Jan. 2019), 120–127.e5. ISSN: 1879-0445. DOI: 10.1016/j.cub.2018.10.065.

[18] Simon Neubauer, Jean-Jacques Hublin, and Philipp Gunz. "The evolution of modern human brain shape". en. In: *Science Advances* 4.1 (Jan. 2018), eaao5961. ISSN: 2375-2548. DOI: 10.1126/sciadv.aao5961.

[19] Emiliano Bruner. "Human Paleoneurology and the Evolution of the Parietal Cortex". In: *Brain, Behavior and Evolution* 91 (2018), pp. 136–147. ISSN: 0006-8977, 1421-9743. DOI: 10.1159/000488889.

[20] Rodrigo S. Lacruz et al. "Ontogeny of the maxilla in Neanderthals and their ancestors". en. In: *Nature Communications* 6.1 (Dec. 2015), p. 8996. ISSN: 2041-1723. DOI: 10.1038/ncomms9996.

[21] Rodrigo S. Lacruz et al. "The evolutionary history of the human face". en. In: *Nature Ecology & Evolution* 3.5 (May 2019), pp. 726–736. ISSN: 2397-334X. DOI: 10.1038/s41559-019-0865-7.

[22] David Gokhman et al. "Differential DNA methylation of vocal and facial anatomy genes in modern humans". en. In: *Nature Communications* 11.1 (Dec. 2020), p. 1189. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15020-6.

[23] Jean-Jacques Hublin et al. “New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens”. en. In: *Nature* 546.7657 (June 2017), pp. 289–292. ISSN: 1476-4687. DOI: 10.1038/nature22336.

[24] Daniel García-Martínez et al. “Early development of the Neanderthal ribcage reveals a different body shape at birth compared to modern humans”. en. In: *Science Advances* 6.41 (Oct. 2020), eabb4377. ISSN: 2375-2548. DOI: 10.1126/sciadv.abb4377.

[25] Asier Gómez-Olivencia et al. “3D virtual reconstruction of the Kebara 2 Neandertal thorax”. en. In: *Nature Communications* 9.1 (Oct. 2018), p. 4387. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06803-z.

[26] David Gokhman et al. “Reconstructing Denisovan Anatomy Using DNA Methylation Maps”. English. In: *Cell* 179.1 (Sept. 2019), 180–192.e10. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2019.08.035.

[27] Noreen von Cramon-Taubadel. “Global human mandibular variation reflects differences in agricultural and hunter-gatherer subsistence strategies”. en. In: *Proceedings of the National Academy of Sciences* 108.49 (Dec. 2011), pp. 19546–19551. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1113050108.

[28] Richard A. Schneider. “Neural crest and the origin of species-specific pattern”. In: *Genesis (New York, N.y. : 2000)* 56.6-7 (2018). ISSN: 1526-954X. DOI: 10.1002/dvg.23219.

[29] Takanori Kochiyama et al. “Reconstructing the Neanderthal brain using computational anatomy”. en. In: *Scientific Reports* 8.1 (Apr. 2018), p. 6296. ISSN: 2045-2322. DOI: 10.1038/s41598-018-24331-0.

[30] Rajiv C. McCoy, Jon Wakefield, and Joshua M. Akey. “Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression”. English. In: *Cell* 168.5 (Feb. 2017), 916–927.e12. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2017.01.038.

[31] Leonard F. Koziol et al. "Consensus Paper: The Cerebellum's Role in Movement and Cognition". en. In: *The Cerebellum* 13.1 (Feb. 2014), pp. 151–177. ISSN: 1473-4230. DOI: 10.1007/s12311-013-0511-x.

[32] Kay Prüfer et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". en. In: *Nature* 505.7481 (Jan. 2014), pp. 43–49. ISSN: 1476-4687. DOI: 10.1038/nature12886.

[33] Matthias Meyer et al. "A High-Coverage Genome Sequence from an Archaic Denisovan Individual". en. In: *Science* 338.6104 (Oct. 2012), pp. 222–226. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1224344.

[34] Kay Prüfer et al. "A high-coverage Neandertal genome from Vindija Cave in Croatia". en. In: *Science* 358.6363 (Nov. 2017), pp. 655–658. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aao1887.

[35] Martin Kuhlwilm and Cedric Boeckx. "A catalog of single nucleotide changes distinguishing modern humans from archaic hominins". en. In: *Scientific Reports* 9.1 (June 2019), p. 8463. ISSN: 2045-2322. DOI: 10.1038/s41598-019-44877-x.

[36] Stéphane Peyrégne et al. "Detecting ancient positive selection in humans using extended lineage sorting". en. In: *Genome Research* 27.9 (Sept. 2017), pp. 1563–1572. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.219493.116.

[37] Sriram Sankararaman et al. "The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans". English. In: *Current Biology* 26.9 (May 2016), pp. 1241–1247. ISSN: 0960-9822. DOI: 10.1016/j.cub.2016.03.037.

[38] Lu Chen et al. "Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals". English. In: *Cell* 180.4 (Feb. 2020), 677–687.e16. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2020.01.012.

[39] Laurits Skov et al. "The nature of Neanderthal introgression revealed by 27,566 Icelandic genomes". en. In: *Nature* 582.7810 (June 2020), pp. 78–83. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2225-9.

[40] Benjamin Vernot et al. "Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals". en. In: *Science* 352.6282 (Apr. 2016), pp. 235–239. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aad9416.

[41] Žiga Avsec et al. "Effective gene expression prediction from sequence by integrating long-range interactions". en. In: *bioRxiv* (Apr. 2021), p. 2021.04.07.438649. DOI: 10.1101/2021.04.07.438649.

[42] Raül Buisan et al. "A distinct expression profile in the cerebellum and striatum for genes under selection within introgression deserts". en. In: *bioRxiv* (Apr. 2021), p. 2021.03.26.437167. DOI: 10.1101/2021.03.26.437167.

[43] Elizabeth Grace Atkinson et al. "No Evidence for Recent Selection at FOXP2 among Diverse Human Populations". English. In: *Cell* 174.6 (Sept. 2018), 1424–1435.e15. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2018.06.048.

[44] Svante Pääbo. "The Human Condition—A Molecular Approach". English. In: *Cell* 157.1 (Mar. 2014), pp. 216–226. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2013.12.036.

[45] Cleber A. Trujillo et al. "Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment". en. In: *Science* 371.6530 (Feb. 2021). ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aax2537.

[46] Laura L. Colbran et al. "Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences". en. In: *Nature Ecology & Evolution* 3.11 (Nov. 2019), pp. 1598–1606. ISSN: 2397-334X. DOI: 10.1038/s41559-019-0996-x.

[47] Harlan R. Barker, Seppo Parkkila, and Martti E.E. Tolvanen. *Evolution is in the details: Regulatory differences in modern human and Neanderthal.* en. preprint. Genomics, Sept. 2020. DOI: 10.1101/2020.09.04.282749.

[48] Stephanie M Yan and Rajiv C McCoy. "Archaic hominin genomics provides a window into gene expression evolution". en. In: *Current Opinion in Genetics & Development* 62 (June 2020), pp. 44–49. ISSN: 0959437X. DOI: 10.1016/j.gde.2020.05.014.

[49] Juan Moriano et al. "Human-derived alleles in SOST and RUNX2 3UTRs cause differential regulation in a bone cell-line model". en. In: *bioRxiv* (Apr. 2021), p. 2021.04.21.440797. DOI: 10.1101/2021.04.21.440797.

[50] Guillaume Dumas, Simon Malesys, and Thomas Bourgeron. "Systematic detection of brain protein-coding genes under positive selection during primate evolution and their roles in cognition". en. In: *Genome Research* 31.3 (Mar. 2021), pp. 484–496. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.262113.120.

[51] F Middleton. "Basal ganglia and cerebellar loops: motor and cognitive circuits". In: *Brain Research Reviews* 31.2-3 (Mar. 2000), pp. 236–250. ISSN: 01650173. DOI: 10.1016/S0165-0173(99)00040-5.

[52] Matteo Zanella et al. "Dosage analysis of the 7q11.23 Williams region identifies BAZ1B as a major human gene patterning the modern human face and underlying self-domestication". en. In: *Science Advances* 5.12 (Dec. 2019), eaaw7908. ISSN: 2375-2548. DOI: 10.1126/sciadv.aaw7908.

[53] Alejandro Andirkó et al. "Fine-grained temporal mapping of derived high-frequency variants supports the mosaic nature of the evolution of Homo sapiens". en. In: *bioRxiv* (Jan. 2021), p. 2021.01.22.427608. DOI: 10.1101/2021.01.22.427608.

[54] Alejandro Andirkó and Cedric Boeckx. “Modern human alleles differentially regulate gene expression across brain regions: implications for brain evolution”. en. In: *bioRxiv* (Nov. 2020), p. 771816. DOI: 10.1101/771816.

[55] Marcel Ruland et al. “An Agent-based model of the gradual emergence of modern linguistic complexity”. en. In: *bioRxiv* (Nov. 2020), p. 2020.11.12.380683. DOI: 10.1101/2020.11.12.380683.

[56] Sven Weyer and Svante Pääbo. “Functional Analyses of Transcription Factor Binding Sites that Differ between Present-Day and Archaic Humans”. en. In: *Molecular Biology and Evolution* 33.2 (Feb. 2016), pp. 316–322. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/msv215.

[57] Juan Moriano and Cedric Boeckx. “Modern human changes in regulatory regions implicated in cortical development”. en. In: *BMC Genomics* 21.1 (Dec. 2020), p. 304. ISSN: 1471-2164. DOI: 10.1186/s12864-020-6706-x.

[58] Fernando Racimo. “Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation”. In: *Genetics* 202.2 (Feb. 2016), pp. 733–750. ISSN: 1943-2631. DOI: 10.1534/genetics.115.178095.

[59] François Aguet et al. “Genetic effects on gene expression across human tissues”. en. In: *Nature* 550.7675 (Oct. 2017), pp. 204–213. ISSN: 1476-4687. DOI: 10.1038/nature24277.

[60] Sarah E. Medland et al. “Ten years of enhancing neuro-imaging genetics through meta-analysis: An overview from the ENIGMA Genetics Working Group”. en. In: *Human Brain Mapping* n/a.n/a (). ISSN: 1097-0193. DOI: https://doi.org/10.1002/hbm.25311.

[61] Barbara R. Pober. “Williams–Beuren Syndrome”. en. In: *New England Journal of Medicine* 362.3 (Jan. 2010), pp. 239–252. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMra0903074.

[62] Stephan J. Sanders et al. “Multiple Recurrent De Novo CNVs, Including Duplications of the 7q11.23 Williams Syndrome Region, Are Strongly Associated with Autism”. English. In: *Neuron* 70.5 (June 2011), pp. 863–885. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2011.05.002.

[63] 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”. eng. In: *Genetics* 197.3 (July 2014), pp. 795–808. ISSN: 1943-2631. DOI: 10.1534/genetics.114.165423.

[64] Brian Hare. “Survival of the Friendliest: Homo sapiens Evolved via Selection for Prosociality”. eng. In: *Annual Review of Psychology* 68 (Jan. 2017), pp. 155–186. ISSN: 1545-2085. DOI: 10.1146/annurev-psych-010416-044201.

[65] Anna V. Kukekova et al. “Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours”. eng. In: *Nature Ecology & Evolution* 2.9 (Sept. 2018), pp. 1479–1491. ISSN: 2397-334X. DOI: 10.1038/s41559-018-0611-6.

[66] Bridgett M. vonHoldt et al. “Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs”. eng. In: *Science Advances* 3.7 (July 2017), e1700398. ISSN: 2375-2548. DOI: 10.1126/sciadv.1700398.

[67] Marcelo R. Sánchez-Villagra and Carel P. van Schaik. “Evaluating the self-domestication hypothesis of human evolution”. eng. In: *Evolutionary Anthropology* 28.3 (May 2019), pp. 133–143. ISSN: 1520-6505. DOI: 10.1002/evan.21777.

[68] Thomas O’Rourke and Cedric Boeckx. “Glutamate receptors in domestication and modern human evolution”. en. In: *Neuroscience & Biobehavioral Reviews* 108 (Jan. 2020), pp. 341–357. ISSN: 01497634. DOI: 10.1016/j.neubiorev.2019.10.004.

[69] Constantina Theofanopoulou et al. "Self-domestication in Homo sapiens: Insights from comparative genomics". en. In: *PLOS ONE* 12.10 (Oct. 2017). Ed. by Michael Klymkowsky, e0185306. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0185306.

[70] Marcelo R. Sánchez-Villagra, Madeleine Geiger, and Richard A. Schneider. "The taming of the neural crest: a developmental perspective on the origins of morphological covariation in domesticated mammals". In: *Royal Society Open Science* 3.6 (June 2016). ISSN: 2054-5703. DOI: 10.1098/rsos.160107.

[71] Sahin Naqvi et al. "Shared heritability of human face and brain shape". en. In: *Nature Genetics* (Apr. 2021), pp. 1–10. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00827-w.

[72] Martin Kuhlwilm, Armaity Davierwala, and Svante Pääbo. "Identification of putative target genes of the transcription factor RUNX2". eng. In: *PloS One* 8.12 (2013), e83218. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0083218.

[73] Toshihisa Komori. "Roles of Runx2 in Skeletal Development". eng. In: *Advances in Experimental Medicine and Biology* 962 (2017), pp. 83–93. ISSN: 0065-2598. DOI: 10.1007/978-981-10-3233-2_6.

[74] Jesus Delgado-Calle, Amy Y. Sato, and Teresita Bellido. "Role and mechanism of action of Sclerostin in bone". In: *Bone* 96 (Mar. 2017), pp. 29–37. ISSN: 8756-3282. DOI: 10.1016/j.bone.2016.10.007.

[75] Patrick K. Albers and Gil McVean. "Dating genomic variants and shared ancestry in population-scale sequencing data". en. In: *PLOS Biology* 18.1 (Jan. 2020), e3000586. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3000586.

[76] Jian Zhou et al. "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk". en. In: *Nature Genetics* 50.8 (Aug. 2018), pp. 1171–1179. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0160-6.

[77] Mingfeng Li et al. “Integrative functional genomic analysis of human brain development and neuropsychiatric risks”. en. In: *Science* 362.6420 (Dec. 2018). ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aat7615.

[78] Pedro Tiago Martins and Cedric Boeckx. “Language evolution and complexity considerations: The no half-Merge fallacy”. en. In: *PLOS Biology* 17.11 (Nov. 2019), e3000389. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3000389.

[79] Fabrizio Mafessoni et al. “A high-coverage Neandertal genome from Chagyrskaya Cave”. en. In: *Proceedings of the National Academy of Sciences* 117.26 (June 2020), pp. 15132–15136. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2004944117.

2 | The effect of regulatory derived variants on brain regions

Brain region-specific effects of nearly fixed sapiens-derived alleles

Alejandro Andirkó[1,2] and Cedric Boeckx[1,2,3,*]

[1]University of Barcelona
[2]University of Barcelona Institute of Complex Systems
[3]ICREA
[*]Corresponding author: cedric.boeckx@ub.edu

April 20, 2021

Abstract

The availability of high-coverage genomes of our extinct relatives, the Neanderthals and Denisovans, and the emergence of large, tissue-specific databases of modern human genetic variation, offer the possibility of probing the effects of derived alleles in specific tissues, such as the brain, and its specific regions. While previous research has explored the effects of introgressed variants in gene expression, the effects of *Homo sapiens*-specific gene expression variability are still understudied. Here we identify derived, *Homo sapiens*-specific high-frequency ($\geq$ 90%) alleles that are associated with differential gene expression across 15 brain structures derived from the GTEx database. We show that regulation by these derived variants targets regions under positive selection more often than expected by chance, and that high-frequency derived alleles lie in functional categories related to transcriptional regulation. Our results highlight the role of these variants in gene regulation in cerebellum and pituitary.

Keywords— Human evolution, brain, cis-eQTL, gene regulation

1 Introduction

Geometric morphometric analysis on endocasts (Gunz et al., 2010, Hublin et al., 2015, Neubauer et al., 2018, Pereira-Pedro et al., 2020, Kochiyama et al., 2018) have revealed significant differences between Neanderthal and *Homo sapiens* skulls that are most likely the result of differential growth of neural tissue. Specific brain regions such as the cerebellum, the parietal and temporal lobes have been hypothesized to have expanded in the *Homo sapiens* lineage, with potential consequences for the evolution and diversification of cognitive skills. Probing the nature of these consequences is challenging, but the availability of several high-quality Neanderthal and Denisovan genomes (Prüfer et al., 2014, 2017, Meyer et al., 2012, Mafessoni et al., 2020) has opened numerous research opportunities for studying the evolution of the *Homo sapiens* brain with unprecedented precision.

Efforts have been made to determine the molecular basis of species differences based on a small number of fixed missense mutations that are *Homo sapiens*-specific (Pääbo, 2014, Trujillo et al., 2021). However, evidence is rapidly emerging in favor of an important evolutionary role of regulatory variants, as originally proposed more than four decades ago (King and Wilson, 1975). For instance, regulatory variants are overrepresented in selective sweep scans to detect areas of the genome that have been significantly affected by natural selection after the split with Neanderthals (Peyrégne et al., 2017).

The increasingly important role of gene regulation in the evolution of *Homo sapiens* has led to the idea of connecting vast datasets of variation in genomic regulation to the genetic sequences obtained from extinct humans. For example, a major study (McCoy et al., 2017) explored the effects of Neanderthal and Denisovan introgressed variants in 44 tissues and found downregulaton by introgressed alleles in the brain, particularly in the cerebellum and the striatum. In a similar vein, another study (Gunz et al., 2019) examined the effects of extinct human introgression on brain and skull shape variability in a modern human population to determine which variants are associated with the globularized brain and skull that is characteristic of our lineage. In consonance with McCoy et al. (2017), the variants with the most salient effects were those found to affect the structure of the cerebellum and the striatum.

Building on these efforts, we decided to relate derived, modern-specific alleles found at very high frequency across modern populations to gene expression in the brain, in order to examine the effects of genetic variation relative to Neanderthals and Denisovans. To this end, we took advantage of a recent systematic review, (Kuhlwilm and Boeckx, 2019), which provides an exhaustive dataset of derived, *Homo sapiens*-specific alleles in modern human population. This dataset includes a subset of nearly-fixed ($\geq 90\%$) variants that can determine common trends in current human populations compared to other extinct human species.

To determine the predicted effect on gene expression of these alleles we exploited the GTEX database. The GTEx data consist of statistically significant allele effects on gene expression dosage in single tissues, obtained from tissues of adult individuals aged 20 to 60 (GTEx Consortium, 2017). By offering information about Expression Quantitative Trait Loci (cis-eQTLs) across tissues, the GTEx database forces us to think beyond variants that affect the structure and function of proteins, as well as to consider those that regulate gene expression.

While the important role genetic regulation in human evolution has been highlighted by previous studies (Gokhman et al., 2020, Colbran et al., 2019, Moriano and Boeckx, 2020), we find that species-specific variants above a high frequency threshold have a previously underexplored role in human brain evolution. We show that regions under putative positive selection are enriched in derived, high-frequency (HF) eQTLs, and that the pituitary and cerebellum have a significantly higher number of regulatory variability compared to other tissues and a control set. We also show that derived alleles tend to have a downregulating effect but only when linkage disequilibrium is not controlled for, a result that contrasts with previous research on introgressed variants (McCoy et al., 2017). Finally, we present a two sample Mendelian randomization analysis that correlates variability in genes related to neurodevelopment and brain volume GWASs.

2 Results

We extracted variation data from (Kuhlwilm and Boeckx, 2019), a dataset that determines *Homo sapiens* allele specificity using three high-coverage archaic human genomes available at the moment (the Altai and Vindija Neanderthals (Prüfer et al., 2014, 2017), and a Denisovan individual (Meyer et al., 2012)).

The variation data was crossed with the list of variants obtained with the GTEX significant cis-eQTL variants dataset to determine if the selected variants affect gene expression, focusing on 15 central nervous system-related tissues. The GTEx data consist of statistically significant allele effects on gene expression dosage in single tissues, obtained from brain samples of adult individuals aged 20 to 60 (GTEx Consortium, 2017). The resulting dataset is composed of *Homo sapiens* derived alleles at high frequency that have a statistically significant effect (at a FDR threshold of 0.05, as defined by the GTEX consortium (The GTEx Consortium et al., 2015)) on gene expression in any of the selected adult human tissues.

2.1 Functional categories and tissue-specificity

In quantitative terms, our data amounts to 8,271 statistically significant SNPs associated with the regulation of a total of 896 eGenes (i.e., genes affected by cis-regulation). When controlling for total eQTL variance between brain regions, a Chi-square test reveals that the proportion of derived, HF eQTLs across tissues is significantly different compared to the rest of non-derived, non-high-frequency eQTLs ($p < 2.2e - 16$). A post-hoc residual analysis indicates that regions such as the pituitary and the cerebellum are among the major contributors to reject the null hypothesis that the distribution is similar between both groups ($p < 0.05$). In other words, the pituitary and the cerebellum are the two brain regions where *Homo sapiens*-specific eQTLs accumulate relative to the control set of variants.

Derived eQTLs at high frequency are significantly different from the categories of the rest of GTEx eQTL variants in brain tissues (Chi-square test, $p < 2.2e - 16$). NMD (nonsense-mediated mRNA decay target) transcript, non coding transcript , and $5'$-UTR (untranslated region) variants are the categories driving significance ($p =< 2.2e - 16$ for the three sets, residual analysis).

To account for linkage disequilibrium and ensure statistical independence, variant clumping was applied through the eQTL mapping p-value at a $r2 = 0.1$. After clumping, the dataset was reduced to 1,270 alleles across tissues, out of which 211 are region-specific (Figure 1B). Because eQTL discovery is highly dependent on the number of tissue samples (The GTEx Consortium et al., 2015), tissues with more samples tend to yield a higher number of significant variants, regardless of tissue specificity (Figure 1C), as shown by a Spearman correlation test ($p = 0.0017$; $r = 0.74$, controlled for linkage disequilibrium). A polynomial regression line fit (blue line in Figure 1C) shows that the cerebellum, adrenal gland and BA9 fall outside the local regression's standard error confidence intervals (in gray in Figure 1C).

We sought to understand if the cerebellum, adrenal gland and BA9 stand out considering that most eQTLs are shared among regions. The distribution of clumped region-specific variants (Figure 1B) does not correlate with GTEx RNAseq sample size ($p = 0.9495$, Pearson correlation test). This lack of correlation might be explained by known effects of genetic regulation disparity between brain regions, reflected in distinct eQTL mappings for cerebellar tissue (Sieberts et al., 2020, Sng et al., 2019). Additionally, we designed a random sampling testing approach ($n = 100$) to see if any particular region tends to draw more clumped unique eQTLs regardless of total eQTL values. The test reveals no significant difference in proportions ($p = 0.3647$, Chi-square independence test). The fact that the adrenal gland and the amygdala have no unique clumped variants might be underlying this result.

2.2 Genomic regions under positive selection are enriched in eQTLs

To determine further the evolutionary significance of any of the variants in our data, we ran two randomization and permutation tests ($N = 1,000$) to test whether the derived HF eQTLs fell within regions under putative positive selection relative to other hominins as identified in two selective sweep studies ((Peyrégne et al., 2017, Racimo et al., 2014)).

We found a significant ($p = 0.001$, observed = 525 overlapping regions, expected = 53) overlap between eQTLs and regions of positive selection as defined by (Peyrégne et al., 2017), as well as in an earlier independent study (Racimo et al., 2014) ($p < 0.02$, observed = 673, expected = 177, Figure 2A and 2B). A Wilcoxon signed-rank test shows that the number of eQTLs found in positive selection regions (visualized per region in Figure 2C) is significantly different between studies ($p = 6.104e - 05$, after controlling for length differences in the windows detected by each study). A Dunn test (after Bonferroni group correction) failed to find a significant difference between the count of alleles per region in each selective sweep, despite the apparent concordance of the studies in the cerebellum (Figure 2C). We take this to mean that positive selection does not reflect a significant accumulation of eQTL variants in any given brain region, but rather seems to affect high-frequency derived eQTLs in general.

2.3 eQTL directionality depends on LD but not allele frequency or brain region

A previous study (McCoy et al., 2017) had suggested that Neanderthal alleles present in the the modern human genetic pool downregulate gene expression in brain tissue. This study also used the GTEx data, but focused on Neanderthal introgressed variants as opposed to *Homo sapiens*-derived ones.

In our derived HF eQTL dataset (Figure 3B), we did not observe any significant deviance from the expected 50% proportion between down and upregulating variants ($p = 0.3656$, Chi-square test). A significant deviance from the expected 50% proportion ($p < 2.2e-16$, Chi-square test) does obtain, however, when linkage disequilibrium is not controlled for (Figure 3A). A hierarchical cluster analysis of the distance of normalized effect size between regions in non-clumped eQTLs shows how the substantia nigra is particularly affected by the downregulating direction skewness effect (Figure 1A). This contrasts with the result found by (McCoy et al., 2017), who found this downregulation effect in cerebellum and the striatum in introgressed dataset, suggesting that variants specific to our lineage do not affect gene expression in the brain in a particular direction.

The same deviation from the expected 50% up and down-regulation proportion was present in major ancestral alleles at a 90% frequency threshold ($p =< 2.2e - 16$, Chi-square test, Figure 3C), discarding the possibility that the asymmetry is due to allele frequency cutoffs. Post-hoc residual analysis shows that downregulating eQTL skewness affects different tissues in the major and minor ancestral eQTL sets. We conclude that asymmetric directionality of eQTL regulation is not specific to a given tissue nor is accounted for by frequency.

2.4 Derived eQTLs are correlated with top hits in brain volume GWASs

As McCoy et al. (2017) had found that some of the introgressed variants from Neanderthals were also top GWAS hits, we hypothesized that derived variants might also reflect some of the changes that are characteristic of our species. We decided to focus on structural changes beyond the cortex since these are much harder to capture by endocasts. By contrast, allelic effect in gene expression can be contrasted with modern brain volume GWAS studies via two sample Mendelian randomization tests. Thus, we chose 10 brain volume GWASs that are part of the UKBiobank and IEU GWAS curated catalogs, and that focus on structures beyond the cerebreal neocortex, harder to capture by endocast analysis and thus underrepresented in the current literature, including four centered on distinct subregions of the cerebellum (left and right white matter tracts and cortices), putamen, hippocampus, amygdala, thalamus, caudate and hippocampus volume (see 4).

We first selected the top eQTL hit per gene and structure based on their eQTL p-value, under the assumption that that is the variant more strongly associated with genetic regulation, and filtered by presence in the catalog of derived alleles by Kuhlwilm and Boeckx (2019). We chose not to use high-frequency variants exclusively, as pleiotropy and linkage disequilibrium may confound the results. Under a pleiotropy model, a variant affects two different phenotypes, mixing the signal of different GWASs, while linkage disequilibrium can affect two sample Mendelian randomization by falsely detecting causality in a high frequency variant that is only in high LD with the real causal variant (one not necessarily being almost fixed or derived). The selected variants were analyzed following Wald ratio tests per gene/structure volume associations.

The results (corrected by Bonferroni) highlight genes associated with neurodevelopment and cerebellar disorders. This is consistent with the kind of phenotypes one would expect for genes associated with brain volume GWASs. However, the importance of these results lies on pinpointing which specific genes have been affected over the course of *Homo sapiens* evolution. Among the genes related to cerebellar volume in the four substructure GWASs we find genes related to ataxia (*PEX7*, *MRPS27*, *PTK2* (Bird, 1993, Jiao et al., 2020, Di Gregorio et al., 2013)), neurodevelopment (*YPEL3*, *CASP6*, *TRIM11*, *GNB5* (Blanco-Sánchez et al., 2020, Ferrer, 1999, Jabbari et al., 2018, Zhang et al., 2011)) and microcephaly (*PDCD6IP*, *USP28* (Khan et al., 2020, Phan et al., 2021)). Of note, hits for other brain structures did not correspond with eQTL regulation in the relevant tissue or have no identified functional role in brain development.

To reveal if the eQTL signal was the same as those of brain volume GWAS top hits, we ran Bayesian colocalization tests for all the eQTL that survived two sample Mendelian Randomization. However, we found that the probability that GWASs and derived eQTLs share the same signal is very low ($< 6\%$). We conclude that there is no causal relationship between eQTL expression changes and subcortical volume GWASs, and that the relationship identified here is of correlation.

3 Discussion

In this study we sought to shed light on the impact of modern-human-specific alleles found at high frequency on gene regulation across brain regions. Our intention was to complement previous work that focused on the effects of introgressed variants from Neanderthals (McCoy et al., 2017, Gunz et al., 2019).

We found that high-frequency derived eQTL indeed constitute a very useful category to understand phenotypical changes specific to our lineage. As reported in the results, these variants accumulate more than expected relative to the control set of eQTLs in the cerebellum and pituitary, are functionally differentiated and overrepresented in windows of the genome associated with signals of positive selection. Also, the enrichment of 5′UTR categories in HF derived eQTLs suggests a role for regulatory variants in *Homo sapiens* evolution (as discussed in Gokhman et al. (2020), Colbran et al. (2019), Moriano and Boeckx (2020)).

Contrary to McCoy et al. (2017) we did not find a significant skewness towards downregulation in derived eQTLs, regardless of frequency. This downregulating effect was previously detected as a characteristic of Neanderthal alleles introgressed in the modern human genetic pool (McCoy et al., 2017). The derived eQTLs examined here did show directional regulatory asymmetry but only when linkage disequilibrium was not controlled for. Additional testing indicates that the effect is not introduced by the high frequency cutoff imposed to the data, nor introduced by the bias of a particular region in either HF or non-HF alleles. We suggest that derived HF variants mapped as eQTLs might affect the modern human genetic regulation landscape by either being drivers of positive selection or being in linkage disequilibrium with causal, positively selected variants.

This idea is reinforced by our results in GWAS colocalization, showing that despite the correlation of eQTLs with subcortical brain volume GWAS top hits, there is no shared genomic signal between GWAS summary data and derived variants affecting gene expression variability. Several reasons could be put forward for this: It could be the case that the underlying causal variants are in high LD with derived eQTL and either (i) derived variants not captured by eQTL mapping, or (ii) non-derived variants that gain functionality by the effects of derived alleles in gene expression. Even if colocalization didn't detect causal variants, some of the eQTLs correlated with GWAS hits might be affecting neural phenotypes that do not leave a clear imprint in endocasts. For example, we find that derived variability in genes related to cerebellar development is correlated with this substructure's volume. The same effect was not found in other subcortical structures, as discussed in section 2.4. However, the pituitary, along with the cerebellum, has a significantly high number of derived eQTLs relative to controls, not explained by LD artifacts (figure 3B). This is relevant in light of claims that the Hypothalamic-pituitary-adrenal (HPA) axis played a role in the evolution of our social cognition (O'Rourke and Boeckx, 2018, Wrangham, 2019).

We wish to stress that our focus on brain(-related) structures in no way is intended to claim that only the brain is the most salient locus of difference between moderns and Neanderthals/Denisovans. While other organs undoubtedly display derived characteristics, we have concentrated on the brain here because our primary interest lies in cognition and behavior, which is most directly affected by brain-related changes.

4 Methods

We accessed the *Homo sapiens* variant annotation data from (Kuhlwilm and Boeckx, 2019). The original complete dataset is publicly available at `https://doi.org/10.6084/m9.figshare.8184038`. This dataset includes archaic-specific variants and all loci showing variation within modern populations, using the 1000 genomes project and ExAc data to derive frequencies and the human genome version *hg19* as reference. As described in the original article, the authors also applied quality filters in the archaic genomes (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 ancestrality was determined using multiple genome aligments (Paten et al., 2008) and the macaque reference sequence (*rheMac3*) (Yan et al., 2011).

For replication purposes, we wrote a script that reproduces the 90% frequency cutoff point used in the original study. We filtered the variants according to the guidelines in (Kuhlwilm and Boeckx, 2019) such that: 1) all variants show 90% allele frequency, 2) the major allele present in *Homo sapiens* is derived (ancestrality is either determined by the criteria in (Paten et al., 2008) or by the macaque reference allele), whereas either archaic reliable genotypes have the ancestral allele, or the Denisovan carries the ancestral allele and one of the Neanderthals the derived allele (accounting for gene flow from *Homo sapiens* to Neanderthal).

Additionally, the original study we relied on (Kuhlwilm and Boeckx, 2019) applies the 90% frequency cutoff point in a global manner: it requires that the global frequency of an allele be more than or equal to 90%, allowing for specific populations to display lower frequencies. Using the metapopulation frequency information provided in the original study, itself derived from the 1000 Genomes Project, we applied a more stringent filter and removed any alleles that where below 90% in any of the five major metapopulations included (African, American, East Asian, European, South Asian). We then harmonized and mapped the high-frequency variants to the data provided by the GTEx database (The GTEx Consortium et al., 2015). In order to do so we pruned out the alleles that did not have an assigned rsIDs.

GTEx offers data for the following tissues of interest: Adrenal Gland, Amygdala, Caudate, Brodmann Area (BA) 9, BA24, Cerebellum, Cerebellar Hemisphere, Cortex, Hippocampus, Hypothalamus, Nucleus Accumbens, Pituitary, Putamen, Spinal Cord, and Substantia Nigra. Of these samples, cerebellar hemisphere and the cerebellum, as well as cortex and BA9, are to be treated as duplicates (GTEx Consortium, 2017). Although not a brain tissue *per se*, the Adrenal Gland was included due to its role in the Hypothalamic-pituitary-adrenal (HPA) axis, an important regulator of the neuroendocrine system that affects behavior.

Post-mostem mRNA degradation affects the number of discovered eQTLs in other tissues. However, we did not control for post-mortem RNA degradation, since the Central Nervous System has been shown to be relatively resistant to this effect (Zhu et al., 2017). However, re-sampled tissues (here labeled 'cerebellar hemisphere' and 'Cortex' following the original GTEx Consortium denominations) do show differences compared to their original samples ('cerebellum' and 'BA 9'). We acknowledge that the resulting data are limited by inherent problems of the GTEx database, such the use of the same individuals for different brain tissue samples, the reduced discovery power of rare variants (GTEx Consortium, 2017), artifacts introduced during RNAseq analysis.

Clumping of the variants to control for Linkage Disequilibrium was done with Plink (version 1.9) through the *ieugwasr* R package (Elsworth et al., 2020), requiring a linkage disequilibrium score of 0.90 (i.e., co-inheritance in 90% of cases) for an SNP to be clumped. The nominal p-value of eQTL mapping was used as the criterion to define a top variant; i.e., haplotypes were clumped around the most robust eQTL candidate variant. Linkage disequilibrium values are extracted from the 1000 Genomes project ftp server (`ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`) by the *ieugwasr* R package.

Distance values for tissue hierarchical clustering were calculated by using the mean values of the normalized effect size of derived HF eQTLs.

We performed the permutation test (n=1,000) with the R package RegioneR (Gel et al., 2015) using the unclumped data, as variants might clump around an eQTL falling outside windows of putative positive selection, underepresenting the number of data points inside such genomic areas and reducing statistical power.

We ran the two sample Mendelian Randomization tests at a $p = 5e-04$ threshold for top hit identification through the *ieugwasr* (Elsworth et al., 2020), *MRinstruments*, and the colocalization tests through the *gwasglue* package. The selected GWASs for colocalization can be consulted in the relevant section of the article's code.

Figures were created with the ggplot2 R package (Wickham, 2009) and RegioneR (Gel et al., 2015). All statistical tests were controlled for power (≥ 0.8). The human selective sweep data was extracted Supplementary Table S5 from Racimo et al. (2014), and Supplementary Table S2 from Peyrégne et al. (2017). GWAS summary data and harmonized top eQTL instruments for two sample Mendelian Randomization were extracted from the IEU GWAS database API Elsworth et al. (2020).

Data availability statement

All data, statistical tests and figures reported in this manuscript, including external data downloads, can be reproduced by the code contained in the following repository: https://github.com/AGMAndirko/GTEX-code

Acknowledgments

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on November 2020.

Author Contributions

Conceptualization: CB & AA; Data Curation: AA; Formal Analysis: AA; Funding Acquisition: CB; Investigation: CB & AA; Methodology: CB & AA; Software: AA; Supervision: CB; Visualization: CB & AA; Writing — Original Draft Preparation: CB & AA; Writing — Review & Editing: CB & AA.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund (BES-2017-080366); the Spanish Ministry of Science and Innovation (grant PID2019-107042GB-I00); a Leonardo fellowship from the BBVA Foundation; the Fundació Bosch i Gimpera; a MEXT/JSPS Grant-in-Aid for Scientific Research on Innovative Areas 4903 (Evolinguistics: JP17H06379); the Generalitat de Catalunya (2017-SGR-341).

Competing interest

Authors declare no competing financial or non-financial interest.

References

T. D. Bird. Hereditary Ataxia Overview. In M. P. Adam, H. H. Ardinger, R. A. Pagon, S. E. Wallace, L. J. Bean, G. Mirzaa, and A. Amemiya, editors, *GeneReviews®*. University of Washington, Seattle, Seattle (WA), 1993. URL http://www.ncbi.nlm.nih.gov/books/NBK1138/.

B. Blanco-Sánchez, A. Clément, S. J. Stednitz, J. Kyle, J. L. Peirce, M. McFadden, J. Wegner, J. B. Phillips, E. Macnamara, Y. Huang, D. R. Adams, C. Toro, W. A. Gahl, M. C. V. Malicdan, C. J. Tifft, E. M. Zink, K. J. Bloodsworth, K. G. Stratton, Undiagnosed Diseases Network, D. M. Koeller, T. O. Metz, P. Washbourne, and M. Westerfield. yippee like 3(ypel3) is a novel gene required for myelinating and perineurial glia development. *PLoS genetics*, 16(6):e1008841, June 2020. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008841.

L. L. Colbran, E. R. Gamazon, D. Zhou, P. Evans, N. J. Cox, and J. A. Capra. Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat Ecol Evol*, 3(11):1598–1606, Nov. 2019. ISSN 2397-334X. doi: 10.1038/s41559-019-0996-x.

E. Di Gregorio, F. T. Bianchi, A. Schiavi, A. M. A. Chiotto, M. Rolando, L. Verdun di Cantogno, E. Grosso, S. Cavalieri, A. Calcia, D. Lacerenza, O. Zuffardi, S. F. Retta, G. Stevanin, C. Marelli, A. Durr, S. Forlani, J. Chelly, F. Montarolo, F. Tempia, H. E. Beggs, R. Reed, S. Squadrone, M. C. Abete, A. Brussino, N. Ventura, F. Di Cunto, and A. Brusco. A de novo X;8 translocation creates a PTK2-THOC2 gene fusion with THOC2 expression knockdown in a patient with psychomotor retardation and congenital cerebellar hypoplasia. *Journal of Medical Genetics*, 50(8): 543–551, Aug. 2013. ISSN 1468-6244. doi: 10.1136/jmedgenet-2013-101542.

B. Elsworth, M. Lyon, T. Alexander, Y. Liu, P. Matthews, J. Hallett, P. Bates, T. Palmer, V. Haberland, G. D. Smith, J. Zheng, P. Haycock, T. R. Gaunt, and G. Hemani. The MRC IEU OpenGWAS data infrastructure. preprint, Genetics, Aug. 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.08.10.244293.

I. Ferrer. Role of caspases in ionizing radiation-induced apoptosis in the developing cerebellum. *Journal of Neurobiology*, 41(4):549–558, Dec. 1999. ISSN 0022-3034. doi: 10.1002/(sici)1097-4695(199912)41:4⟨549::aid-neu10⟩3.0.co;2-g.

B. Gel, A. Díez-Villanueva, E. Serra, M. Buschbeck, M. A. Peinado, and R. Malinverni. regioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, page btv562, Sept. 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btv562.

D. Gokhman, M. Nissim-Rafinia, L. Agranat-Tamir, G. Housman, R. García-Pérez, E. Lizano, O. Cheronet, S. Mallick, M. A. Nieves-Colón, H. Li, S. Alpaslan-Roodenberg, M. Novak, H. Gu, J. M. Osinski, M. Ferrando-Bernal, P. Gelabert, I. Lipende, D. Mjungu, I. Kondova, R. Bontrop, O. Kullmer, G. Weber, T. Shahar, M. Dvir-Ginzberg, M. Faerman, E. E. Quillen, A. Meissner, Y. Lahav, L. Kandel, M. Liebergall, M. E. Prada, J. M. Vidal, R. M. Gronostajski, A. C. Stone, B. Yakir, C. Lalueza-Fox, R. Pinhasi, D. Reich, T. Marques-Bonet, E. Meshorer, and L. Carmel. Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat Commun*, 11(1):1189, Dec. 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15020-6.

GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, Oct. 2017. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature24277.

P. Gunz, S. Neubauer, B. Maureille, and J.-J. Hublin. Brain development after birth differs between Neanderthals and modern humans. *Current Biology*, 20(21):R921–R922, Nov. 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.10.018.

P. Gunz, A. K. Tilot, K. Wittfeld, A. Teumer, C. Y. Shapland, T. G. van Erp, M. Dannemann, B. Vernot, S. Neubauer, T. Guadalupe, G. Fernández, H. G. Brunner, W. Enard, J. Fallon, N. Hosten, U. Völker, A. Profico, F. Di Vincenzo, G. Manzi, J. Kelso, B. St. Pourcain, J.-J. Hublin, B. Franke, S. Pääbo, F. Macciardi, H. J.

Grabe, and S. E. Fisher. Neandertal Introgression Sheds Light on Modern Human Endocranial Globularity. *Current Biology*, 29(1):120–127.e5, Jan. 2019. ISSN 09609822. doi: 10.1016/j.cub.2018.10.065.

J.-J. Hublin, S. Neubauer, and P. Gunz. Brain ontogeny and life history in Pleistocene hominins. *Phil. Trans. R. Soc. B*, 370(1663):20140062, Mar. 2015. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2014.0062.

E. Jabbari, J. Woodside, M. M. X. Tan, M. Shoai, A. Pittman, R. Ferrari, K. Y. Mok, D. Zhang, R. H. Reynolds, R. de Silva, M.-J. Grimm, G. Respondek, U. Müller, S. Al-Sarraj, S. M. Gentleman, A. J. Lees, T. T. Warner, J. Hardy, T. Revesz, G. U. Höglinger, J. L. Holton, M. Ryten, and H. R. Morris. Variation at the TRIM11 locus modifies progressive supranuclear palsy phenotype. *Annals of Neurology*, 84 (4):485–496, Oct. 2018. ISSN 1531-8249. doi: 10.1002/ana.25308.

B. Jiao, Z. Zhou, Z. Hu, J. Du, X. Liao, Y. Luo, J. Wang, X. Yan, H. Jiang, B. Tang, and L. Shen. Homozygosity mapping and next generation sequencing for the genetic diagnosis of hereditary ataxia and spastic paraplegia in consanguineous families. *Parkinsonism & Related Disorders*, 80:65–72, Nov. 2020. ISSN 1873-5126. doi: 10.1016/j.parkreldis.2020.09.013.

A. Khan, M. Alaamery, S. Massadeh, A. Obaid, A. A. Kashgari, C. A. Walsh, and W. Eyaid. PDCD6IP, encoding a regulator of the ESCRT complex, is mutated in microcephaly. *Clinical Genetics*, 98(1):80–85, July 2020. ISSN 1399-0004. doi: 10.1111/cge.13756.

M. King and A. Wilson. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, Apr. 1975. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1090005.

T. Kochiyama, N. Ogihara, H. C. Tanabe, O. Kondo, H. Amano, K. Hasegawa, H. Suzuki, M. S. P. de León, C. P. E. Zollikofer, M. Bastir, C. Stringer, N. Sadato, and T. Akazawa. Reconstructing the Neanderthal brain using computational anatomy. *Scientific Reports*, 8(1):6296, Apr. 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-24331-0.

M. Kuhlwilm and C. Boeckx. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci Rep*, 9(1):8463, Dec. 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-44877-x.

F. Mafessoni, S. Grote, C. d. Filippo, V. Slon, K. A. Kolobova, B. Viola, S. V. Markin, M. Chintalapati, S. Peyrégne, L. Skov, P. Skoglund, A. I. Krivoshapkin, A. P. Derevianko, M. Meyer, J. Kelso, B. Peter, K. Prüfer, and S. Pääbo. A high-coverage Neandertal genome from Chagyrskaya Cave. *Proceedings of the National Academy of Sciences*, 117(26):15132–15136, June 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2004944117. URL `https://www.pnas.org/content/117/26/15132`. Publisher: National Academy of Sciences Section: Biological Sciences.

R. C. McCoy, J. Wakefield, and J. M. Akey. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell*, 168(5):916–927.e12, Feb. 2017. ISSN 00928674. doi: 10.1016/j.cell.2017.01.038.

M. Meyer, M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andrés, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*, 338(6104):222–226, Oct. 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1224344.

J. Moriano and C. Boeckx. Modern human changes in regulatory regions implicated in cortical development. *BMC Genomics*, 21(1), Apr. 2020. doi: 10.1186/s12864-020-6706-x. URL `https://doi.org/10.1186/s12864-020-6706-x`.

S. Neubauer, J.-J. Hublin, and P. Gunz. The evolution of modern human brain shape. *Sci. Adv.*, 4(1):eaao5961, Jan. 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aao5961.

T. O'Rourke and C. Boeckx. Converging roles of glutamate receptors in domestication and prosociality. Preprint, Evolutionary Biology, Oct. 2018.

S. Pääbo. The Human Condition—A Molecular Approach. *Cell*, 157(1):216–226, Mar. 2014. ISSN 00928674. doi: 10.1016/j.cell.2013.12.036.

B. Paten, J. Herrero, S. Fitzgerald, K. Beal, P. Flicek, I. Holmes, and E. Birney. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research*, 18(11):1829–1843, Nov. 2008. ISSN 1088-9051. doi: 10.1101/gr.076521.108.

A. S. Pereira-Pedro, E. Bruner, P. Gunz, and S. Neubauer. A morphometric comparison of the parietal lobe in modern humans and Neanderthals. *Journal of Human Evolution*, 142:102770, May 2020. ISSN 00472484. doi: 10.1016/j.jhevol.2020.102770.

S. Peyrégne, M. J. Boyle, M. Dannemann, and K. Prüfer. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res.*, 27(9):1563–1572, Sept. 2017. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.219493.116.

T. P. Phan, A. L. Maryniak, C. A. Boatwright, J. Lee, A. Atkins, A. Tijhuis, D. C. Spierings, H. Bazzi, F. Foijer, P. W. Jordan, T. H. Stracker, and A. J. Holland. Centrosome defects cause microcephaly by activating the 53BP1-USP28-TP53 mitotic surveillance pathway. *The EMBO journal*, 40(1):e106118, Jan. 2021. ISSN 1460-2075. doi: 10.15252/embj.2020106118.

K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. F. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, and S. Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, Jan. 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12886.

K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. M. Andrés, J. Kelso, M. Meyer, and S. Pääbo. A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science*, 358(6363):655–658, Nov. 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aao1887.

F. Racimo, M. Kuhlwilm, and M. Slatkin. A Test for Ancient Selective Sweeps and an Application to Candidate Sites in Modern Humans. *Molecular Biology and Evolution*, 31(12):3344–3358, Dec. 2014. ISSN 1537-1719, 0737-4038. doi: 10.1093/molbev/msu255.

S. K. Sieberts, , T. M. Perumal, M. M. Carrasquillo, M. Allen, J. S. Reddy, G. E. Hoffman, K. K. Dang, J. Calley, P. J. Ebert, J. Eddy, X. Wang, A. K. Greenwood, S. Mostafavi, L. Omberg, M. A. Peters, B. A. Logsdon, P. L. D. Jager, N. Ertekin-Taner, and L. M. M. and. Large eQTL meta-analysis reveals differing patterns between cerebral cortical and cerebellar brain regions. *Scientific Data*, 7 (1), Oct. 2020. doi: 10.1038/s41597-020-00642-8. URL `https://doi.org/10.1038/s41597-020-00642-8`.

L. M. F. Sng, P. C. Thomson, and D. Trabzuni. Genome-wide human brain eQTLs: In-depth analysis and insights using the UKBEC dataset. *Sci Rep*, 9(1):19201, Dec. 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-55590-0.

The GTEx Consortium, K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, J. B. Maller, T. Tukiainen, M. Lek, L. D. Ward, P. Kheradpour, B. Iriarte, Y. Meng, C. D. Palmer, T. Esko, W. Winckler, J. N. Hirschhorn, M. Kellis, D. G. MacArthur, G. Getz, A. A. Shabalin, G. Li, Y.-H. Zhou, A. B. Nobel, I. Rusyn, F. A. Wright, T. Lappalainen, P. G. Ferreira, H. Ongen, M. A. Rivas, A. Battle, S. Mostafavi, J. Monlong, M. Sammeth, M. Mele, F. Reverter, J. M. Goldmann, D. Koller, R. Guigo, M. I. McCarthy, E. T. Dermitzakis, E. R. Gamazon, H. K. Im, A. Konkashbaev, D. L. Nicolae, N. J. Cox, T. Flutre, X. Wen, M. Stephens, J. K. Pritchard, Z. Tu, B. Zhang, T. Huang, Q. Long, L. Lin, J. Yang, J. Zhu, J. Liu, A. Brown, B. Mestichelli, D. Tidwell, E. Lo, M. Salvatore, S. Shad, J. A. Thomas, J. T. Lonsdale, M. T. Moser, B. M. Gillard, E. Karasik, K. Ramsey, C. Choi, B. A. Foster, J. Syron, J. Fleming, H. Magazine, R. Hasz, G. D. Walters, J. P. Bridge, M. Miklos, S. Sullivan, L. K. Barker, H. M. Traino, M. Mosavel, L. A. Siminoff, D. R. Valley, D. C. Rohrer, S. D. Jewell, P. A. Branton, L. H. Sobin, M. Barcus, L. Qi, J. McLean, P. Hariharan, K. S. Um, S. Wu, D. Tabor, C. Shive, A. M. Smith, S. A. Buia, A. H. Undale, K. L. Robinson, N. Roche, K. M. Valentino, A. Britton, R. Burges, D. Bradbury, K. W. Hambright, J. Seleski, G. E. Korzeniewski, K. Erickson, Y. Marcus, J. Tejada, M. Taherian, C. Lu, M. Basile, D. C. Mash, S. Volpi, J. P. Struewing, G. F. Temple, J. Boyer, D. Colantuoni, R. Little, S. Koester, L. J. Carithers, H. M. Moore, P. Guan, C. Compton, S. J. Sawyer, J. P. Demchok, J. B. Vaught, C. A. Rabiner, N. C. Lockhart, K. G. Ardlie, G. Getz, F. A. Wright, M. Kellis, S. Volpi, and E. T. Dermitzakis. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1262110.

C. A. Trujillo, E. S. Rice, N. K. Schaefer, I. A. Chaim, E. C. Wheeler, A. A. Madrigal, J. Buchanan, S. Preissl, A. Wang, P. D. Negraes, R. A. Szeto, R. H. Herai, A. Huseynov, M. S. A. Ferraz, F. S. Borges, A. H. Kihara, A. Byrne, M. Marin, C. Vollmers, A. N. Brooks, J. D. Lautz, K. Semendeferi, B. Shapiro, G. W. Yeo, S. E. P. Smith, R. E. Green, and A. R. Muotri. Reintroduction of the archaic variant of NOVA1 in cortical organoids alters neurodevelopment. *Science*, 371 (6530), Feb. 2021. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax2537. URL `https://science.sciencemag.org/content/371/6530/eaax2537`.

H. Wickham. *Ggplot2: Elegant Graphics for Data Analysis.* Use R! Springer, New York, 2009. ISBN 978-0-387-98140-6. OCLC: ocn382399721.

R. W. Wrangham. *The Goodness Paradox: The Strange Relationship between Virtue and Violence in Human Evolution.* Pantheon Books, New York, first edition edition, 2019. ISBN 978-1-101-87090-7.

G. Yan, G. Zhang, X. Fang, Y. Zhang, C. Li, F. Ling, D. N. Cooper, Q. Li, Y. Li, A. J. van Gool, H. Du, J. Chen, R. Chen, P. Zhang, Z. Huang, J. R. Thompson, Y. Meng, Y. Bai, J. Wang, M. Zhuo, T. Wang, Y. Huang, L. Wei, J. Li, Z. Wang, H. Hu, P. Yang, L. Le, P. D. Stenson, B. Li, X. Liu, E. V. Ball, N. An, Q. Huang, Y. Zhang, W. Fan, X. Zhang, Y. Li, W. Wang, M. G. Katze, B. Su, R. Nielsen, H. Yang, J. Wang, X. Wang, and J. Wang. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nature Biotechnology*, 29(11):1019–1023, Nov. 2011. ISSN 1546-1696. doi: 10.1038/nbt.1992. URL `https://www.nature.com/articles/nbt.1992`.

J.-H. Zhang, M. Pandey, E. M. Seigneur, L. M. Panicker, L. Koo, O. M. Schwartz, W. Chen, C.-K. Chen, and W. F. Simonds. Knockout of G protein beta5 impairs brain development and causes multiple neurologic abnormalities in mice. *Journal of Neurochemistry*, 119(3):544–554, Nov. 2011. ISSN 1471-4159. doi: 10.1111/j.1471-4159.2011.07457.x.

Y. Zhu, L. Wang, Y. Yin, and E. Yang. Systematic analysis of gene expression patterns associated with postmortem interval in human tissues. *Scientific Reports*, 7(1): 5435, July 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-05882-0. URL `https://www.nature.com/articles/s41598-017-05882-0`.

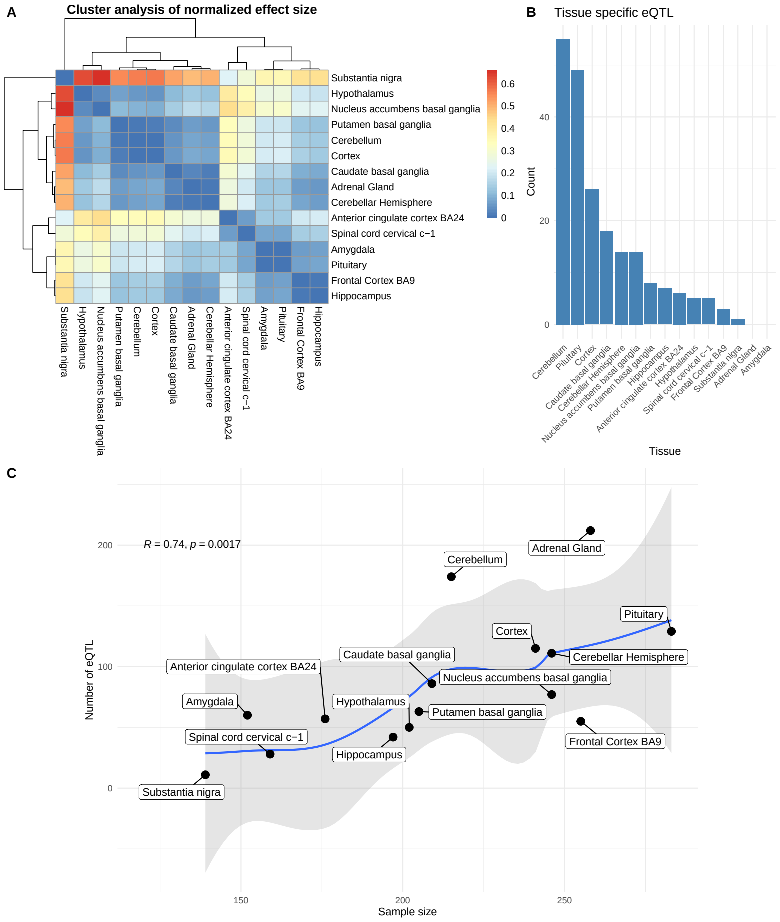


Figure 1: A: Hierarchical clustering analysis of eQTL normal effect size, not controlled for linkage disequilibrium (LD). Color denotes hierarchical distance. B: Number of tissue-specific eQTLs after clumping. Adrenal gland and Amygdala do not contain tissue-specific eQTL in our dataset. C: Brain region sample size and eQTL count correlate in our dataset. The blue line marks a polynomial regression line fit, with regression's standard error confidence intervals (95%) in gray.

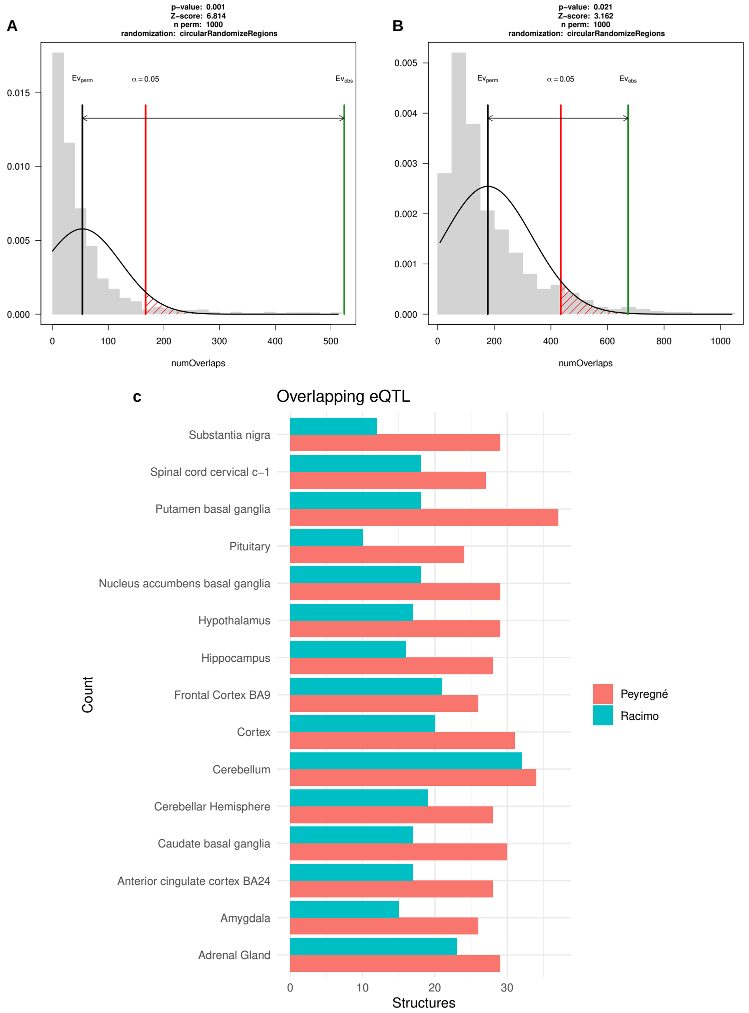


Figure 2: Derived, HF eQTLs are present more than expected by chance in selective sweeps from (Peyrégne et al., 2017) (A) and (Racimo et al., 2014) (B). C shows the count of eQTL overlapping with regions under putative positive selection per region.

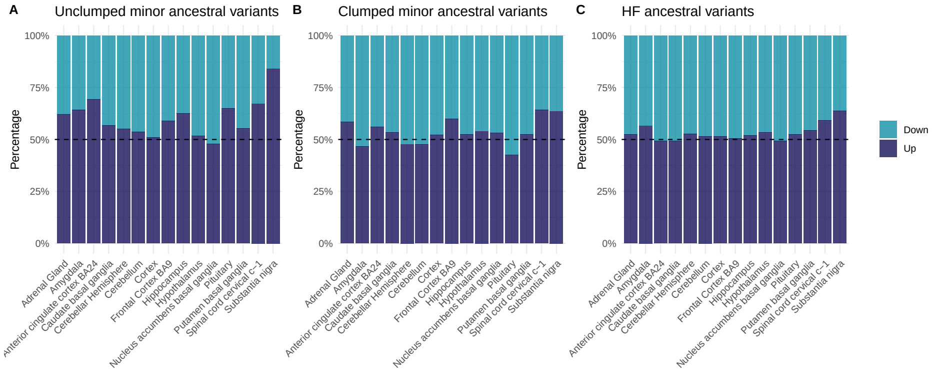


Figure 3: Distribution of up and down-regulating ancestral variants across different subsets of the data, in all eGenes. We include here data before (A) and after (B) controlling for linkage disequilibrium in minor alleles (≥10% frequency). A control using major ancestral alleles (at ≥90% frequency) is included (C).

3 | Williams-Beuren syndrome, BAZ1B and the human face

EVOLUTIONARY BIOLOGY

Dosage analysis of the 7q11.23 Williams region identifies *BAZ1B* as a major human gene patterning the modern human face and underlying self-domestication

Matteo Zanella[1,2*†], Alessandro Vitriolo[1,2*], Alejandro Andirko[3,4], Pedro Tiago Martins[3,4], Stefanie Sturm[3,4], Thomas O'Rourke[3,4], Magdalena Laugsch[5,6,7], Natascia Malerba[8], Adrianos Skaros[1,2], Sebastiano Trattaro[1,2], Pierre-Luc Germain[1,2,9], Marija Mihailovic[1,2], Giuseppe Merla[8], Alvaro Rada-Iglesias[5,10,11], Cedric Boeckx[3,4,12], Giuseppe Testa[1,2,13‡]

We undertook a functional dissection of chromatin remodeler BAZ1B in neural crest (NC) stem cells (NCSCs) from a uniquely informative cohort of typical and atypical patients harboring 7q11.23 copy number variants. Our results reveal a key contribution of BAZ1B to NCSC in vitro induction and migration, coupled with a crucial involvement in NC-specific transcriptional circuits and distal regulation. By intersecting our experimental data with new paleogenetic analyses comparing modern and archaic humans, we found a modern-specific enrichment for regulatory changes both in BAZ1B and its experimentally defined downstream targets, thereby providing the first empirical validation of the human self-domestication hypothesis and positioning BAZ1B as a master regulator of the modern human face. In so doing, we provide experimental evidence that the craniofacial and cognitive/behavioral phenotypes caused by alterations of the Williams-Beuren syndrome critical region can serve as a powerful entry point into the evolution of the modern human face and prosociality.

INTRODUCTION

Anatomically modern humans (AMHs) exhibit a suite of craniofacial and prosocial characteristics that are reminiscent of traits distinguishing domesticated species from their wild counterparts (*1*–*3*). This has led to the formulation of a self-domestication hypothesis according to which modern humans (*3*) went through a domestication process in the course of their evolution. Recent evidence, along with the well-warranted distinction between domestication and selective breeding (*4*), is also extending this notion to other species that might have undergone a self-domestication phase, such as cats, dogs, and bonobos (*3*). Thus, as self-domestication represents a special case of domestication, the most parsimonious hypothesis must posit the same core mechanisms to underlie both. For this reason, the self-domestication hypothesis also entails the prediction that key aspects of modern humans' anatomy and cognition can be illuminated by studies of the so-called "domestication syndrome," the core set of domestication-related traits that was recently proposed to result from mild neural crest (NC) deficits (*5*). However, both the neurocristopathic basis of domestication and its extension to the evolution of AMHs remain to be tested experimentally.

Williams-Beuren syndrome [WBS; OMIM (Online Mendelian Inheritance in Man) 194050] and Williams-Beuren region duplication syndrome (7dupASD; OMIM 609757), caused respectively by the hemideletion or hemiduplication of 28 genes at the 7q11.23 region [WBS critical region (WBSCR)], represent a paradigmatic pair of neurodevelopmental conditions whose NC-related craniofacial dysmorphisms and cognitive/behavioral traits (*6*, *7*) bear directly on domestication-related traits relevant for AMHs (facial reduction and retraction, pronounced friendliness, and reduced reactive aggression) (fig. S1A). Structural variants in WBS genes, for example in the case of GTF2I and its paralogs, have been shown to underlie stereotypical hypersociability in domestic dogs and foxes (*8*, *9*).

Among the WBSCR genes, we focus here on the chromatin regulator *BAZ1B* (also known as Williams syndrome transcription factor, *WSTF*), on the basis of the following lines of evidence that implicate it in domestication-relevant craniofacial features: (i) its established role in NC maintenance and migration in *Xenopus laevis* and the craniofacial defects observed in knockout mice (*10*, *11*); (ii) the observation that its expression is affected by domestication-related events in canids (*12*); (iii) the first formulation of the neurocristopathic hypothesis of domestication, which included *BAZ1B* among the genes influencing NC development (*5*); (iv) the most comprehensive studies focusing on regions of the modern human genome associated with selective sweep signals compared to Neanderthals/Denisovans (hereafter "archaics") (*13*, *14*), one of which specifically included *BAZ1B* within the detected portions of the WBSCR; and (v) the thus far most detailed study systematically exploring high-frequency (HF) (>90%) changes in modern humans for which archaic humans carry the

[1]Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy. [2]Laboratory of Stem Cell Epigenetics, IEO, European Institute of Oncology, IRCCS, Milan, Italy. [3]University of Barcelona, Barcelona, Spain. [4]University of Barcelona Institute of Complex Systems (UBICS), Barcelona, Spain. [5]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany. [6]Institute of Human Genetics, University Hospital Cologne, Cologne, Germany. [7]Institute of Human Genetics, University Hospital Heidelberg, Heidelberg, Germany. [8]Division of Medical Genetics, Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Foggia, Italy. [9]D-HEST Institute for Neuroscience, ETH Zürich, Switzerland. [10]Cluster of Excellence Cellular Stress Responses in Aging-Associated Diseases (CECAD), University of Cologne, Germany. [11]Institute of Biomedicine and Biotechnology of Cantabria, University of Cantabria, Cantabria, Spain. [12]Catalan Institute for Advanced Studies and Research (ICREA), Barcelona, Spain. [13]Human Technopole, Center for Neurogenomics, Via Cristina Belgioioso 171, Milan, Italy.
*These authors contributed equally to this work.
†Present address: Evotec SE, Hamburg, Germany.
‡Corresponding author. Email: giuseppe.testa@unimi.it, giuseppe.testa@ieo.it, giuseppe.testa@htechnopole.it

ancestral state, which found *BAZ1B* enriched for mutations in modern humans (most of which fall in the regulatory regions of the gene) (*15*).

Our previous work had established the largest cohort of 7q11.23 patient-derived induced pluripotent stem cell (iPSC) lines and revealed major disease-relevant transcriptional dysregulation that was already apparent at the pluripotent state and became further exacerbated upon differentiation (*16*). Here, we first harness this resource to dissect the impact of *BAZ1B* dosage on the NC of patients with WBS and 7dupASD, both in terms of function (i.e., NC migration and induction) and of transcriptional and chromatin dysregulation, thereby defining the BAZ1B dosage–dependent circuits controlling the NC. Next, we apply these experimentally determined BAZ1B-dependent circuits underlying craniofacial morphogenesis to interrogate the evidence from paleogenomic analyses, which were thus far only of a correlative nature. We find major convergence between the BAZ1B control and the genes harboring regulatory changes in the modern human lineage. Together, the definition of the role of BAZ1B dosage in craniofacial neurocristopathy and its application to domestication-relevant paleogenomics demonstrate a major contribution of BAZ1B to the modern human face and offer experimental validation for the prediction at the heart of NC-based accounts of (self-) domestication: that the modern human face acquired its shape as an instance of mild neurocristopathy.

RESULTS

Establishment and validation of an extensive cohort of patient-specific BAZ1B-interfered NC stem cell lines

To dissect the role of BAZ1B in the craniofacial dysmorphisms that characterize WBS and 7dupASD, we started from our previous characterization of WBS patient– and 7dupASD patient–specific iPSC lines and differentiated derivatives (*16*) and selected a cohort of 11 NC stem cell (NCSC) lines (four from patients with WBS, three from patients with 7dupASD, and four from control individuals), which also represent the largest cohort of patient-specific NCSCs described so far. Given the centrality of the cranial NC for the development of the face, we first validated the cranial identity of our NCSC cohort by transcriptomic profiling through a manually curated gene expression signature (fig. S2A), confirming their suitability for the study of craniofacial dysregulations. We then knocked down BAZ1B via RNA interference in all lines across the three genetic conditions, including also NCSCs derived from a particularly informative patient with atypical WBS (hereafter atWBS) bearing a partial deletion of the region that spares *BAZ1B* and six additional genes (Fig. 1A) (*17*). To establish a high-resolution gradient of *BAZ1B* dosages, we selected two distinct short hairpin RNA (shRNA) against BAZ1B (i.e., sh1 and sh2) along with a scrambled shRNA sequence (hereafter scr) as negative control, for a total of 32 NCSC lines. Knockdown (KD) efficiency was evaluated at the RNA level by quantitative polymerase chain reaction (qPCR) (Fig. 1B and fig. S1C), confirming the attainment of the desired gradient with an overall reduction of about 40% for sh1 and 70% for sh2, as well as reduction at the protein level, as detected by Western blot (fig. S1E).

BAZ1B dosage imbalance impairs NCSC migration and induction

NCSCs need to migrate to reach specific target regions in the developing embryo and give rise to distinct cell types and tissues, including craniofacial structures that are major areas of change in human evolution. Since BAZ1B KD was shown to affect the migration of the NC in *X. laevis* and to promote cancer cell invasion in different lung cancer cell lines (*10*, *18*), we hypothesized that the *BAZ1B* dosage imbalances entailed in the 7q11.23 syndromes could result in a defective regulation of NCSC migration and might underlie the NC-related alterations typical of patients with WBS and 7dupASD. To test this, we compared the migration properties of patient-specific BAZ1B KD NCSC lines (sh2) to their respective control NCSC line (scr) by the well-established wound-healing assay. The 7dupASD NCSC KD lines took longer to fill the wound when compared to the respective control lines (scr), as indicated by images taken at 8 and 16 hours after a gap was created on the plate surface (Fig. 1C and fig. S1F). We instead observed an opposite behavior for the WBS BAZ1B KD lines, which were faster than the respective scr lines in closing the gap (Fig. 1C and fig. S1F). In contrast to the previous observations from *X. laevis* (*10*), we also observed a minor delay in NC induction as a consequence of BAZ1B KD (Fig. 1D and fig. S1D), by means of a differentiation protocol based on NC delamination from adherent embryoid bodies (EBs), which recapitulates the initial steps of NC generation (*19*). In particular, starting from 2 to 3 days after attachment of EBs, we observed a lower number of outgrowing cells in the KD line (Fig. 1D, days 7 and 10), coupled with an evidently higher cell mortality. Cells were eventually able to acquire the typical NC morphology, although lower differentiation efficiency was evident, as shown by images taken at day 12. In addition, the delay in NC formation was associated with a down-regulation of well-established critical regulators of NC migration and maintenance, including *NR2F1*, *NR2F2*, *TFAP2A*, and *SOX9* (Fig. 1E). These results show that BAZ1B regulates the developing NC starting from its earliest migratory stages and that the symmetrically opposite 7q11.23 dosages alterations prime NCSCs to symmetrically opposite deficits upon BAZ1B interference. In turn, the central role of the NC in the development of facial morphology allows relating such findings to the symmetrically opposite craniofacial dysmorphisms of the two 7q11.23 syndromes.

BAZ1B interference disrupts key NC-specific transcriptional circuits

Having defined the functional impact of BAZ1B dosage on NC function, we predicted that a main molecular readout of its dosage imbalances would be at the level of transcriptional regulation, given its critical role as transcriptional regulator in different cell and animal models (*20*–*22*). To test this hypothesis and gain mechanistic insights into the specific BAZ1B dosage–dependent downstream circuits, we subjected 32 interfered NCSC lines to high-coverage RNA sequencing (RNA-seq) analysis. As shown in fig. S2A, a manually curated signature from an extensive literature review (*23*–*28*) validated the cranial identity of our NCSC lines, while clustering by Pearson correlation excluded the presence of any genotype- or hairpin-specific expression change. Confirming our previous observations in the two largest cohorts of iPSC lines (*29*), a principal component analysis (PCA) corroborated the significant impact of individual genetic backgrounds on transcriptional variability, with most "KD lines" clustering with their respective control "scr line." This was consistent with the narrow range of experimentally interfered BAZ1B dosages and pointed to a selective BAZ1B dosage–dependent transcriptional vulnerability (fig. S2B).

To dissect it, we thus resorted to a combination of classical pairwise comparative analysis, contrasting shBAZ1B-interfered NCSC lines

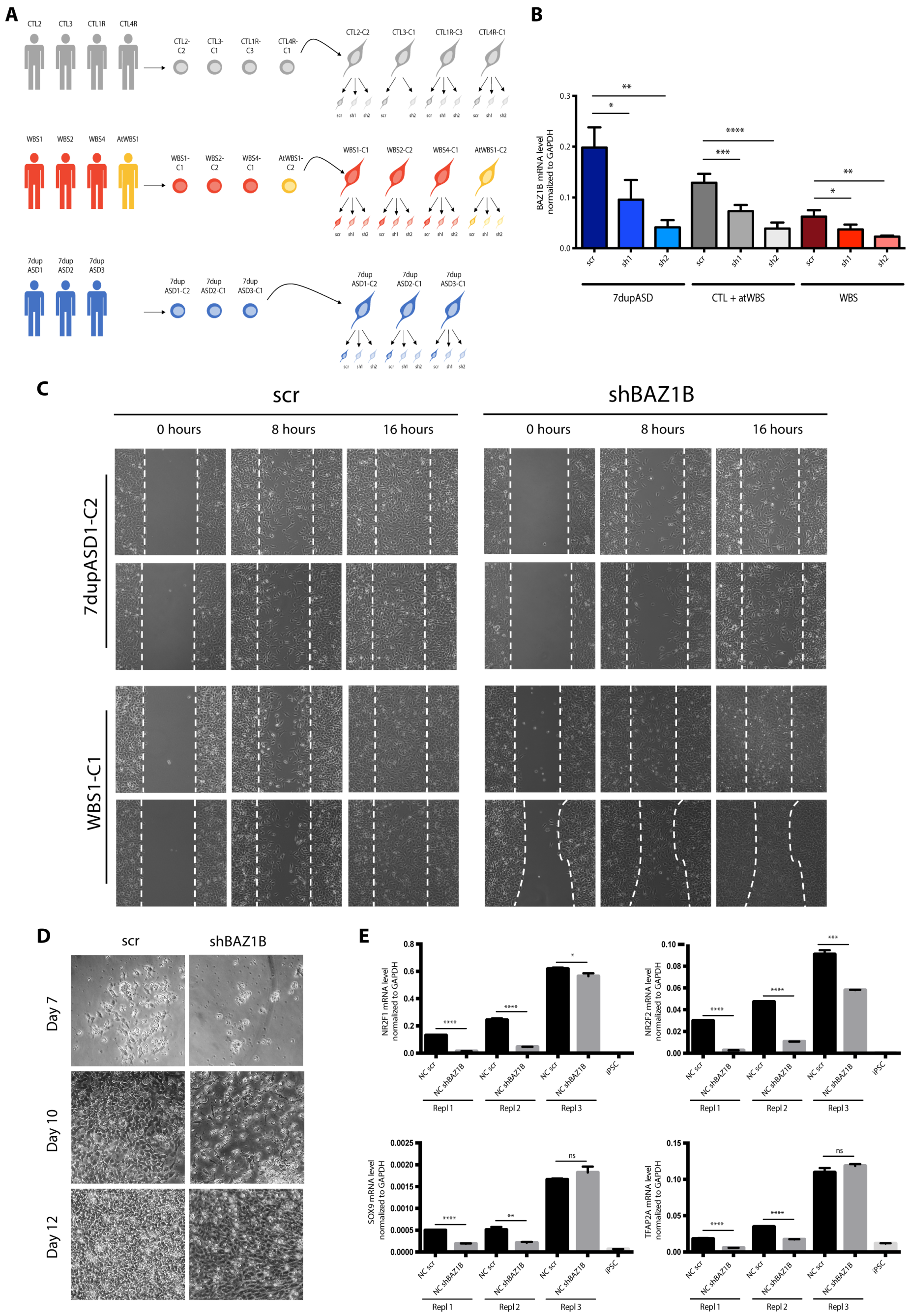


Fig. 1. *BAZ1B KD* impairs migration and induction of patient-specific iPSC-derived NCSCs. (**A**) Schematic representation of the KD strategy on our iPSC-derived NCSC cohort. (**B**) BAZ1B mRNA levels in all the interfered lines (scr, sh1, and sh2) as measured by qPCR. Data represent aggregates of samples with the same number of *BAZ1B* copies (7dup, CTL + atWBS, and WBS). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is used as a normalizer. (**C**) Eight- and 16-hour time points from the wound-healing assay analyses performed on a 7dupASD and a WBS NCSC line upon BAZ1B KD. Cells from the same line infected with the scr sh were used as references for the migration ($n = 2$). (**D**) Days 7, 10, and 12 of NC differentiation from embryoid bodies (EBs) of an scr-interfered iPSC line and its respective BAZ1B KD ($n = 3$). (**E**) mRNA levels of NC markers at day 12 of differentiation in three individual experimental replicates [bright-field images are reported in (D)]. An iPSC line is included as a negative control. Student's *t* test was used (ns, not significant; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, and $****P < 0.0001$).

(sh1 + sh2) with their respective controls (scr), with a complementary regression analysis using *BAZ1B* expression levels as independent variables, subtracting the contribution of individual genetic backgrounds. This design increases robustness and sensitivity in the identification of genes that, across multiple genetic backgrounds and target gene dosages, might have a different baseline (scr) across individuals while still being robustly dysregulated upon BAZ1B interference.

The two analyses identified a total of 448 genes with false discovery rate (FDR) < 0.1 (1192 with $P < 0.01$ and FDR < 0.25) whose transcriptional levels followed BAZ1B dosage, in either a direct (202; 539 with $P < 0.01$ and FDR < 0.25) or an inverse (246; 653 with $P < 0.01$ and FDR < 0.25) fashion. In addition, genes identified in the regression analysis included around 90% of the differentially expressed genes (DEGs) (27 of 29, FDR < 0.1) found in the comparative analysis (Fig. 2A). Consistent with the differential efficiency of the two short hairpins, we found a globally stronger transcriptional impact for the group of samples targeted by sh2 (fig. S2C) and a milder but nevertheless clearly distinguishable effect of sh1, resulting in particularly informative gradient of dosages over the scr control lines.

Particularly noteworthy among the genes that we found correlated with *BAZ1B* levels were (i) crucial regulators of cranial NC, further highlighting a convergent BAZ1B dosage–dependent dysregulation of the foundational *CUL3*-centered regulatory axis orchestrating NC-mediated craniofacial morphogenesis (*30*), and (ii) genes associated with variation of human facial shape or causative of dysmorphic facial features and mild intellectual disability when mutated (Fig. 2B and table S1).

Gene Ontology (GO) analysis performed on genes directly following BAZ1B levels suggested specific enrichments in biological processes such as histone phosphorylation, chromosome localization, RNA processing, and splicing. Genes inversely following BAZ1B levels were instead enriched in categories particularly relevant for NC and NC-derivative functions, such as cell migration and cardiovascular and skeletal development (Fig. 2C). By querying the OMIM database, we found that several DEGs were associated with genetic disorders whose phenotypes include "mental retardation," "intellectual disability," and/or "facial dysmorphisms" (Fig. 2D), underscoring the pertinence of BAZ1B-dependent dysregulation across both the neurocristopathic and cognitive axes.

Last, a master regulator analysis identified candidate regulators of BAZ1B DEGs, including factors involved in enhancer marking [CEBPB, p300, RBBP5, HDAC2 (histone deacetylase 2), KDM1A, and TCF12], promoter activation [TBP (TATA box–binding protein), TAF1 (TBP-associated factor 1), and POL2 (polymerase 2)], and chromatin remodeling (CTCF, RAD21, and YY1) (Fig. 2E and fig. S2D), several of which are themselves causative genes of intellectual disability syndromes with neurocristopathic involvement, as in the case of our recently identified Gabriele–de Vries syndrome caused by YY1 haploinsufficiency (*31*, *32*). Chromatin remodeling was indeed the most prominently enriched group within the overall domain of transcriptional regulation. Two master regulators are particularly noteworthy, as they are themselves regulated by BAZ1B dosage. The first is EGR1 (FDR < 0.1), which is itself among the genes inversely correlated with BAZ1B levels, which is implicated in cranial development (in animal models) (*33*, *34*) and whose promoter has been recently shown to feature a bivalent state in human embryonic cranial NC (*23*, *35*). The second is MXI1, identified as master regulator of genes directly following BAZ1B levels (FDR < 0.001), which is itself found among the genes inversely correlated with BAZ1B and is itself a regulator of BAZ1B, pointing to a cross-talk between the two (fig. S2C). Notably, two differentially expressed targets of MXI1, *TGFB2* and *NFIB*, are also involved in intellectual disability and craniofacial defects (*30*, *36*, *37*).

BAZ1B regulates the NC epigenome in a dosage-dependent manner

The transcriptional readout and functional impact of BAZ1B dosage (at the level of NC induction and migration) established its role as a master controller of the NC. We thus predicted, on the basis of its molecular function, that BAZ1B would directly bind to key NC target genes and that for some of these, the binding would be dosage sensitive. These genes would be, in turn, the most likely direct targets to mediate the dosage-dependent transcriptional and functional phenotypes described above. To test this prediction, we set out to both identify BAZ1B direct targets and characterize their promoter and enhancer states, so as to mechanistically link their transcriptional dysregulation with BAZ1B dosage–dependent chromatin binding. Given the absence of chromatin immunoprecipitation (ChIP)–grade BAZ1B antibodies, to carry out our ChIP coupled with sequencing (ChIP-seq) on scr and KD lines, we first designed a tagging strategy to establish, by CRISPR-Cas9 editing, a series of in-frame 3xFLAG endogenously tagged *BAZ1B* alleles in representative iPSCs of the four genotypes (Fig. 3A and fig. S3, A and B). These were then differentiated to NCSCs (fig. S3C) and subjected to ChIP-seq with anti-FLAG antibody, enabling a faithful characterization of BAZ1B genome-wide occupancy across dosages (one tagged allele in WBS, two tagged alleles in atWBS and CTL, and two tagged alleles in the context of 1.5-fold dosages in 7dupASD).

PCA shows a clear separation of the samples by *BAZ1B* copy number, with CTL and atWBS samples clustering more closely and WBS and 7dupASD samples clustering at opposing positions (Fig. 3B). To call NC-specific enhancer regions and promoter-enhancer associations, we exploited for chromatin annotation the unprecedented resolution afforded by the patients' cohort with its underlying variability and proceeded to (i) select chromosomal regions featuring H3K4me1 and H3K27ac in at least two individuals; (ii) exclude regions marked with H3K4me3 in at least two individuals;(iii) eliminate regions bearing a transcription start site (TSS); and (iv) associate each putative enhancer to the closest TSS, identifying a total of 30,8470 putative enhancer regions. Notably, BAZ1B binds 75% of its targets at their enhancer regions (6747 genes), with the remaining 2297 targets bound at promoters (Fig. 3C). In addition, 40% of genes expressed in NC are bound by BAZ1B, either exclusively at enhancers (27.4%) or exclusively at promoters (3.5%) or at both regions (9%). This highlights its pervasiveness within the NC epigenome (Fig. 3C) and is also reflected in the key functional enrichments observed for the BAZ1B direct targets that are also expressed and that include "axon guidance," "tube development," "dendrite development," "outflow tract morphogenesis," "odontogenesis," "wound healing," and "endochondral bone morphogenesis" (Fig. 3D). Many of the phenomena captured by these GO categories (e.g., odontogenesis and endochondral bone morphogenesis) are linked to recent changes in the bone structure of modern (versus archaic) humans, with *Homo sapiens* having characteristically smaller teeth than its extinct relatives.

Last and consistent with the enrichments in NC-defining categories uncovered above, the analysis of BAZ1B bound regions revealed major convergence with the binding motifs of critical NC regulators, including two motifs similar to those of TFAP2A and NEUROG2, and one equally associated to TAL1, TCF12, AP4, and ASCL1

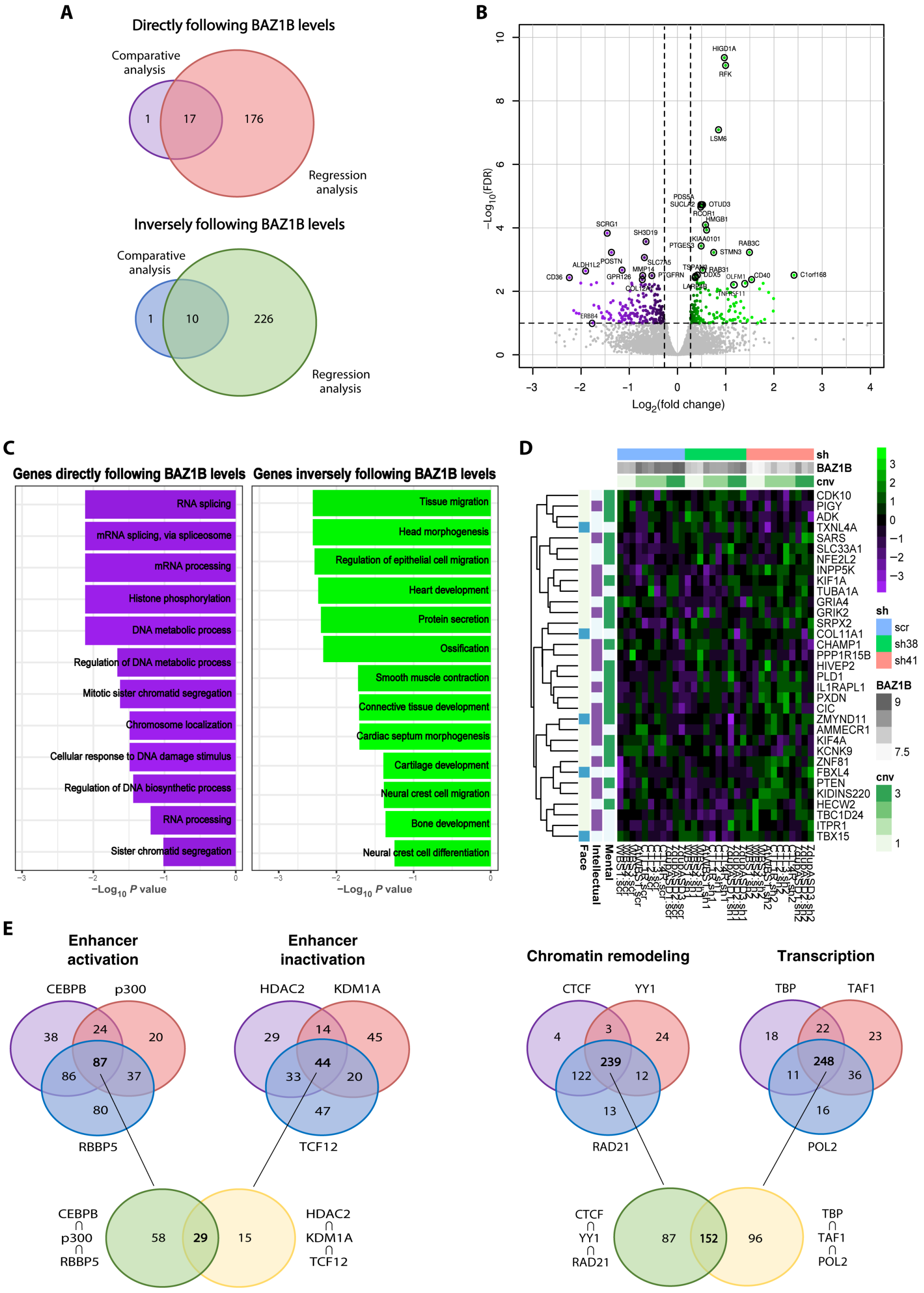


Fig. 2. BAZ1B KD is responsible for transcriptional alterations in NC-related pathways. (**A**) Overlap between genes directly or inversely following BAZ1B levels identified in the pairwise comparative analysis (scr versus shBAZ1B) and in the regression analysis on BAZ1B-level sensitive genes on iPSC-derived NCSCs (FDR < 0.1). (**B**) Volcano plot reporting DEGs identified in the RNA-seq analysis on iPSC-derived NCSCs [fold change (FC) > 1.25; FDR < 0.1]. (**C**) Top most specific enrichments for GO biological processes among the DEGs in the RNA-seq analysis on iPSC-derived NCSCs. (**D**) Heat map representing DEGs that are dysregulated in genetic disorders involving mental retardation ("Mental"), intellectual disability ("Intellectual"), and/or facial dysmorphisms ("Face") according to OMIM database classification. cnv, copy number variant. (**E**) Putative regulators of genes that follow BAZ1B levels identified by a master regulator analysis. Regulators were divided in four different groups based on their main functions.

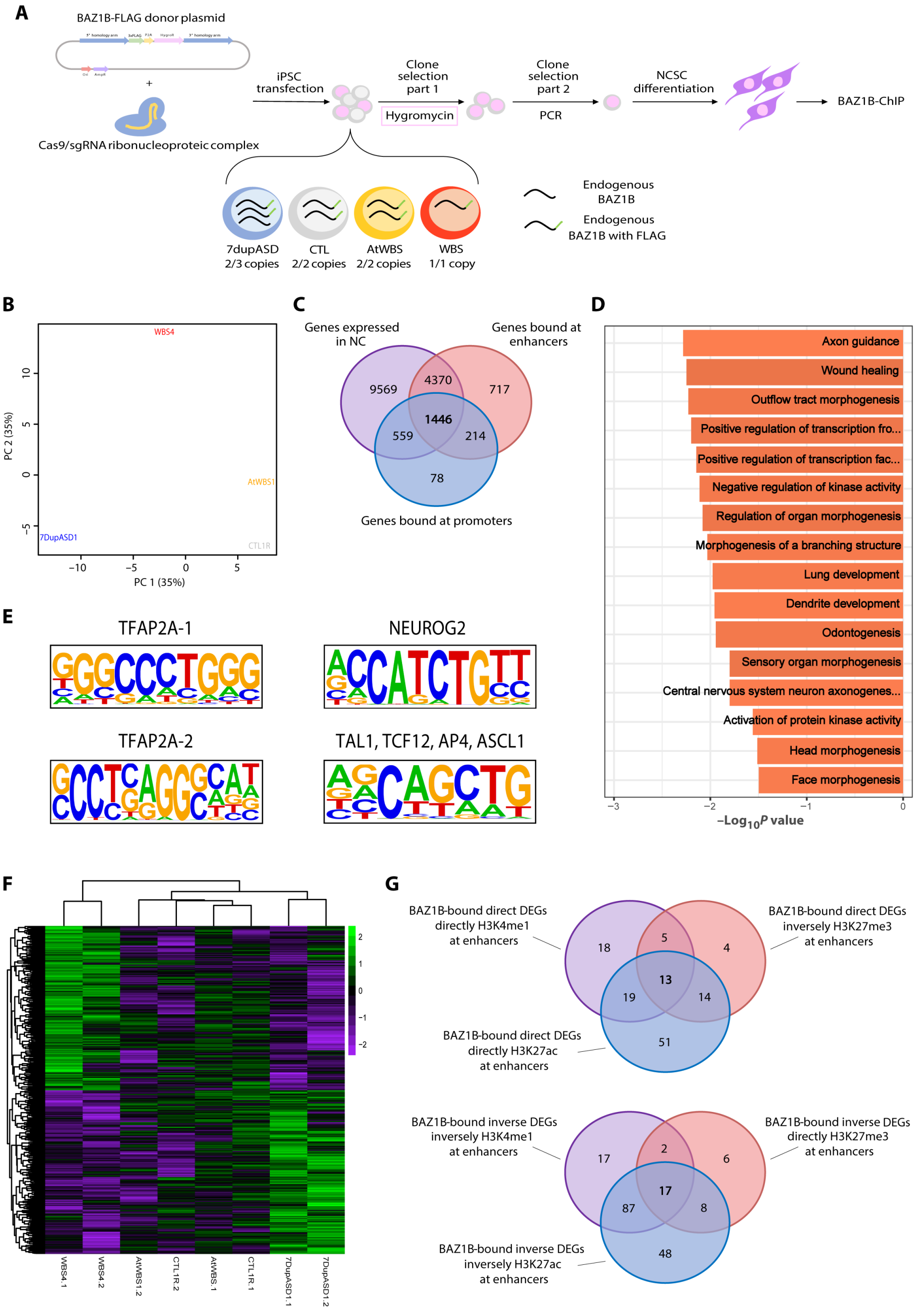


Fig. 3. BAZ1B preferentially binds its targets at their enhancer regions and its KD causes a redistribution of enhancer histone marks. (**A**) Schematic representation of the strategy for CRISPR-Cas9–mediated tagging of endogenous *BAZ1B*. Briefly, iPSCs from the four genotypes were electroporated with the donor plasmid and the Cas9/single-guide RNA ribonucleoprotein complex; clones were selected via hygromycin and PCR, differentiated to NCSCs, and then subjected to ChIP-seq. (**B**) PCA showing the distribution of the four *BAZ1B*-tagged NCSC lines according to their chromatin profiles. (**C**) Overlap between genes expressed in our NCSC lines (purple) and genes bound by BAZ1B at their enhancer (red) or promoter (blue) regions. (**D**) Top most specific enrichments for GO biological processes among the genes that are bound by BAZ1B and expressed in our NCSC cohort. (**E**) Most represented BAZ1B DNA binding motifs identified by HOMER show high similarity to neural and NCSC-specific transcription factors motifs. (**F**) BAZ1B differentially bound regions according to its copy number (FDR < 0.1; $n = 2$). (**G**) Overlap between genes that are differentially expressed have their enhancers differentially marked concordantly (H3K27ac, H3K4me1, and H3K27me3) and are bound by BAZ1B at enhancers.

(Fig. 3E and text S1A). Thus, BAZ1B binding regions are enriched for target sites of major regulators of NC and its neural derivatives (*38*, *39*), among which TFAP2A stands out given its core role in neural border formation and NC induction and differentiation (*40*) through the binding and stabilization of NC-specific enhancers, in concert with NR2F1, NR2F2, and EP300 (*41*).

Last, we identified 81 regions that are quantitatively bound by BAZ1B depending on its copy number (FDR < 0.1) (Fig. 3F), 153 regions differentially bound concordantly in WBS and 7dupASD compared to control and atWBS samples (FDR < 0.1) (fig. S4A), and 176 and 25 regions differentially bound preferentially in WBS (fig. S4B) and 7dupASD (FDR < 0.1) (fig. S4C), respectively.

Given the prominence of its binding to distal regulatory regions, we then set out to define the BAZ1B dosage–dependent impact on NCSC-specific enhancers by integrating H3K27ac, H3K4me1, H3K27me3, and H3K4me3 profiles. We thus performed a regression analysis on *BAZ1B* levels for the distribution of the three histone marks in the aforementioned regions and found H3K27ac to be the most affected, with 7254 genes differentially acetylated at their enhancers, followed by a differential distribution of the H3K4me1 (4048) and H3K27me3 (2136) marks (fig. S4D). This enabled the overlay of epigenomic and transcriptomic profiles, uncovering that among the 1192 DEGs identified in the regression RNA-seq analysis, 21.3% (257 of 1192) are associated to enhancers that are both bound by BAZ1B and differentially H3K27-acetylated in a manner concordant with *BAZ1B* levels (fig. S4E), with a stronger overlap for genes whose expression is inversely correlated with *BAZ1B* levels (160 versus 97). The same held for DEGs that have a concordant differential distribution of H3K4me1 mark at enhancers (123 versus 55), underscoring the consistency of the impact of BAZ1B dosage on distal regulation (fig. S4F). In contrast, a lower number of genes (*36*) showed a concordant differential distribution of the H3K27me3 mark and, at the same time, were bound by BAZ1B at enhancers (fig. S4G), indicating that BAZ1B preferentially affects active chromatin. From this integrative analysis, we could thus lastly identify a core set of 30 bona fide direct targets of BAZ1B, which are genes whose expression tightly follows BAZ1B levels and whose enhancers are bound by BAZ1B and clearly differentially modified (Fig. 3G, fig. S4H, and text S1B). Together, this first dosage-faithful analysis of BAZ1B occupancy in a diverse cohort of human NCSCs establishes its pervasive and mostly distal targeting of the NC-specific epigenome, with a preferential activator role on the critical transcriptional circuits that define NC fate and function.

Intersection with paleogenomic datasets uncovers a key evolutionary role for *BAZ1B*

Mild NC deficits have been put forth as a unifying explanatory framework for the defining features of the so-called domestication syndrome, with *BAZ1B* listed among the putative underlying genes because of its previously reported role in the NC of model organisms (*5*, *10*, *11*). The recent observation that its expression is affected by domestication-related mobile element insertion methylation in gray wolves (*12*) further supported its role in domestication, offering an intriguing parallel to the paleogenomic results that had detected *BAZ1B* within the regions of the modern genome reflective of selective sweeps and found it enriched for putatively regulatory mutations in AMHs (*15*).

Having defined the molecular circuits through which BAZ1B regulates NC, and since NC changes have been implicated in the domestication syndrome (*5*), since craniofacial differences correlate with self-domestication (*1*), and since 7q11.23 dosage-related craniofacial differences in humans relate to the *H. sapiens* versus Neanderthal comparison (fig. S1A), we set out to test the role of BAZ1B dosage in the differences between modern and archaic humans. For this, we carried out a systematic integrative analysis of the overlaps between our empirically defined BAZ1B dosage–sensitive genes (blue Venn in Fig. 4B) and a combination of uniquely informative datasets highlighting differences between modern humans and archaics (Neanderthals/Denisovans) (represented in Fig. 4A by skulls illustrating the more "gracile" and "juvenile" profile in AMH relative to Neanderthals visible in the overall shape of the neurocranium, reduced prognathism, brow ridges, and nasal projections) (*1*, *13*–*15*). Specifically, as shown in Fig. 4B, these datasets include (i) genes associated with signals of positive selection in the modern branch compared to archaic lineages (purple Venn) (*13*, *14*); (ii) genes harboring (nearly) fixed mutations in moderns versus archaics (pink Venn); and (iii) genes associated with signals of positive selection in the four paradigmatic domesticated species dog, cat, cattle, and horse (*1*) (orange Venn), to reveal statistically significant overlaps between them and genes associated with signals of positive selection in the modern branch compared to archaic lineages. In turn, the list of genes harboring (nearly) fixed mutations in moderns versus archaics contains three classes: (i) genes harboring high-frequency changes (*15*), (ii) genes harboring high-frequency missense mutations (red barplot), and (iii) genes enriched for high-frequency mutations in regulatory regions (green barplot) [data based on (*15*)] (Fig. 4C). As shown in the barplots, the obviously very limited number of high-quality coverage archaic genomes available results in a much higher number of nearly fixed changes identified in archaics (left/negative side of the plot) versus modern humans (right side) (Fig. 4C), setting a comparatively much higher threshold for the identification of nearly fixed modern changes.

These analyses are visualized in Fig. 4D (and detailed in tables S2 and S3) through a matrix that intersects all *BAZ1B* dosage–dependent genes (partitioned in the two categories of directly and inversely correlated and ordered across the full range of biological significance and regulatory proximity, from simply DEGs to bona fide direct targets) with the evolutionary changes underlying domestication and self-domestication, yielding the following key insights (color coded for degree of overlap and marked for significance in hypergeometric tests). First, the most significant pattern was obtains at the intersection with the top 10% genes showing an excess of (nearly) fixed mutations in the regulatory regions of modern humans compared to archaics, across both directly and inversely BAZ1B level–dependent genes (table S2). This same category of nearly fixed modern regulatory changes is also the only one that returns a statistically significant overlap with the most stringent core of BAZ1B dosage–dependent targets (i.e., DEGs whose enhancers are both directly bound by BAZ1B and differentially marked upon its decrease), demonstrating that BAZ1B directly controls, in an exquisitely dosage-dependent manner, this coherent and particularly relevant set of genes that underwent regulatory changes in human evolution. Second, the overall strongest overlaps map to the class of genes that are inversely correlated to BAZ1B levels, which we found to be strongly and specifically enriched for head morphogenesis and NC categories (Fig. 2C), thereby confirming craniofacial morphogenesis as the key domain of functionally relevant overlap between BAZ1B dosage and (self-) domestication changes relevant to the evolution of AMHs. Third, despite the spuriously inflated number of apparently fixed mutations in archaics (*15*), the

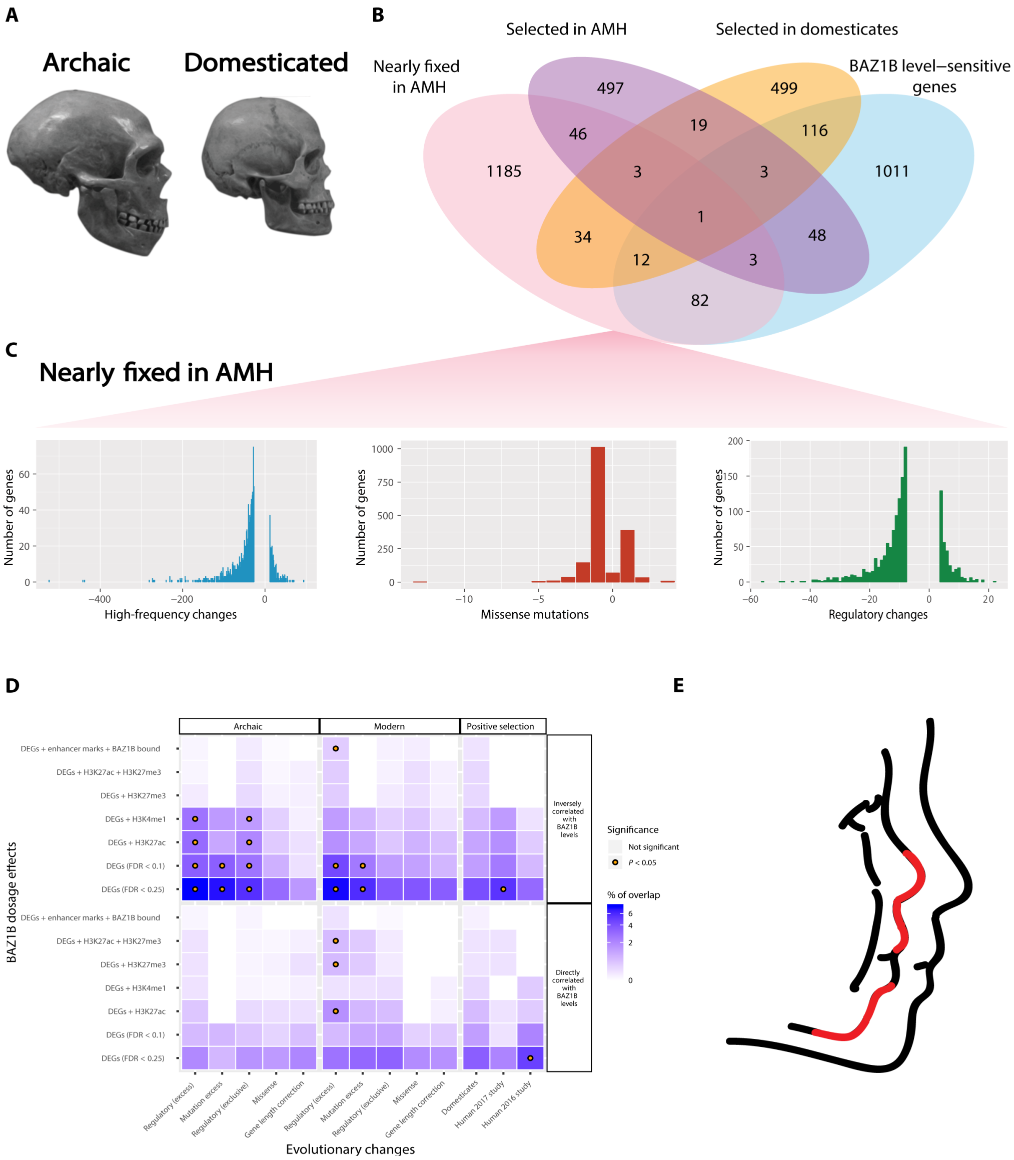


Fig. 4. Exploration of paleogenomic datasets supports a key evolutionary role for *BAZ1B* and validates the self-domestication hypothesis. (**A**) Archaic (Neanderthal) and modern skulls, illustrating the target domesticated phenotype that was captured by our analysis. Skull images were derived from work under a CC BY-SA 2.0 license (https://creativecommons.org/licenses/by-sa/2.0/deed.en) by hairymuseummatt. (**B**) Overlap between *BAZ1B* level–sensitive genes and datasets, which bring out differences between AMHs and archaics, as well as genes under positive selection in modern humans and domesticates. (**C**) Barplots showing the occurrence of high-frequency changes, missense mutations, and mutations in regulatory regions in genes from the AMH (nearly) fixed mutation dataset (pink Venn in B). (**D**) Heat map representing the amount of overlaps for each list selected from (B). Gene overlaps and detailed list descriptions are reported in table S2. (**E**) Rendering of a typical WBS face (left) against the background of a typical modern face (right). Red segments indicate areas of the lower face where the two faces most sharply depart (nose, philtrum, and lower front of the mandible). The lower midface region is most often associated with mutations in genes figuring prominently in our intersections, as discussed in the text and table S3.

overall extent of overlap between genes affected by BAZ1B dosage and our modern and archaic sets does not reveal significantly more hits for archaics. Globally, we found consistently more overlapping genes between the BAZ1B targets and the modern human data and even no statistically significant overlap for any list of the archaic-specific mutations when crossed against genes directly correlated to *BAZ1B* level. We find this noteworthy, given the evidence that the Neanderthal face also displays derived characteristics (*42*) that could be the result of modifications of genes that could overlap with those highlighted in this work. Last, the (lower) midface emerges as a particularly salient area of functionally relevant overlap (as illustrated in Fig. 4E and detailed in table S3), given the specific genes that our analysis unearthed: (i) *COL11A1*, one of the few craniofacial genes highlighted across domestication studies (dog, house sparrow, and pig breeds), which lies in a region of the human genome that resisted archaic introgression (*13*) and is associated with Marshall syndrome; (ii) *XYLT1*, one of the five genes (along with *ACAN*, *SOX9*, *COL2A1*, and *NFIX*) that affect lower and midfacial protrusion, are among the top differentially methylated genes compared to archaics and were also highlighted in a recent study on regulatory changes that shaped the modern human facial and vocal anatomy (tables S1 and S3) (*43*); and (iii) *NFIB*, which belongs to the same gene family as *NFIX* and shares some of its functions. In sum, the direct and dosage-sensitive control by BAZ1B of genes that underwent regulatory changes in human evolution and whose altered expression underlies neurocristopathic facial dysmorphisms is consistent with the hypothesis of mild neurocristopathy as the mechanistic core selected in the self-domestication of the modern human face.

DISCUSSION

As recently reconstructed (*3*), the idea of human self-domestication dates back, at least in terms of scientific record, to Johann Friedrich Blumenbach at the onset of the 19th century. Following on his seminal account of domestication systematized in *Variations of Animals and Plants under Domestication* (*44*), Charles Darwin also considered the analogy between modern humans and domesticated species in *The Descent of Man* (*45*), yet his emphasis on controlled breeding as a key aspect of domestication led him to frame domestication and self-domestication as distinct phenomena and thereby leave Blumenbach's intuition largely undeveloped (*46*). Since then, the possibility that the anatomical and cognitive-behavioral hallmarks of AMHs could result from an evolutionary process bearing such significant similarities to the domestication of animals as to share the same underlying cause has been refined into the full-fledged self-domestication hypothesis (*1*, *2*). As recently argued (*1*, *3*), convergent lines of evidence also indicate that self-domestication is temporally aligned with the emergence of AMH, although the process may have acquired further momentum with the gradual expansion of our species (*1*, *3*). However, despite spurring considerable interest, the self-domestication hypothesis has thus far failed to marshal conclusive evidence largely because of two factors: (i) the lack of a coherent explanation, even at a theoretical level, of what developmental and genetic mechanisms could underlie domestication in general and (ii) the absence of suitable experimental systems in which those mechanisms could be specifically tested in the case of human self-domestication. The first problem was tackled by the recent proposition of mild NC deficits as a central and unifying functional layer underlying domestication (*5*). This constituted a major conceptual advance, particularly because it generated the testable hypothesis of an altered NC gene expression program in domesticated species relative to their wild-type ancestors. For humans, given the obvious lack of gene expression data from archaic hominins, we reasoned that this hypothesis could be verified by examining the genetic changes between archaic and modern humans in light of the gene regulatory networks directly inferred from human neurocristopathies. We thus set out to test whether specific human neurodevelopmental disorders, carefully selected on the basis of both craniofacial and cognitive-behavioral traits relevant to domestication, could illuminate the regulatory circuits shaping the modern human face and hence be harnessed for an empirical validation of the self-domestication hypothesis. Specifically, we reasoned that WBS and 7dupASD, through their uniquely informative set of symmetrically opposite phenotypes at the level of face morphology (fig. S1A) and sociality, constituted a paradigmatic test case to probe the heuristic potential of neurodevelopmental disease modeling for the experimental understanding of human evolution. The following key insights confirm the validity of this approach.

First, we identified the 7q11.23 region *BAZ1B* gene as a master regulator of the modern human face on the basis of a molecular and functional dissection in the thus far largest cohort of WBS patient– and 7dupASD patient–specific NCSCs and across an exhaustive range of BAZ1B dosages. Notably, our cohort also included NCSCs from a patient with rare WBS featuring a much milder WBS gestalt and harboring an atypical, *BAZ1B*-sparing deletion that served as a particularly informative control, as confirmed by the clustering of atypical NCSC lines with controls when probed for BAZ1B occupancy. In particular, exploiting the fine-grained resolution of BAZ1B dosages recapitulated in our cohort, we could couple classical pairwise comparisons with a more sophisticated regression analysis on BAZ1B levels, thereby revealing major BAZ1B dosage–dependent transcriptional alterations pivoting around clusters of pathways that are crucial for NC development and maintenance, as well as for its downstream skeletal and cardiac outputs.

Second, we repurposed the versatility of CRISPR-Cas9 to generate an allelic series of endogenously tagged *BAZ1B* across 7q11.23 dosages (including the *BAZ1B*-sparing atypical patient as uniquely relevant control) to define its dosage-dependent genome-wide occupancy. Taking advantage of previous extensive work on the NCSC chromatin landscape (*41*, *47*–*49*), we were able to define a pivotal role for BAZ1B in NCSC enhancer regulation, consistent with its preferential binding of distal regulatory regions, and to partition its dosage-dependent regulation into bona fide direct and indirect targets. The overall balance between the numbers of genes up- or down-regulated upon BAZ1B KD—together with the greater overlap, sheer size, and significance of enrichments in chromatin remodeling categories over other domains of transcription regulation—further corroborates the inclusion of BAZ1B among the factors acting upstream of enhancer and promoter modulations to enable or reinforce rather than specify their net outcome. Last, this molecular readout was translated to the functional level with the definition of an impairment in both NCSC migration and outgrowth from EBs upon decrease in BAZ1B, providing the first validation of BAZ1B involvement in key functions of the developing human NC.

Third, our investigation provides the first experimental evidence for the neurocristopathic hypothesis that had been put forth to explain the domestication syndrome and had pointed to BAZ1B as one of the candidates underlying this syndrome (*5*). Among the key NC hubs affected by *BAZ1B* dosage, we uncovered three additional critical genes—*EDN3*, *MAGOH*, and *ZEB2*—that had also been predicted in the same model because they are associated with behavioral changes found in domesticates, thereby defining a regulatory hierarchy for this coherent set of genes underlying domestication.

Last, the empirical determination of BAZ1B dosage–sensitive genes in NC models from AMHs with accentuated domestication-relevant traits allowed us to expose, in a functionally relevant manner, the genetic differences between modern versus archaic. This brought to the fore the significant convergence between BAZ1B-dependent circuits and genes harboring regulatory changes in the human lineage, reinforcing the notion that regulatory regions contain some of the most significant changes relevant for the modern lineage. This is also reinforced by the recent identification of AMH-specific hypermethylation in the regulatory region of *BAZ1B* itself (*43*).

Last, it is noteworthy that genes implicated in NC development also play significant roles in the establishment of brain circuits that are critical for cognitive processes like language or theory of mind prominently affected in 7q11.23 syndromes. Among the genes downstream of BAZ1B that we uncovered in this study, *FOXP2*, *ROBO1*, and *ROBO2* have long been implicated in brain wiring processes critical for vocal learning in several species (*50*, *51*), including humans, and will warrant further mechanistic dissection in light of the distinctive linguistic profile of WBS individuals. In conclusion, our findings establish the heuristic power of neurodevelopmental disease modeling for the study of human evolution.

MATERIALS AND METHODS

Human samples

Ethics approvals were reported in the study that established the original iPSC cohort (*16*) and also apply to the additional samples included in this study (7dupASD3 and CTL4R).

Fibroblast reprogramming and iPSC culture

WBS1, WBS2, WBS3, WBS4, 7dupASD1, atWBS1, and CTL2 fibroblasts were reprogrammed using the mRNA Reprogramming Kit (Stemgent), while the 7dupASD2 and CTL1R lines were reprogrammed with the microRNA Booster Kit (Stemgent). The CTL3 line was reprogrammed by transfection with the STEMCCA polycistronic lentiviral vector followed by Cre-mediated excision of the integrated polycistron. 7dupASD3 and CTL4R fibroblasts were reprogrammed using the Simplicon RNA Reprogramming Kit (Millipore).

Before differentiation, iPSC lines were cultured on Matrigel hESC-qualified Matrix (BD Biosciences)–coated plates, diluted 1:40 in Dulbecco's minimum essential medium/F-12, and grown in mTeSR 1 medium (STEMCELL Technologies). They were passaged upon treatment with Accutase (Sigma-Aldrich) and then plated in mTeSR 1 medium supplemented with 5 μM Y-27632 (Sigma-Aldrich).

Differentiation

Differentiation into NCSCs was performed as previously described (*52*), with the exception of NCSCs used in the experiment reported in Fig. 1 (D and E) (*19*).

Flow cytometry

NCSCs were detached using Accutase and counted, and 1×10^6 cells per experimental condition were fixed in 4% paraformaldehyde and then blocked in 10% bovine serum albumin. Cells were incubated for 1 hour with primary antibodies conjugated to fluorophores (HNK1–fluorescein isothiocyanate and nerve growth factor receptor–Alex Fluor 647; BD Biosciences). Analyses were performed on a FACSCalibur instrument (BD Biosciences), and data were analyzed with FCS express software (Tree Star). Fluorescence-activated cell sorting characterization for 7dupASD3 and CTL4R lines is reported in fig. S1B; for all the other lines, see (*16*).

Lentiviral vector production and NCSC transfection

BAZ1B KD was performed using validated pLKO.1 TRC vector TRCN0000013338 (referred to as sh1) and TRCN0000013341 (referred to as sh2). A pLKO.1 TRC vector containing a scrambled short hairpin sequence was used as a negative control.

Second generation lentiviral vectors were produced through calcium phosphate transfection of human embryonic kidney 293T cells and ultracentrifugation (2 hours, 20°C, 20,000 rpm).

NCSCs (3 to 4×10^5) were infected upon splitting and then selected by adding puromycin (1 μg/ml) to the medium.

RNA extraction, retrotranscription, and real-time qPCR

RNA was extracted using the RNeasy Micro Plus Kit (QIAGEN) according to the manufacturer's instructions. Retrotranscribed cDNA was obtained from 0.5 to 1 μg of total RNA using the SuperScript VILO retrotranscription kit (Thermo Fisher Scientific).

Real-time qPCR was performed on a 7500 Fast Real-Time PCR system (Applied Biosystems) using SYBR Green Master Mix (Applied Biosystems) as the detecting reagent. A total cDNA amount corresponding to 15 ng of starting RNA was used for each reaction. Each sample was analyzed in triplicate and normalized to *GAPDH*. Relative mRNA quantity was calculated by the comparative cycle threshold (Ct) method using the formula $2^{-\Delta Ct}$ [*BAZ1B*, CCTCGCAGTAAGAAAGCAAAC (forward) and ACTCATCCAGCTCCTTTTGAC (reverse); *GAPDH*, GCACCGTCAAGGCTGAGAAC (forward) and AGGGATCTCGCTCCTGGAA (reverse); *NR2F1*, AGAAGCTCAAGGCGCTACAC (forward) and GGGTACTGGCTCCTCACGTA (reverse); *NR2F2*, GCAAGTGGAGAAGCTCAAGG (forward) and GCTTTCCACATGGGCTACAT (reverse); *TFAP2A*, GCCTCTCGCTCCTCAGCTCC (forward) and CGTTGGCAGCTTTACGTCTCCC (reverse); and *SOX9*, AGTACCCGCACTTGCACAAC (forward) and GTAATCCGGGTGGTCCTTCT (reverse)].

RNA-seq libraries preparation

Library preparation for RNA-seq was performed according to the TruSeq Total RNA sample preparation protocol (Illumina), starting from 250 ng to 1 μg of total RNA. cDNA library quality was assessed in an Agilent 2100 Bioanalyzer using the High Sensitivity DNA Kit. Libraries were sequenced with the Illumina HiSeq machine at a read length of 50–base pair (bp) paired end and a coverage of 35 million of reads per sample.

Protein extraction and Western blot

NCSCs were lysed in radioimmunoprecipitation assay buffer [10 mM tris (pH 8.0), 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, 140 mM NaCl, and 1 mM EDTA] supplemented with protease inhibitor cocktail (Sigma-Aldrich) and 0.5 mM phenylmethylsulfonyl fluoride (Sigma-Aldrich) for 1 hour at 4°C.

Protein extracts (30 to 50 μg per sample) were supplemented with NuPAGE LDS sample buffer (Thermo Fisher Scientific) and 50 mM dithiothreitol (Thermo Fisher Scientific) and denatured at 95°C for 3 min. Then, extracts were run on a precast NuPAGE 4 to 12% bis-tris Gel (Thermo Fisher Scientific) in NuPAGE MOPS SDS Running Buffer (Thermo Fisher Scientific) and transferred to a 0.45-μm nitrocellulose membrane (GE Healthcare) for 1 hour at 100 V in a buffer containing 20% absolute ethanol and 10% 0.25 M tris base and 1.9 M glycine. The membranes were blocked in TBST [50 mM tris (pH 7.5), 150 mM NaCl, and 0.1% Tween 20] and 5% milk for 1 hour, incubated with primary antibodies overnight at 4°C and with secondary antibodies for 1 hour at room temperature. Primary [BAZ1B (Abcam) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH; Millipore)] and secondary antibodies were diluted in TBST and 5% milk. Blots were detected with the ECL Prime Western Blotting Detection Reagents (Sigma-Aldrich) and scanned using the ChemiDoc system (Bio-Rad).

Wound-healing assay

Cells (5×10^4 to 7×10^4) were plated in each of the two Matrigel-coated wells of silicone culture-inserts (Ibidi) attached to six-well culture plates. After 24 hours, the insert was removed, medium was changed to remove dead cells, and time lapse was performed for 24 hours at the rate of one image every 10 min at ×10 magnification; each condition was analyzed in duplicate. Images were acquired with the BX61 upright microscope equipped with a motorized stage from Olympus or the Nikon Eclipse Ti inverted microscope equipped with a motorized stage from Nikon and analyzed with ImageJ.

Endogenous *BAZ1B* tagging via CRISPR-Cas9

iPSCs were pretreated with 10 μM rho kinase inhibitor for 4 hours, and then 2×10^6 cells were electroporated using the Neon system with the Cas9/single-guide RNA ribonucleoprotein complex and the donor plasmid (synthesized by GeneArt). The donor plasmid contained three FLAG tags followed by a self-cleaving peptide (P2A) and a hygromycin resistance (HygroR). The 3xFLAG-P2A-HygroR cassette was flanked by *BAZ1B*-specific homology arms (5′ HA and 3′ HA) to promote homologous recombination and then subcloned into a bacterial backbone (Fig. 3A).

After 48 hours, iPSC medium was supplemented with hygromycin B (50 μg/μl), and selection medium was maintained for 15 days. Fifteen to 20 clones per iPSC line were then subjected to PCR to (i) evaluate the presence of the cassette and the insertion in the correct genomic locus and (ii) distinguish heterozygously tagged from homozygously tagged clones (fig. S3A). We could isolate a clone with a homozygous integration from the CTL, the atWBS, and the typical WBS but not from the 7dupASD line. In the 7dupASD clone, the FLAG tag was present in two of three copies, as shown by a digital PCR analysis (fig. S3B).

Digital PCR

DNA (60 ng) was amplified in a reaction volume containing the following reagents: QuantStudio 3D Digital PCR Master Mix v2 (Thermo Fisher Scientific), Custom TaqMan Copy Number Assays SM 20× FAM labeled (Thermo Fisher Scientific), and TaqMan Copy Number Reference Assay 20× (Thermo Fisher Scientific) VIC labeled (Thermo Fisher Scientific). The mix was loaded on a chip using the QuantStudio 3D Digital PCR Chip Loader. The chips were then loaded on the ProFlex PCR System (Thermo Fisher Scientific), and data were analyzed using the "QuantStudio 3D AnalysisSuite Cloud Software." The entire process was performed by the qPCR Service at Cogentech, Milano [Custom (FLAG) TaqMan Copy Number Assays: forward primer, TGGACAGTCCAGAGGACGAA; reverse primer, CACCCTTGTCGTCATCGTCTT; and probe, FAMACAGAAGA-AGGACTACAAAGACG and TaqMan Copy Number Reference Assay: TERT (VIC) (catalog number 4403316)].

ChIP coupled with sequencing

Approximately 2×10^5 cells were used (~100 μg of chromatin) for histone mark IP, and 1 mg of chromatin was used for BAZ1B-FLAG IP. Cells were fixed with phosphate-buffered saline, containing 1% formaldehyde (Sigma-Aldrich), for 10 min to cross-link proteins and DNA, when the reaction was then stopped by adding 125 mM glycine for 5 min. Cells were lysed with SDS buffer containing 100 mM NaCl, 50 mM tris-HCl (pH 8.0), 5 mM EDTA (pH 8.0), and 10% SDS, at which point chromatin pellets were resuspended in IP buffer containing 1 volume of SDS buffer and 0.5 volume of Triton dilution buffer [100 mM tris-HCl (pH 8.5), 5 mM EDTA (pH 8.0), and 5% Triton X-100]. Chromatin was then sonicated using the S220 Focused-ultrasonicator (Covaris) to generate <300 bp DNA fragments (for histone mark IPs) or the Branson Digital Sonifier to generate 500 to 800 bp DNA fragments (for BAZ1B-FLAG IP).

Sonicated chromatin was incubated overnight at 4°C with primary antibodies [H3K27ac (Abcam), H3K4me1 (Abcam), H3K4me3 (Abcam), H3K27me3 (Cell Signaling Technology), and FLAG (Sigma-Aldrich)] and then for 3 hours with Dynabeads Protein G (Thermo Fisher Scientific). Beads were washed three times with low-salt wash buffer [0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM tris-HCl (pH 8.0), and 150 mM NaCl] and once with high-salt wash buffer [0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM tris-HCl (pH 8.0), and 500 mM NaCl]. Immunocomplexes were eluted in decross-linking buffer (1% SDS and 100 mM NaHCO3) at 65°C for 2 hours. DNA was purified using QIAquick PCR columns (QIAGEN) and quantified with a Qubit dsDNA HS assay kit (Thermo Fisher Scientific). DNA libraries were prepared by the sequencing facility at European Institute of Oncology according to the protocol described by Blecher-Gonen and colleagues (*53*), and DNA was sequenced on the Illumina HiSeq 2000 platform. For the FLAG ChIP, samples were run in duplicate.

RNA-seq analysis

RNA-seq data were quantified using Salmon 0.91 to calculate read counts and transcripts per million in a transcript- and gene-wise fashion, using the quasi-mapping offline algorithm (*54*) on the GRCh38 (National Center for Biotechnology Information) database. edgeR was used for differential gene expression analysis (DEA), using generalized linear regression methods, to identify pattern of differential expression following two different schemes:

1) A factorial analysis based on the definition of one group of scrambled and one group of KD samples to identify genes dysregulated similarly across short hairpins characterized by different efficiencies.

2) A numerical analysis in which log-normalized [Trimmed Mean of M-values (TMM)] BAZ1B levels, as quantified by RNA-seq, was used as independent variable.

All analyses were performed dropping individual variations (~individual+KD or ~individual+BAZ1B) to account for the genetic background of each individual. In particular, this design is expected to permit the identification of genes, which change expression level upon KD even in situations in which genotype-specific makeups would lead BAZ1B-dependent genes to have unique expression levels in scramble lines. In the factorial analysis, DEGs were identified and characterized by filtering for fold change (FC) > 1.25 and FDR < 0.05 unless explicitly indicated.

To our knowledge, performing a regression analysis at a gene-specific level has never been performed. We were able to do this because of the availability of a large set of samples (11 individuals) and because of the two short hairpins robustly respectively reducing BAZ1B expression levels, respectively by ~40 and ~70% in all individuals lines. To validate the quality of our numerical differential expression analysis, we took advantage of HipSci data (*55*, *56*) and iPSCpoweR tools (*29*). We took 50 of 105 possible combinations of 13 random individual RNA-seq data from the healthy HipSci cohort, representing both sexes and having at least two technical replicates per individual. Unfortunately, HipSci does not contain at least 13 individuals with three clones per individual. Thus, we performed four alternative DEAs with edgeR (table S4) on the 50 different random combinations of 13 individuals identified (200 DEAs in total, on 22 samples, two clones per individual), using the same model matrix used for the regression analysis (~individual+BAZ1B) and using BAZ1B levels of scramble and sh2 lines. All analyses identified very low number of spurious DEGs (fig. S2E). Thus, we used the "Edg2" pipeline (table S4) because it does not discard genes with higher variability (Edg2 and Edg4 versus Edg1 and Edg3), and it is based on a better suited algorithm (Edg2 versus Edg4). With our model matrix, filtering by $P < 0.01$ (and FDR < 0.25), using Edg2 on a random HipSci data, we obtained an average of 93.32 DEGs (on average) with a median equal to 43 (table S5). GO enrichments were performed using topGO R package version 2.28.0.

Master regulatory analysis was performed via hypergeometric test by measuring gene set enrichments in lists of transcription factor targets provided by the TFBS tools database (*57*). Both GO and transcription factor enrichment analyses were performed considering background genes expressed in at least two samples in our NCSC cohort.

ChIP-seq analysis

ChIP-seq experiments were analyzed both qualitatively and quantitatively. Reads were trimmed with the FASTX-Toolkit (-Q33 -t 20 -l 22), aligned with Bowtie 1.0 (-v 2 -m 1) on the Human hg38 reference genome, and peaks were called using MACS 2.1.1. H3K4me1, H3K27ac, H3K4me3, and H3K27me3 peaks were called with --broad using default parameters and $q < 0.05$.

Qualitative analysis, including intersection and comparison of bed files, was performed using BedTools version 2.23.

To define enhancer regions, we intersected those marked by H3K4me1 and H3K27ac in at least two samples, discarded regions with H3K4me3 in at least two samples, and discarded regions overlapping with TSS. Motif enrichment was performed by using HOMER v4.10.

Quantification of reads per region was performed with DeepTools 3.0.2. Differential mark deposition was conducted by means of edgeR 3.24.1 inside R 3.3.3. To define mark deposition following BAZ1B levels, we used the same design as for RNA-seq data (~individual+BAZ1B).

To identify BAZ1B bound regions and to avoid losing identification of lowly covered regions, we resorted to (i) aggregation of all sample aligned reads and (ii) peak calling with MACS2 using –extsize 800 and $q < 0.25$. BAZ1B binding coverage was calculated with DeepTools, with the same parameters used for histone marks, on the identified peak regions. Differentially bound regions were identified with edgeR.

Assembly of archaic and modern human lists

The archaic/modern lists were generated from the material presented in (*15*). We used high-coverage genotypes for three archaic individuals: one Denisovan (*58*), one Neanderthal from the Denisova cave in Altai mountains (*59*), and another Neanderthal from Vindija cave, Croatia (*60*). The data are publicly available at http://cdna.eva.mpg.de/neandertal/Vindija/VCF/, with the human genome version hg19 as reference. High-frequency (HF) differences were defined as positions where more than 90% of present-day humans carry a derived allele, while at least the Denisovan and one Neanderthal carry the ancestral allele. High-frequency changes in archaics were defined as occurring at less than 1% in present-day humans, while at least two archaic individuals carry the derived allele. The HF lists used here were examined as presented in (*15*), with the exception of the HF lists in regulatory regions, which were extracted from the same dataset but not presented as such in the original paper.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/12/eaaw7908/DC1
Fig. S1. BAZ1B KD validation in iPSC-derived NCSCs and evaluation of its impact on NCSC migration.
Fig. S2. BAZ1B KD affects the transcriptome of iPSC-derived NCSCs.
Fig. S3. Generation of BAZ1B-FLAG iPSC lines and differentiation to NCSCs.
Fig. S4. BAZ1B KD induces a significant chromatin remodeling at distal regions.
Text S1A. Detailed description of HOMER motif enrichments performed on BAZ1B ChIP-seq data.
Text S1B. List of key direct targets of BAZ1B involved in neural- and NC-related development and relevant associated literature.
Table S1. Genes relevant for NC and NC-derived features whose expression follows BAZ1B levels.
Table S2A. Significant genes in human evolution.
Table S2B. Regulatory excess in archaic humans, overlap with BAZ1B targets.
Table S2C. Mutation excess in archaic humans, overlap with BAZ1B targets.
Table S2D. Regulatory changes (exclusive) in archaic humans, overlap with BAZ1B targets.
Table S2E. Missense mutations in archaic humans, overlap with BAZ1B targets.
Table S2F. Mutation excess in archaic humans corrected for length, overlap with BAZ1B targets.
Table S2G. Regulatory excess in modern humans, overlap with BAZ1B targets.
Table S2H. Mutation excess in modern humans, overlap with BAZ1B targets.
Table S2I. Regulatory changes (exclusive) in modern humans, overlap with BAZ1B targets.
Table S2J. Missense mutations in modern humans, overlap with BAZ1B targets.
Table S2K. Mutation excess in modern humans corrected for length, overlap with BAZ1B targets.
Table S2L. Genes under positive selection in domesticated animals, overlap with BAZ1B targets.
Table S2M. Genes under positive selection from Peyrégne *et al.* (*13*) in modern humans, overlap with BAZ1B targets.
Table S2N. Genes under positive selection from Racimo (*14*) in modern humans, overlap with BAZ1B targets.
Table S3. Crucial genes identified in the overlap between BAZ1B datasets and archaic versus modern human datasets reported in this study.
Table S4. Alternative differential expression analysis functions tested with iPSCpower to assess the efficacy of our design matrix (~individual+BAZ1B). R code provided.
Table S5. Number of genes differentially expressed following BAZ1B data in our numerical analysis compared to an analysis conducted on randomized HipSci data, using Edg2 function (see table S4).
References (*61*–*89*)

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. C. Theofanopoulou, S. Gastaldon, T. O'Rourke, B. D. Samuels, A. Messner, P. T. Martins, F. Delogu, S. Alamri, C. Boeckx, Self-domestication in *Homo sapiens*: Insights from comparative genomics. *PLOS ONE* **12**, e0185306 (2017).
2. B. Hare, Survival of the friendliest: *Homo sapiens* evolved via selection for prosociality. *Annu. Rev. Psychol.* **68**, 155–186 (2017).
3. R. W. Wrangham, *The Goodness Paradox: How Evolution Made Us Both More and Less Violent* (Profile Books Ltd., 2019).
4. M. R. Sánchez-Villagra, C. P. van Schaik, Evaluating the self-domestication hypothesis of human evolution. *Evol. Anthropol.* **28**, 133–143 (2019).
5. A. S. Wilkins, R. W. Wrangham, W. T. Fitch, The "domestication syndrome" in mammals: A unified explanation based on neural crest cell behavior and genetics. *Genetics* **197**, 795–808 (2014).
6. B. R. Pober, Williams-Beuren syndrome. *N. Engl. J. Med.* **362**, 239–252 (2010).
7. S. J. Sanders, A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-de-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, C. E. Mason, K. Bilguvar, P. B. S. Celestino-Soper, M. Choi, E. L. Crawford, L. Davis, N. R. Davis Wright, R. M. Dhodapkar, M. DiCola, N. M. DiLullo, T. V. Fernandez, V. Fielding-Singh, D. O. Fishman, S. Frahm, R. Garagaloyan, G. S. Goh, S. Kammela, L. Klei, J. K. Lowe, S. C. Lund, A. D. McGrew, K. A. Meyer, W. J. Moffat, J. D. Murdoch, B. J. O'Roak, G. T. Ober, R. S. Pottenger, M. J. Raubeson, Y. Song, Q. Wang, B. L. Yaspan, T. W. Yu, I. R. Yurkiewicz, A. L. Beaudet, R. M. Cantor, M. Curland, D. E. Grice, M. Günel, R. P. Lifton, S. M. Mane, D. M. Martin, C. A. Shaw, M. Sheldon, J. A. Tischfield, C. A. Walsh, E. M. Morrow, D. H. Ledbetter, E. Fombonne, C. Lord, C. L. Martin, A. I. Brooks, J. S. Sutcliffe, E. H. Cook Jr., D. Geschwind, K. Roeder, B. Devlin, M. W. State, Multiple

recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* **70**, 863–885 (2011).

8. B. M. vonHoldt, E. Shuldiner, I. J. Koch, R. Y. Kartzinel, A. Hogan, L. Brubaker, S. Wanser, D. Stahler, C. D. L. Wynne, E. A. Ostrander, J. S. Sinsheimer, M. A. R. Udell, Structural variants in genes associated with human Williams-Beuren syndrome underlie stereotypical hypersociability in domestic dogs. *Sci. Adv.* **3**, e1700398 (2017).
9. A. V. Kukekova, J. L. Johnson, X. Xiang, S. Feng, S. Liu, H. M. Rando, A. V. Kharlamova, Y. Herbeck, N. A. Serdyukova, Z. Xiong, V. Beklemischeva, K. P. Koepfli, R. G. Gulevich, A. V. Vladimirova, J. P. Hekman, P. L. Perelman, A. S. Graphodatsky, S. J. O'Brien, X. Wang, A. G. Clark, G. M. Acland, L. N. Trut, G. Zhang, Red fox genome assembly identifies genomic regions associated with tame and aggressive behaviours. *Nat. Ecol. Evol.* **2**, 1479–1491 (2018).
10. C. Barnett, O. Yazgan, H. C. Kuo, S. Malakar, T. Thomas, A. Fitzgerald, W. Harbour, J. J. Henry, J. E. Krebs, Williams syndrome transcription factor is critical for neural crest cell function in *Xenopus laevis*. *Mech. Dev.* **129**, 324–338 (2012).
11. A. Ashe, D. K. Morgan, N. C. Whitelaw, T. J. Bruxner, N. K. Vickaryous, L. L. Cox, N. C. Butterfield, C. Wicking, M. E. Blewitt, S. J. Wilkins, G. J. Anderson, T. C. Cox, E. Whitelaw, A genome-wide screen for modifiers of transgene variegation identifies genes with critical roles in development. *Genome Biol.* **9**, R182 (2008).
12. B. M. vonHoldt, S. S. Ji, M. L. Aardema, D. R. Stahler, M. A. R. Udell, J. S. Sinsheimer, Activity of genes with functions in Human Williams-Beuren syndrome is impacted by mobile element insertions in the gray wolf genome. *Genome Biol. Evol.* **10**, 1546–1553 (2018).
13. S. Peyrégne, M. J. Boyle, M. Dannemann, K. Prüfer, Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res.* **2017**, 1563–1572 (2017).
14. F. Racimo, Testing for ancient selection using cross-population allele frequency differentiation. *Genetics* **202**, 733–750 (2016).
15. M. Kuhlwilm, C. Boeckx, A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci. Rep.* **9**, 8463 (2019).
16. A. Adamo, S. Atashpaz, P. L. Germain, M. Zanella, G. D'Agostino, V. Albertin, J. Chenoweth, L. Micale, C. Fusco, C. Unger, B. Augello, O. Palumbo, B. Hamilton, M. Carella, E. Donti, G. Pruneri, A. Selicorni, E. Biamino, P. Prontera, R. McKay, G. Merla, G. Testa, 7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages. *Nat. Genet.* **47**, 132–141 (2015).
17. C. Fusco, L. Micale, B. Augello, M. Teresa Pellico, D. Menghini, P. Alfieri, M. Cristina Digilio, B. Mandriani, M. Carella, O. Palumbo, S. Vicari, G. Merla, Smaller and larger deletions of the Williams Beuren syndrome region implicate genes involved in mild facial phenotype, epilepsy and autistic traits. *Eur. J. Hum. Genet.* **22**, 64–70 (2014).
18. J. Meng, X. T. Zhang, X. L. Liu, L. Fan, C. Li, Y. Sun, X. H. Liang, J. B. Wang, Q. B. Mei, F. Zhang, T. Zhang, WSTF promotes proliferation and invasion of lung cancer cells by inducing EMT via PI3K/Akt and IL-6/STAT3 signaling pathways. *Cell. Signal.* **28**, 1673–1682 (2016).
19. R. Bajpai, D. A. Chen, A. Rada-Iglesias, J. Zhang, Y. Xiong, J. Helms, C. P. Chang, Y. Zhao, T. Swigut, J. Wysocka, CHD7 cooperates with PBAF to control multipotent neural crest formation. *Nature* **463**, 958–962 (2010).
20. C. Barnett, J. E. Krebs, WSTF does it all: A multifunctional protein in transcription, repair, and replication. *Biochem. Cell Biol.* **89**, 12–23 (2011).
21. R. A. Poot, L. Bozhenok, D. L. C. van den Berg, S. Steffensen, F. Ferreira, M. Grimaldi, N. Gilbert, J. Ferreira, P. D. Varga-Weisz, The Williams syndrome transcription factor interacts with PCNA to target chromatin remodelling by ISWI to replication foci. *Nat. Cell Biol.* **6**, 1236–1244 (2004).
22. A. E. Culver-Cochran, B. P. Chadwick, Loss of WSTF results in spontaneous fluctuations of heterochromatin formation and resolution, combined with substantial changes to gene expression. *BMC Genomics* **14**, 740 (2013).
23. A. Wilderman, J. VanOudenhove, J. Kron, J. P. Noonan, J. Cotney, High-resolution epigenomic atlas of human embryonic craniofacial development. *Cell Rep.* **23**, 1581–1597 (2018).
24. R. F. Spokony, Y. Aoki, N. Saint-Germain, E. Magner-Fink, J.-P. Saint-Jeannet, The transcription factor Sox9 is required for cranial neural crest development in *Xenopus*. *Development* **129**, 421–432 (2002).
25. F. Santagati, F. M. Rijli, Cranial neural crest and the building of the vertebrate head. *Nat. Rev. Neurosci.* **4**, 806–818 (2003).
26. S. Bhatt, R. Diaz, P. A. Trainor, Signals and switches in mammalian neural crest cell differentiation. *Cold Spring Harb. Perspect. Biol.* **5**, a008326 (2013).
27. S. O. Ko, I. H. Chung, X. Xu, S. Oka, H. Zhao, E. S. Cho, C. Deng, Y. Chai, Smad4 is required to regulate the fate of cranial neural crest cells. *Dev. Biol.* **312**, 435–447 (2007).
28. Y. Mishina, T. N. Snider, Neural crest cell signaling pathways critical to cranial bone development and pathology. *Exp. Cell Res.* **325**, 138–147 (2014).
29. P.-L. Germain, G. Testa, Taming human genetic variability: Transcriptomic meta-analysis guides the experimental design and interpretation of iPSC-based disease modeling. *Stem Cell Rep.* **8**, 1784–1796 (2017).
30. I. Schanze, J. Bunt, J. W. C. Lim, D. Schanze, R. J. Dean, M. Alders, P. Blanchet, T. Attié-Bitach, S. Berland, S. Boogert, S. Boppudi, C. J. Bridges, M. T. Cho, W. B. Dobyns, D. Donnai, J. Douglas, D. L. Earl, T. J. Edwards, L. Faivre, B. Fregeau, D. Genevieve, M. Gérard, V. Gatinois, M. Holder-Espinasse, S. F. Huth, K. Izumi, B. Kerr, E. Lacaze, P. Lakeman, S. Mahida, G. M. Mirzaa, S. M. Morgan, C. Nowak, H. Peeters, F. Petit, D. T. Pilz, J. Puechberty, E. Reinstein, J. B. Rivière, A. B. Santani, A. Schneider, E. H. Sherr, C. Smith-Hicks, I. Wieland, E. Zackai, X. Zhao, R. M. Gronostajski, M. Zenker, L. J. Richards, *NFIB* haploinsufficiency is associated with intellectual disability and macrocephaly. *Am. J. Hum. Genet.* **103**, 752–768 (2018).
31. M. Gabriele, A. T. Vulto-van Silfhout, P. L. Germain, A. Vitriolo, R. Kumar, E. Douglas, E. Haan, K. Kosaki, T. Takenouchi, A. Rauch, K. Steindl, E. Frengen, D. Misceo, C. R. J. Pedurupillay, P. Stromme, J. A. Rosenfeld, Y. Shao, W. J. Craigen, C. P. Schaaf, D. Rodriguez-Buritica, L. Farach, J. Friedman, P. Thulin, S. D. McLean, K. M. Nugent, J. Morton, J. Nicholl, J. Andrieux, A. Stray-Pedersen, P. Chambon, S. Patrier, S. A. Lynch, S. Kjaergaard, P. M. Tørring, C. Brasch-Andersen, A. Ronan, A. van Haeringen, P. J. Anderson, Z. Powis, H. G. Brunner, R. Pfundt, J. H. M. Schuurs-Hoeijmakers, B. W. M. van Bon, S. Lelieveld, C. Gilissen, W. M. Nillesen, L. E. L. M. Vissers, J. Gecz, D. A. Koolen, G. Testa, B. B. A. de Vries, *YY1* haploinsufficiency causes an intellectual disability syndrome featuring transcriptional and chromatin dysfunction. *Am. J. Hum. Genet.* **100**, 907–925 (2017).
32. M. Gabriele, A. Lopez Tobon, G. D'Agostino, G. Testa, The chromatin basis of neurodevelopmental disorders: Rethinking dysfunction along the molecular and temporal axes. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **84**, 306–327 (2018).
33. A. P. McMahon, J. E. Champion, J. A. McMahon, V. P. Sukhatme, Developmental expression of the putative transcription factor Egr-1 suggests that Egr-1 and c-fos are coregulated in some tissues. *Development* **108**, 281–287 (1990).
34. J. Dalcq, V. Pasque, A. Ghaye, A. Larbuisson, P. Motte, J. A. Martial, M. Muller, RUNX3, EGR1 and SOX9B form a regulatory cascade required to modulate BMP-signaling during cranial cartilage development in zebrafish. *PLOS ONE* **7**, e50140 (2012).
35. A. Werner, S. Iwasaki, C. A. McGourty, S. Medina-Ruiz, N. Teerikorpi, I. Fedrigo, N. T. Ingolia, M. Rape, Cell-fate determination by ubiquitin-dependent regulation of translation. *Nature* **525**, 523–527 (2015).
36. L. P. Sanford, I. Ormsby, A. C. Gittenberger-de Groot, H. Sariola, R. Friedman, G. P. Boivin, E. L. Cardell, T. Doetschman, TGFβ2 knockout mice have multiple developmental defects that are non-overlapping with other TGFβ knockout phenotypes. *Development* **124**, 2659–2670 (1997).
37. C. Boileau, D. C. Guo, N. Hanna, E. S. Regalado, D. Detaint, L. Gong, M. Varret, S. K. Prakash, A. H. Li, H. d'indy, A. C. Braverman, B. Grandchamp, C. S. Kwartler, L. Gouya, R. L. Santos-Cortez, M. Abifadel, S. M. Leal, C. Muti, J. Shendure, M. S. Gross, M. J. Rieder, A. Vahanian, D. A. Nickerson, J. B. Michel; National, Lung Heart Project Blood Institute (NHLBI) Go Exome Sequencing, G. Jondeau, D. M. Milewicz, TGFβ2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome. *Nat. Genet.* **44**, 916–921 (2012).
38. O. L. Wapinski, T. Vierbuchen, K. Qu, Q. Y. Lee, S. Chanda, D. R. Fuentes, P. G. Giresi, Y. H. Ng, S. Marro, N. F. Neff, D. Drechsel, B. Martynoga, D. S. Castro, A. E. Webb, T. C. Südhof, A. Brunet, F. Guillemot, H. Y. Chang, M. Wernig, Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. *Cell* **155**, 621–635 (2013).
39. S. J. Hong, H. J. Choi, S. Hong, Y. Huh, H. Chae, K. S. Kim, Transcription factor GATA-3 regulates the transcriptional activity of dopamine beta-hydroxylase by interacting with Sp1 and AP4. *Neurochem. Res.* **33**, 1821–1831 (2008).
40. N. de Crozé, F. Maczkowiak, A. H. Monsoro-Burq, Reiterative AP2a activity controls sequential steps in the neural crest gene regulatory network. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 155–160 (2011).
41. A. Rada-Iglesias, S. L. Prescott, J. Wysocka, Human genetic variation within neural crest enhancers: molecular and phenotypic implications. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20120360 (2013).
42. R. S. Lacruz, C. B. Stringer, W. H. Kimbel, B. Wood, K. Harvati, P. O'Higgins, T. G. Bromage, J.-L. Arsuaga, The evolutionary history of the human face. *Nat. Ecol. Evol.* **3**, 726–736 (2019).
43. D. Gokhman, L. Agranat-Tamir, G. Housman, M. Nissim-Rafinia, M. Nieves-Colón, H. Gu, Recent regulatory changes shaped human facial and vocal anatomy. bioRxiv 106955 [**Preprint**]. https://doi.org/10.1101/106955.
44. C. Darwin, *The Variation of Animals and Plants under Domestication* (J. Murray, 1868).
45. C. Darwin, *The Descent of Man: And Selection in Relation to Sex* (J. Murray, 1871).
46. M. Brüne, On human self-domestication, psychiatry, and eugenics. *Philos. Ethics Humanit. Med.* **2**, 21 (2007).
47. P. Claes, J. Roosenboom, J. D. White, T. Swigut, D. Sero, J. Li, M. K. Lee, A. Zaidi, B. C. Mattern, C. Liebowitz, L. Pearson, T. González, E. J. Leslie, J. C. Carlson, E. Orlova, P. Suetens, D. Vandermeulen, E. Feingold, M. L. Marazita, J. R. Shaffer, J. Wysocka, M. D. Shriver, S. M. Weinberg, Genome-wide mapping of global-to-local genetic effects on human facial shape. *Nat. Genet.* **50**, 414–423 (2018).
48. A. Rada-Iglesias, R. Bajpai, S. Prescott, S. A. Brugmann, T. Swigut, J. Wysocka, Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell Stem Cell* **11**, 633–648 (2012).
49. C. Buecker, J. Wysocka, Enhancers as information integration hubs in development: Lessons from genomics. *Trends Genet.* **28**, 276–284 (2012).

50. S. C. Vernes, P. L. Oliver, E. Spiteri, H. E. Lockstone, R. Puliyadi, J. M. Taylor, J. Ho, C. Mombereau, A. Brewer, E. Lowy, J. Nicod, M. Groszer, D. Baban, N. Sahgal, J. B. Cazier, J. Ragoussis, K. E. Davies, D. H. Geschwind, S. E. Fisher, Foxp2 regulates gene networks implicated in neurite outgrowth in the developing brain. *PLOS Genet.* **7**, e1002145 (2011).
51. R. Wang, C. C. Chen, E. Hara, M. V. Rivas, P. L. Roulhac, J. T. Howard, M. Chakraborty, J. N. Audet, E. D. Jarvis, Convergent differential regulation of SLIT-ROBO axon guidance genes in the brains of vocal learners. *J. Comp. Neurol.* **523**, 892–906 (2015).
52. L. Menendez, M. J. Kulik, A. T. Page, S. S. Park, J. D. Lauderdale, M. L. Cunningham, S. Dalton, Directed differentiation of human pluripotent cells to neural crest stem cells. *Nat. Protoc.* **8**, 203–212 (2013).
53. R. Blecher-Gonen, Z. Barnett-Itzhaki, D. Jaitin, D. Amann-Zalcenstein, D. Lara-Astiaso, I. Amit, High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. *Nat. Protoc.* **8**, 539–554 (2013).
54. P.-L. Germain, A. Vitriolo, A. Adamo, P. Laise, V. Das, G. Testa, RNAontheBENCH: Computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Res.* **44**, 5054–5067 (2016).
55. I. Streeter, P. W. Harrison, A. Faulconbridge; The HipSci Consortium, P. Flicek, H. Parkinson, L. Clarke, The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res.* **45**, D691–D697 (2017).
56. H. Kilpinen, A. Goncalves, A. Leha, V. Afzal, K. Alasoo, S. Ashford, S. Bala, D. Bensaddek, F. P. Casale, O. J. Culley, P. Danecek, A. Faulconbridge, P. W. Harrison, A. Kathuria, D. McCarthy, S. A. McCarthy, R. Meleckyte, Y. Memari, N. Moens, F. Soares, A. Mann, I. Streeter, C. A. Agu, A. Alderton, R. Nelson, S. Harper, M. Patel, A. White, S. R. Patel, L. Clarke, R. Halai, C. M. Kirton, A. Kolb-Kokocinski, P. Beales, E. Birney, D. Danovi, A. I. Lamond, W. H. Ouwehand, L. Vallier, F. M. Watt, R. Durbin, O. Stegle, D. J. Gaffney, Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature* **546**, 370–375 (2017).
57. G. Tan, B. Lenhard, TFBSTools: An R/Bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555–1556 (2016).
58. M. Meyer, M. Kircher, M. T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prufer, C. de Filippo, P. H. Sudmant, C. Alkan, Q. Fu, R. Do, N. Rohland, A. Tandon, M. Siebauer, R. E. Green, K. Bryc, A. W. Briggs, U. Stenzel, J. Dabney, J. Shendure, J. Kitzman, M. F. Hammer, M. V. Shunkov, A. P. Derevianko, N. Patterson, A. M. Andres, E. E. Eichler, M. Slatkin, D. Reich, J. Kelso, S. Paabo, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
59. K. Prüfer, F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, G. Renaud, P. H. Sudmant, C. de Filippo, H. Li, S. Mallick, M. Dannemann, Q. Fu, M. Kircher, M. Kuhlwilm, M. Lachmann, M. Meyer, M. Ongyerth, M. Siebauer, C. Theunert, A. Tandon, P. Moorjani, J. Pickrell, J. C. Mullikin, S. H. Vohr, R. E. Green, I. Hellmann, P. L. Johnson, H. Blanche, H. Cann, J. O. Kitzman, J. Shendure, E. E. Eichler, E. S. Lein, T. E. Bakken, L. V. Golovanova, V. B. Doronichev, M. V. Shunkov, A. P. Derevianko, B. Viola, M. Slatkin, D. Reich, J. Kelso, S. Pääbo, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
60. K. Prüfer, C. de Filippo, S. Grote, F. Mafessoni, P. Korlević, M. Hajdinjak, B. Vernot, L. Skov, P. Hsieh, S. Peyrégne, D. Reher, C. Hopfe, S. Nagel, T. Maricic, Q. Fu, C. Theunert, R. Rogers, P. Skoglund, M. Chintalapati, M. Dannemann, B. J. Nelson, F. M. Key, P. Rudan, Ž. Kućan, I. Gušić, L. V. Golovanova, V. B. Doronichev, N. Patterson, D. Reich, E. E. Eichler, M. Slatkin, M. H. Schierup, A. M. Andrés, J. Kelso, M. Meyer, S. Pääbo, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
61. M. Barembaum, T. A. Moreno, C. LaBonne, J. Sechrist, M. Bronner-Fraser, Noelin-1 is a secreted glycoprotein involved in generation of the neural crest. *Nat. Cell Biol.* **2**, 219–225 (2000).
62. D. Huang, Y. Wang, L. Xu, L. Chen, M. Cheng, W. Shi, H. Xiong, D. Zalli, S. Luo, GLI2 promotes cell proliferation and migration through transcriptional activation of *ARHGEF16* in human glioma cells. *J. Exp. Clin. Cancer Res.* **37**, 247 (2018).
63. T. C. Südhof, Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* **455**, 903–911 (2008).
64. G. Chen, J. Sima, M. Jin, K. Y. Wang, X. J. Xue, W. Zheng, Y. Q. Ding, X. B. Yuan, Semaphorin-3A guides radial migration of cortical neurons during development. *Nat. Neurosci.* **11**, 36–44 (2008).
65. E. Betters, Y. Liu, A. Kjaeldgaard, E. Sundström, M. I. García-Castro, Analysis of early human neural crest development. *Dev. Biol.* **344**, 578–592 (2010).
66. Q. Wu, W. Tang, Z. Luo, Y. Li, Y. Shu, Z. Yue, B. Xiao, L. Feng, DISC1 regulates the proliferation and migration of mouse neural stem/progenitor cells through Pax5, Sox2, Dll1 and Neurog2. *Front. Cell. Neurosci.* **11**, 261 (2017).
67. M. C. Horowitz, Y. Xi, D. L. Pflugh, D. G. T. Hesslein, D. G. Schatz, J. A. Lorenzo, A. L. M. Bothwell, Pax5-deficient mice exhibit early onset osteopenia with increased osteoclast progenitors. *J. Immunol.* **173**, 6583–6591 (2004).
68. H. Hsu, D. L. Lacey, C. R. Dunstan, I. Solovyev, A. Colombero, E. Timms, H. L. Tan, G. Elliott, M. J. Kelley, I. Sarosi, L. Wang, X. Z. Xia, R. Elliott, L. Chiu, T. Black, S. Scully, C. Capparelli, S. Morony, G. Shimamoto, M. B. Bass, W. J. Boyle, Tumor necrosis factor receptor family member RANK mediates osteoclast differentiation and activation induced by osteoprotegerin ligand. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3540–3545 (1999).
69. C. A. McGourty, D. Akopian, C. Walsh, A. Gorur, A. Werner, R. Schekman, D. Bautista, M. Rape, Regulation of the CUL3 ubiquitin ligase by a calcium-dependent co-adaptor. *Cell* **167**, 525–538.e14 (2016).
70. K. Oka, M. J. Honda, E. Tsuruga, Y. Hatakeyama, K. Isokawa, Y. Sawa, Roles of collagen and periostin expression by cranial neural crest cells during soft palate development. *J. Histochem. Cytochem.* **60**, 57–68 (2011).
71. H. Rios, S. V. Koushik, H. Wang, J. Wang, H. M. Zhou, A. Lindsley, R. Rogers, Z. Chen, M. Maeda, A. Kruzynska-Frejtag, J. Q. Feng, S. J. Conway, Periostin null mice exhibit dwarfism, incisor enamel defects, and an early-onset periodontal disease-like phenotype. *Mol. Cell. Biol.* **25**, 11131–11144 (2005).
72. J. P. Golding, P. Trainor, R. Krumlauf, M. Gassmann, Defects in pathfinding by cranial neural crest cells in mice lacking the neuregulin receptor ErbB4. *Nat. Cell Biol.* **2**, 103–109 (2000).
73. S. Burden, Y. Yarden, Neuregulins and their receptors: A versatile signaling module in organogenesis and oncogenesis. *Neuron* **18**, 847–855 (1997).
74. G. Andreoletti, E. G. Seaby, J. M. Dewing, I. O'Kelly, K. Lachlan, R. D. Gilbert, S. Ennis, *AMMECR1*: Single point mutation causes developmental delAay, midface hypoplasia and elliptocytosis. *J. Med. Genet.* **54**, 269–277 (2017).
75. B. Tumiene, Ž. Čiuladaitė, E. Preikšaitienė, R. Mameniškienė, A. Utkus, V. Kučinskas, Phenotype comparison confirms *ZMYND11* as a critical gene for 10p15.3 microdeletion syndrome. *J. Appl. Genet.* **58**, 467–474 (2017).
76. J. M. Cobben, M. M. Weiss, F. S. van Dijk, R. de Reuver, C. de Kruiff, W. Pondaag, R. C. Hennekam, H. G. Yntema, A *de novo* mutation in *ZMYND11*, a candidate gene for 10p15.3 deletion syndrome, is associated with syndromic intellectual disability. *Eur. J. Med. Genet.* **57**, 636–638 (2014).
77. K. F. Oram, E. A. Carver, T. Gridley, Slug expression during organogenesis in mice. *Anat. Rec. A Discov. Mol. Cell. Evol. Biol.* **271**, 189–191 (2003).
78. W. S. Wu, S. Heinrichs, D. Xu, S. P. Garrison, G. P. Zambetti, J. M. Adams, A. T. Look, Slug antagonizes p53-mediated z progenitors by repressing *puma*. *Cell* **123**, 641–653 (2005).
79. L. Garbes, K. Kim, A. Rieß, H. Hoyer-Kuhn, F. Beleggia, A. Bevot, M. J. Kim, Y. H. Huh, H. S. Kweon, R. Savarirayan, D. Amor, P. M. Kakadia, T. Lindig, K. O. Kagan, J. Becker, S. A. Boyadjiev, B. Wollnik, O. Semler, S. K. Bohlander, J. Kim, C. Netzer, Mutations in *SEC24D*, encoding a component of the COPII machinery, cause a syndromic form of osteogenesis imperfecta. *Am. J. Hum. Genet.* **96**, 432–439 (2015).
80. A. J. Griffith, L. K. Sprunger, D. A. Sirko-Osadsa, G. E. Tiller, M. H. Meisler, M. L. Warman, Marshall syndrome associated with a splicing defect at the *COL11A1* locus. *Am. J. Hum. Genet.* **62**, 816–823 (1998).
81. E. Roessler, Y. Ma, M. V. Ouspenskaia, F. Lacbawan, C. Bendavid, C. Dubourg, P. A. Beachy, M. Muenke, Truncating loss-of-function mutations of *DISP1* contribute to holoprosencephaly-like microform features in humans. *Hum. Genet.* **125**, 393–400 (2009).
82. J. Punetha, A. Kesari, E. P. Hoffman, M. Gos, A. Kamińska, A. Kostera-Pruszczyk, I. Hausmanowa-Petrusewicz, Y. Hu, Y. Zou, C. G. Bönnemann, M. JĘdrzejowska, NovelCol12A1variant expands the clinical picture of congenital myopathies with extracellular matrix defects. *Muscle Nerve* **55**, 277–281 (2017).
83. T. O'Rourke, C. Boeckx, Converging roles of glutamate receptors in domestication and prosociality. bioRxiv 439869 [**Preprint**]. https://doi.org/10.1101/439869.
84. S. Srivastava, H. Engels, I. Schanze, K. Cremer, T. Wieland, M. Menzel, M. Schubach, S. Biskup, M. Kreiß, S. Endele, T. M. Strom, D. Wieczorek, M. Zenker, S. Gupta, J. Cohen, A. M. Zink, S. B. Naidu, Loss-of-function variants in *HIVEP2* are a cause of intellectual disability. *Eur. J. Hum. Genet.* **24**, 556–561 (2016).
85. H. Lei, Z. Yan, X. Sun, Y. Zhang, J. Wang, C. Ma, Q. Xu, R. Wang, E. D. Jarvis, Z. Sun, Axon guidance pathways served as common targets for human speech/language evolution and related disorders. *Brain Lang.* **174**, 1–8 (2017).
86. E. Lausch, P. Hermanns, H. F. Farin, Y. Alanay, S. Unger, S. Nikkel, C. Steinwender, G. Scherer, J. Spranger, B. Zabel, A. Kispert, A. Superti-Furga, TBX15 mutations cause craniofacial dysmorphism, hypoplasia of scapula and pelvis, and short stature in Cousin syndrome. *Am. J. Hum. Genet.* **83**, 649–655 (2008).
87. M. K. Lee, J. R. Shaffer, E. J. Leslie, E. Orlova, J. C. Carlson, E. Feingold, M. L. Marazita, S. M. Weinberg, Genome-wide association study of facial morphology reveals novel associations with *FREM1* and *PARK2*. *PLOS ONE* **12**, e0176566 (2017).
88. J. Iwata, C. Parada, Y. Chai, The mechanism of TGF-β signaling during palate development. *Oral Dis.* **17**, 733–744 (2011).
89. Y. Zhang, J. Su, J. Yu, X. Bu, T. Ren, X. Liu, L. Yao, An essential role of discoidin domain receptor 2 (DDR2) in osteoblast differentiation and chondrocyte maturation via modulation of *Runx2* activation. *J. Bone Miner. Res.* **26**, 604–617 (2011).

Acknowledgments: We thank the IFOM imaging unit and the IEO genomic unit for the help with, respectively, wound-healing assays and sequencing, and G. Scita and his group for suggestions on migration experiments. **Funding:** This work was funded by the Telethon Foundation (grant number GGP14265 to G.T. and G.M.), the EPIGEN Flagship Project of the

Italian National Research Council (to G.T.), the European Research Council (consolidator grant number 616441-DISEASEAVATARS to G.T.), the Horizon 2020 Innovative Training Network EpiSyStem (to G.T.), Ricerca Corrente granted by the Italian Ministry of Health (to G.T. and G.M.) Giovani Ricercatori granted by the Italian Ministry of Health (to G.T.), Daunia Plast (to G.M.), Fondazione Umberto Veronesi (to P.-L.G.), the Foundation IEO-CCM (fellowship to A.V.), the Spanish Ministry of Economy and Competitiveness/FEDER funds (grant FFI2016-78034-C2-1-P to C.B.), the Generalitat de Catalunya (grant 2017-SGR-341 to C.B. and doctoral fellowships to T.O. and S.S.), the MEXT/JSPS Grant-in-Aid for Scientific Research on Innovative Areas 4903 [Evolinguistics: JP17H06379 (C.B.; PI: K. Okanoya)], a Marie Curie International Reintegration Grant from the European Union (PIRG-GA-2009-256413 to C.B.), the European Social Fund (grant BES-2017-080366 to A.A.), and the Portuguese Foundation for Science and Technology (PhD grant number SFRH/BD/131640/2017 to P.T.M.). M.Z., A.V., A.S., and S.T. conducted this study as fulfillment of their Ph.D. within the European School of Molecular Medicine (SEMM), Milan, Italy. **Author contributions:** M.Z. carried out all the experiments including the generation and characterization of both BAZ1B-interfered lines and BAZ1B-tagged lines, the preparation of RNA-seq libraries, ChIPs, and wound-healing assays. A.V. performed RNA-seq and ChIP-seq bioinformatic analyses with the supervision of P.-L.G. A.A., P.T.M., S.S., T.O., and C.B. performed analyses in Fig. 4. M.L. and A.R.-I. characterized the NC induction phenotype reported in Fig. 1 (D and E). N.M. reprogrammed 7dupASD3 and CTL4R fibroblasts and generated the corresponding NC lines. A.S. helped with cell culture. M.Z. and S.T. designed the tag strategy. M.M. provided theoretical support in IP experiments whose results reinforced the rationale for *BAZ1B* endogenous tagging. G.M. provided 7DupASD3 and CTL4R fibroblast samples through the Genomic and Genetic Disorders Biobank, Telethon Network of Genetic Biobanks, and EuroBioBank Network for biobanking biospecimens (Telethon Foundation, grant number GTB12001G). M.Z., C.B., and G.T. wrote the manuscript with contributions from A.V. and A.S. G.T. conceived, designed, and supervised the study. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The cell lines can be provided by G.T. pending scientific review and a completed material transfer agreement. Requests for the cell lines should be submitted to giuseppe.testa@ieo.it. Additional data related to this paper may be requested from the authors.

Submitted 24 January 2019
Accepted 26 September 2019
Published 4 December 2019
10.1126/sciadv.aaw7908

Citation: M. Zanella, A. Vitriolo, A. Andirko, P. T. Martins, S. Sturm, T. O'Rourke, M. Laugsch, N. Malerba, A. Skaros, S. Trattaro, P.-L. Germain, M. Mihailovic, G. Merla, A. Rada-Iglesias, C. Boeckx, G. 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. *Sci. Adv.* **5**, eaaw7908 (2019).

Science Advances

Dosage analysis of the 7q11.23 Williams region identifies *BAZ1B* as a major human gene patterning the modern human face and underlying self-domestication

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

Sci Adv **5** (12), eaaw7908.
DOI: 10.1126/sciadv.aaw7908

ARTICLE TOOLS	http://advances.sciencemag.org/content/5/12/eaaw7908
SUPPLEMENTARY MATERIALS	http://advances.sciencemag.org/content/suppl/2019/12/02/5.12.eaaw7908.DC1
REFERENCES	This article cites 84 articles, 13 of which you can access for free http://advances.sciencemag.org/content/5/12/eaaw7908#BIBL
PERMISSIONS	http://www.sciencemag.org/help/reprints-and-permissions

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

4 | A chronology of *Homo sapiens*

variants

Fine-grained temporal mapping of derived high-frequency variants supports the mosaic nature of the evolution of *Homo sapiens*

Alejandro Andirkó[1,2], **Juan Moriano**[1,2], **Alessandro Vitriolo**[3,4], **Martin Kuhlwilm**[5], **Giuseppe Testa**[3,4,6], **and Cedric Boeckx**[1,2,7,*]

[1]Universitat de Barcelona, Spain
[2]Universitat de Barcelona Institute of Complex Systems, Spain
[3]University of Milan, Italy
[4]European Institute of Oncology, Italy
[5]Institut de Biologia Evolutiva, CSIC-Universitat Pompeu Fabra, Spain
[6]Human Technopole, Italy
[7]Catalan Institute for Research and Advanced Studies (ICREA), Spain
[*]Correspondence: cedric.boeckx@ub.edu

ABSTRACT

Large-scale estimations of the time of emergence of variants are essential offer precise answers to time-sensitive hypotheses concerning human evolution. 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 in various brain regions. We find a recurrent bimodal distribution of high-frequency variants, but also evidence for specific enrichments of gene categories in various time windows, which brings into prominence the 300-500k time slice. We also find evidence for very early mutations impacting the facial phenotype, and much more recent molecular events linked to specific brain regions such as the cerebellum or the precuneus. Additionally, we present a case study of an evolutionarily relevant gene, *BAZ1B*, and its targets.

Author summary

The timing of mutations is a key question for most research in Human Evolution, but it's a too often overlooked perspective in gene expression studies. We assigned dates to genetic variantion unique to Homo sapiens compared to Neanderthals and Denisovans, as well as other sets of genetic mutations that are interesting in evolution, such as variants inherited from extinct humans. With a temporal classification of genetic mutations in hand, we then applied a machine learning tool and other analysis to predict what genes have changed more in certain time windows, what do those genes do and when. We also used clinical data from a gene known to affect facial bone development in humans to understand how far back in time do mutations affecting it go.

1 Introduction

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 [1]. 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' modernity [2]. We have also come to appreciate the extent of (multidirectional) gene flow between Sapiens and Neanderthals and Denisovans, raising interesting questions about speciation [3, 4, 5, 6]. Last, but not least, it is now well established that our species has a long history. Robust genetic analyses [7] indicate a divergence time between us and other hominins for which 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) [8].

Such a long period of time allows for the distinction between early and late members of our species [8]. Genomic analysis of ancient human remains in Africa reveal deep population splits and complex admixture patterns among populations well before the coalescence of modernity in the fossil record [9, 10]. At the same time, reanalysis of archaic fossils in Africa [11] point to the extended presence of multiple hominins on this continent, with the possibility of 'super-archaic' admixture [12, 13]. Lastly, our deeper understanding of other hominins point to derived characteristics in these lineages that make some of our species' traits more ancestral (less 'modern') than previously believed [14].

In the context of this significant rewriting of our history, we decided to explore the temporal structure of an extended catalog of single nucleotide changes found at high frequency (HF ≥90%) across major modern populations we previously generated on the basis of 3 high-coverage archaic genomes [15]. This catalog aims to offer a richer picture of molecular events setting us apart from our closest extinct relatives. To do so, we took advantage of the Genealogical Estimation of Variant Age (GEVA) tool [16]. 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 section 4). Our overall objective here is to use the temporal resolution afforded by GEVA to to estimate the age of emergence of polymorphic sites, and gain further insights into the complex evolutionary trajectory.

Here, we reveal a bimodal temporal distribution of modern human derived high-frequency variants and provide 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, as well as estimate the age of enhancer regulatory variants for different brain regions. Our enrichment analyses uncovers GO-terms unique to specific temporal windows, prominently facial and behavioral-related terms between 300k and 500k years. With a finer-grained level of scrutiny, our machine learning-based analyses predicting differential gene expression regulation of mapped variants (through [17]) reveals a trend towards downregulation in the aforementioned period (300k-500k years; corresponding to the early emergence of our species). We further identify variant-associated genes whose differential regulation may specifically affect brain structures thought to be derived in late *Homo sapiens* such as the cerebellum and the precuneus. Finally, we delved into the study of *BAZ1B*, for its contribution to our understanding of craniofacial development and human evolution [18]. We found a cluster of variants linked to a specific set of *BAZ1B* targets dated around 300-500k years (within the suggested period of appearance of distinctive facial traits in our species), and characterized a set of older variants that further shed light into the timing of the emergence of the 'modern' human face.

2 Results

The distribution of alleles over time follows a bimodal distribution regardless of the frequency cutoff (Figure 1A; Figure S1), with a global maximum around 40kya (for complete allele counts, see section 4). The two modes of the distribution 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 around 100kya [19], while the older distribution contains the period associated with the divergence between *Homo sapiens* and other *Homo* species [7, 20]. When dividing the modes (at the 300kya time mark), the distribution of variants over time is statistically different between the set of overall derived variants and each of the two HF filtered sets ($p < 0.01$, Kolmogorov–Smirnov test).

In order to divide the data for downstream analysis we considered a k-means clustering analysis (at $k = 3$ and $k = 4$, Figure S2). This clustering method yields a division clear enough to distinguish between early and late *Homo sapiens* specimens after the split with other human species. 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 (Figure 1B). Our basic division consisted of three periods: 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*; a later period from 300kya to 500kya, the period associated with earlier members of our species such as the Jebel Irhoud fossil [21] ; and a third, older period, from 500kya to 1 million year ago, corresponding to the time of the most recent common ancestor with the Neanderthal and Denisovan lineage [22]. Finer-grained time slices were adopted for further analyses (see, e.g., section 2.3).

We note that the distribution goes as far back as 2.5 million years ago (see Figure 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 4), as changes over time in human recombination rates might affect the timing of older variants [16], or to the fact that we don't have genomes for older *Homo* species. Some of these very old variants may have been inherited from them, and lost further down the archaic lineages. In this context, we note that 40% of the genes that exhibit an excess of mutations in the modern lineage and totally lack HF derived variants in other hominins in [15] do not exhibit any single 'recent' (<400kya) HF variant (Fig. S3).

2.1 Variant subset distributions

In an attempt to see if specific subsets of variants had strikingly different distributions over time, 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') [23, 24], and regions under putative positive selection [25], and mapped the HF variants from [15] 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 archaic lineages ('length' and 'excess' lists from [15] – see sec. 4). Finally, we plotted introgressed alleles [23, 26]. A bimodal distribution is clearly visible in all the subsets except the introgression datasets (Figure 1C). Introgressed variants peak locally in the earlier period (0-100kya). The distribution roughly fades after 250kya, in consonance with the possible timing of introgression events [4, 12, 24, 27]. As a case example, we plotted those introgressed variants associated with phenotypes highlighted in Table 1 of [28]. As shown in Figure S4, 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 [26] and [23], follow a significantly different distribution over time ($p < 2.2 - 16$, Kolmogorov–Smirnov test) (Figure 1C).

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 (Figure S5). 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 introgression. 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 (0-100kya) that is absent from the deserts of introgression datasets. Also, as shown in 1E, 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 S6 illustrates this point for two genes that have figured prominently in early discussions of selective sweeps since [3]: *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 in [15]. In addition, we are aware that variants enter the genome at one stage and are likely selected for at a (much) later stage [29, 30]. As such our study differs from the chronological atlas of natural selection in our species presented in [31] (as well as from other studies focusing on more recent periods of our evolutionary history, such as [32]). This may explain some important discrepancies between the overall temporal profile of genes highlighted in [31] and the distribution of HF variants for these genes in our data (Figure S7).

Having said this, our analysis recaptures earlier observations about prominent selected variants, located around the most recent peak, concerning genes such as *CADPS2* ([33], Fig. S8). This study also identifies a large set of old variants, well before 300kya, associated with genes belonging to putative positively-selected regions before the deepest divergence of *Homo sapiens* populations [34], such as *LPHN3*, *FBXW7*, and *COG5* (figure S9).

Finally, we estimated the age of putative regulatory variants of 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 Consortium [35]. We did the same for the modern human HF missense mutations [15]. A comparative plot reveals a similar pattern between the three structures, with no obvious differences in variant distribution (see Fig. S10). The cerebellum contains a slightly higher number of variants assigned to the more recent peak when the proportion to total mapped variants is computed: 15.59% to 14.97% (PFC) and 15.20% (TC). 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. S10).We caution, though, that the overall number of missense variants is considerably lower in comparison to the other three datasets.

2.2 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 package [36]. For the three time windows analyzed (corresponding to the recent peak: 0-300kya; divergence time and earlier peak: 500kya-1mya; and time slot between them: 300kya-500kya), we identified unique and shared gene ontology terms (see Figure 2A and sec. 4). Of note, when we compared the most recent period against the two earlier windows together (from 300kya-1mya), 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 (300-500k) period (Table S2; Figure S11). A summary of terms shared across the three time windows can be seen in Figure S12.

2.3 Gene expression predictions

To see if term-enriched genes are associated with particular expression profiles, we made use of ExPecto [17], a sequence-based tool to predict gene expression *in silico* (see description in section 4). We found that there is a significant skewness towards extreme negative values in the 300kya to 500kya time period that is not so salient in the other windows (as shown in quantile-quantile plots in Fig. S14). This skewness is present but not so salient in the overall set of tissue HF variant-specific expression predictions. A series of Kruskal-Wallis tests show that variants coming from GO-enriched genes have significant differences in their average expression levels in each period (0-300kya, 300-500kya and 500-800kya) compared to the others ($p = 3.411e-05$, $p = 4.032e-08$ and $p = 4.032e-08$, adjusted by Bonferroni).

We applied the ExPecto tool as well to the overall derived HF variant dataset derived from [15], with a particular focus on expression changes in brain tissues.

To examine if certain tissues had a specially high predicted expression value in certain key time windows, we further divided the variants in six chronological groups ranging from the present to an estimated 800kya according to the GEVA set dating (Fig. 3A – see Fig. S15 for full details). Of note is the presence of the cerebellum in a period preceding the last major Out-of-Africa event (as predicted by [37]) in a landscape otherwise dominated by tissues such as the Adrenal Gland, the Pituitary, Astrocytes, and Neural Progenitor Cells.

The six windows (0-60, 60-100, 100-200, 200-300, 300-500 and 500-800kya) attempt to capture events in a finer-grained fashion (see sec. 4). We found that the sum of predicted gene expression values differs across timing windows, as determined by an approximate Kruskal-Wallis Test with random sampling ($n = 1000$) test, but not across tissues. A post-hoc Dunn test shows that expression values predicted by ExPecto are significantly different between the 60-100 and the 200-300 and 300-500 windows ($p = 0.001$ and $p = 0.0012$, p-values adjusted with Benjamini-Hochberg) and between 0-60 and 60-100 ($p = 0.0102$, adjusted). We performed an additional analysis to check whether there is an association between exact dates predicted by the GEVA tool and expression (as opposed to a time window division). The correlation between these two values is not significant ($p = 0.3287$, Pearson correlation test).

The authors of the article describing the ExPecto tool [17] 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 (Figure 3B). Among these we find genes such as *DLL4*, a Notch ligand implicated in arterial formation [38]; *FGF14*, which regulates the intrinsic excitability of cerebellar Purkinje neurons [39]; *SLC6A15*, a gene that modulates stress vulnerability through the glutamate system [40]; 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 species [15].

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-analysis [41] 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 (Splenium), 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 influence human cortical structure [42], one variant (rs75255901) in Table 1, linked to *DAAM1*, 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 [37, 43, 44, 45], and their associated variants are dated amongst the most recent ones in the table.

2.4 Case study

As a case example of the potential of the GEVA dataset when applied to evolutionary questions, we examined HF variants found in *BAZ1B* and target genes. *BAZ1B* is a gene implicated in craniofacial defects in Williams-Beuren syndrome. We recently positioned this gene upstream in the developmental hierarchy of the modern human face on the basis of empirical evidence gathered from neural crest models with interfered gene function [18]. We wanted to determine if HF mutations harbored by *BAZ1B* are temporally accompanied by HF variant changes in a range of target genes that we previously demonstrated cluster in statistically significant ways when examined in an evolutionary context [18]. These targets fall in two broad groups: those genes whose expression patterns change in the same direction as that of *BAZ1B* (labeled “DIR"), and those whose expression patterns go in the opposite direction (labeled “INV"). Experimental validation further refined these two sets of genes and identified *bona fide* direct targets of *BAZ1B* (27DIR and 25INV genes, and, with further filtering, 13DIR and 17INV). We already observed that these two sets of targets overlap significantly with genes harboring (regulatory) HF mutations in modern

Location	rsid	Nearest gene(s)	GWAS trait	Age (GEVA)
20:49070644	rs75994450	PTPN1	Fractional anisotropy measurement, Splenium (Corpus Callosum)	36735.46
14:59669037	rs75255901	DAAM1	Functional connectivity (rfMRI)	39543.24
1:22498451	rs2807369	WNT4	Volume of gray matter in Cerebellum (left)	50060.96
2:63144695	rs17432559	EHBP1	Volume of Corpus Callosum (Posterior)	52290.48
12:2231744	rs75557252	CACNA1C	Functional connectivity (rfMRI)	93924.62
10:92873811	rs17105731	PCGF5	Volume of inferiortemporal gyrus (right)	255792.5
17:59312894	rs73326893	BCAS3	Functional connectivity (rfMRI)	418742.6
22:27195261	rs72617274	CRYBA4	Functional connectivity (rfMRI)	445477.7
2:230367803	rs56049535	DNER	Functional connectivity (rfMRI)	523629.8
16:3687973	rs78315731	DNASE1	Volume of Pars triangularis (left)	698856.5

Table 1. Big40 Brain volume GWAS [41] 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 in [46]).

human genomes compared to archaic human genomes, although for the broadest set of “INV" targets, the overlap resulted statistically significant for extinct human species as well [18].

Out of a total of 289 HF mutations harbored by direct targets of *BAZ1B*, 238 could be mapped via GEVA (Figure 4A-B). We observe that close to 25% of all HF variants associated with both INV and DIR targets are found in the oldest time slices defined by the occurrence of *BAZ1B* HF variants, around 1.3mya. 13% of all these ‘target’ variants are found in the 300-500k time window, and about the same percentage (15%) in the most recent (0-300k) period. In other words, unlike the general variant distribution found throughout this study, we do not find a recent peak of variants associated with *BAZ1B* targets. This is in line with the GO-enrichment results presented above, where we don’t find any enrichment for ‘face’-related terms in the most recent periods.

These results invited us to look more closely into the 300-500k period, which as been independently linked to the emergence of modern facial traits (Jebel Irhoud fossil, [21]), and possibly mark a change in our prosociality captured by the “self-domestication hypothesis" ([47, 48]). This period shows a local increase in HF variants for genes harboring an “excess" of mutations compared to archaics, controlling for gene length [15] (Fig 4C). Mutations in other genes we have previously linked to the earliest stages of self-domestication [49] cluster around this period, as shown in Fig 4C. Among them are other genes belonging to the Williams-Beuren Syndrome critical region (*STX1A*, *GTF2I*), prominent targets of *BAZ1B* implicated in Neural Crest processes (*OLFM1*, *EDN3*, *TGFBR2*), as well as specific classes of genes that modulate glutamate signaling (*GRIK3*, *GRIK2*, *GRM7*, *NETO2*) and hormones (*OXTR*, *AVPR1B*). Interestingly, the most recent HF variants in *FOXP2* we could map belong to that period.

It is noteworthy that HF variants harbored by genes associated with face and vocal tract anatomy that were singled out for their extensive methylation changes in [50] (*SOX9*, *ACAN*, *COL2A1*, *NFIX* and *XYLT1*) cluster (together with other *BAZ1B* HF mutations) in our dataset in a more recent time window (Fig S16), pointing to further refinement of the modern facial phenotype, in line with the authors’ own claims in [50]. It is also worth pointing out that *BAZ1B* (and its targets) harbor several HF mutations going back to as early as 900k, which may indicate that aspects of the ‘modern’ face are indeed as old as some have recently claimed, relying on a characterization of both proteomic and phenotypic characterizations of *Homo antecessor* [14, 51].

3 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 archaeology [52] and as detailed a map as possible of the evolutionary history of our species. 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 consequences [52], 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 in [15].

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. Indeed, the emerging picture is very much mosaic-like in its character, in consonance with recent work in archeology [1].

Our analysis highlights the importance of a temporal window between 300-500k that may well correspond to a significant behavioral shift in our lineage, corresponding to the Jebel Irhoud fossil, but also in other parts of the African continent, to increased ecological resource variability [53], and evidence of long-distance stone transport and pigment use [54]. Other aspects of our cognitive and anatomical modernity emerged much more recently, in the last 150000 years, and for these our analysis points to the relevance of gene expression regulation differences in recent human evolution, in line with [55, 56, 57]. These two salient temporal windows are well represented by the density of HF mutations in genes such as *PTEN*, one of the genes highlighted in [15] as harboring an excess of derived HF mutations on the modern compared to extinct human lineages (Fig S17).

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.

4 Methods

Homo sapiens variant catalog. We made use of a publicly available dataset [15] that takes advantage of the Neanderthal and Denisovan genomes to compile a genome-wide catalog of *Homo sapiens*-specific variation (genome version *hg19*, 1000 genomes project frequency data, dbSNP database). In addition to the full data, the authors offered a subset of the data that includes derived variants at a ≥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 ≥90%, we also considered including a metapopulation-wide variant ≥90% frequency cutoff dataset to this study (Fig 1A). All files (the original full and high-frequency sets and the modified, stricter high-frequency one) are provided in the accompanying code.

GEVA. The Genealogical Estimation of Variant Age (GEVA) tool [16] 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 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 of [16]), although it decreases the amount of sites that can be looked at. We give estimated dates after assuming 29 years per generation, as suggested in [58]. 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 4437804 for our total set of variants, 2294023 where mapped in the GEVA dataset (51% of the original total). For the HF subsets, the mapping improves: 101417 (74% of total) and 48424 (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* tool [17]. *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. We then selected brain and brain-related tissues (as detailed in the code), and divided the variants by time period (0-60kya, 60-100kya, 100-200kya, 200-300kya, 300-500kya and 500-800kya – Fig. S15 and Fig. 3A).

gProfiler2. Enrichment analysis was performed using *gProfiler2* package [36] (hypergeometric test; multiple comparison correction, 'gSCS' method; p-values .01 and .05). Dated variants were subdivided in three time windows (0-300kya, 300kya-500kya and 500kya-1mya) and variant-associated genes (retrieved from [15]) were used as input (all annotated genes for *H. sapiens* in the Ensembl database were used as background). Following [17], variation potential directionality scores were calculated as the sum of all variant effects in a range of 1kb from the TSS. Summary GO figures presented in Figure S12 were prepared with *GO Figure* [59].

For enrichment analysis, the Hallmark curated annotated sets [60] were also consulted, but the dated set of HF variants as a whole did not return any specific enrichment.

Data and Code Availability

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

Author Contributions

Conceptualization: CB & AA & JM; Methodology: CB & AA & JM; Data Curation: AA & JM; Software: AA & JM; Formal analysis: AA & JM; Visualization: CB & AA & JM & AV & MK & GT; Investigation: CB & AA & JM & AV & MK & GT; Writing – original draft preparation: CB & AA & JM; Writing – review and editing: CB & AA & JM & AV & MK & GT; Supervision: CB; Funding acquisition: CB.

Funding statement

CB acknowledges support from the Spanish Ministry of Economy and Competitiveness (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 BBVA Foundation (Leonardo Fellowship). 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). M.K. is supported by "la Caixa" Foundation (ID 100010434), fellowship code LCF/BQ/PR19/11700002.

Competing interests

The authors declare no competing interests.

References

1. 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, DOI: 10.1016/j.tree.2018.05.005 (2018).

2. Sykes, R. W. *Kindred: 300,000 Years of Neanderthal Life and Afterlife.* (Bloomsbury Publishing USA, 2020). OCLC: 1126396038.

3. Green, R. E. *et al.* A Draft Sequence of the Neandertal Genome. *Science* **328**, 710–722, DOI: 10.1126/science.1188021 (2010).

4. Kuhlwilm, M. *et al.* Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* **530**, 429–433, DOI: 10.1038/nature16544 (2016).

5. 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, DOI: 10.1016/j.cell.2018.02.031 (2018).

6. Gokcumen, O. Archaic hominin introgression into modern human genomes. *Am. J. Phys. Anthropol.* **171**, 60–73, DOI: https://doi.org/10.1002/ajpa.23951 (2020).

7. Posth, C. *et al.* Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nat. Commun.* **8**, 16046, DOI: 10.1038/ncomms16046 (2017).

8. Stringer, C. The origin and evolution of Homo sapiens. *Philos. Transactions Royal Soc. B: Biol. Sci.* **371**, 20150237, DOI: 10.1098/rstb.2015.0237 (2016).

9. 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, DOI: 10.1126/science.aao6266 (2017).

10. Prendergast, M. E. *et al.* Ancient DNA reveals a multistep spread of the first herders into sub-Saharan Africa. *Science* **365**, DOI: 10.1126/science.aaw6275 (2019).

11. Grün, R. *et al.* Dating the skull from Broken Hill, Zambia, and its position in human evolution. *Nature* **580**, 372–375, DOI: 10.1038/s41586-020-2165-4 (2020).

12. 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, DOI: 10.1371/journal.pgen.1008895 (2020).

13. Durvasula, A. & Sankararaman, S. Recovering signals of ghost archaic introgression in African populations. *Sci. Adv.* **6**, eaax5097, DOI: 10.1126/sciadv.aax5097 (2020).

14. Lacruz, R. S. *et al.* The evolutionary history of the human face. *Nat. Ecol. & Evol.* **3**, 726–736, DOI: 10.1038/s41559-019-0865-7 (2019).

15. Kuhlwilm, M. & Boeckx, C. A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. *Sci. Reports* **9**, 8463, DOI: 10.1038/s41598-019-44877-x (2019).

16. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLOS Biol.* **18**, e3000586, DOI: 10.1371/journal.pbio.3000586 (2020).

17. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179, DOI: 10.1038/s41588-018-0160-6 (2018).

18. Zanella, M. *et al.* Dosage analysis of the 7q11.23 Williams region identifies BAZ1B as a major human gene patterning the modern human face and underlying self-domestication. *Sci. Adv.* **5**, eaaw7908, DOI: 10.1126/sciadv.aaw7908 (2019).

19. Groucutt, H. S. *et al.* Rethinking the dispersal of Homo sapiens out of Africa. *Evol. Anthropol.* **24**, 149–164, DOI: 10.1002/evan.21455 (2015).

20. Gómez-Robles, A. Dental evolutionary rates and its implications for the Neanderthal–modern human divergence. *Sci. Adv.* **5**, eaaw1268, DOI: 10.1126/sciadv.aaw1268 (2019).

21. Hublin, J.-J. *et al.* New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature* **546**, 289–292, DOI: 10.1038/nature22336 (2017).

22. Bermúdez de Castro, J. M. *et al.* A hominid from the lower Pleistocene of Atapuerca, Spain: possible ancestor to Neandertals and modern humans. *Sci. (New York, N.Y.)* **276**, 1392–1395, DOI: 10.1126/science.276.5317.1392 (1997).

23. Sankararaman, S., Mallick, S., Patterson, N. & Reich, D. The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans. *Curr. biology: CB* **26**, 1241–1247, DOI: 10.1016/j.cub.2016.03.037 (2016).

24. 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, DOI: 10.1016/j.cell.2020.01.012 (2020).

25. Peyrégne, S., Boyle, M. J., Dannemann, M. & Prüfer, K. Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res.* **27**, 1563–1572, DOI: 10.1101/gr.219493.116 (2017).

26. Vernot, B. *et al.* Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239, DOI: 10.1126/science.aad9416 (2016).

27. Petr, M. *et al.* The evolutionary history of Neanderthal and Denisovan Y chromosomes. *Science* **369**, 1653–1656, DOI: 10.1126/science.abb6460 (2020).

28. 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).

29. Zhang, X. *et al.* The history and evolution of the Denisovan-EPAS1 haplotype in Tibetans. *bioRxiv* 2020.10.01.323113, DOI: 10.1101/2020.10.01.323113 (2020).

30. Yair, S., Lee, K. M. & Coop, G. The timing of human adaptation from Neanderthal introgression. *bioRxiv* 2020.10.04.325183, DOI: 10.1101/2020.10.04.325183 (2020).

31. 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).

32. Tilot, A. K. *et al.* The Evolutionary History of Common Genetic Variants Influencing Human Cortical Surface Area. *Cereb. Cortex* DOI: 10.1093/cercor/bhaa327 (2020).

33. Racimo, F. Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. *Genetics* **202**, 733–750, DOI: 10.1534/genetics.115.178095 (2016).

34. 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, DOI: 10.1093/molbev/msaa140 (2020).

35. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464, DOI: 10.1126/science.aat8464 (2018).

36. 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, DOI: 10.1093/nar/gkm226 (2007).

37. Neubauer, S., Hublin, J.-J. & Gunz, P. The evolution of modern human brain shape. *Sci. Adv.* **4**, eaao5961, DOI: 10.1126/sciadv.aao5961 (2018).

38. Pitulescu, M. E. *et al.* Dll4 and Notch signalling couples sprouting angiogenesis and artery formation. *Nat. Cell Biol.* **19**, 915–927, DOI: 10.1038/ncb3555 (2017).

39. 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. *The J. Neurosci. The Off. J. Soc. for Neurosci.* **35**, 6752–6769, DOI: 10.1523/JNEUROSCI.2663-14.2015 (2015).

40. 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, DOI: 10.3109/10253890.2015.1105211 (2016).

41. Smith, S. M. *et al.* Enhanced Brain Imaging Genetics in UK Biobank. *bioRxiv* 2020.07.27.223545, DOI: 10.1101/2020.07.27.223545 (2020).

42. Grasby, K. L. *et al.* The genetic architecture of the human cerebral cortex. *Science* **367**, DOI: 10.1126/science.aay6690 (2020).

43. Theofanopoulou, C. Brain asymmetry in the white matter making and globularity. *Front. Psychol.* **6**, DOI: 10.3389/fpsyg.2015.01355 (2015).

44. Bruner, E. Human Paleoneurology and the Evolution of the Parietal Cortex, DOI: 10.1159/000488889 (2018).

45. Lombard, M. & Högberg, A. Four-Field Co-evolutionary Model for Human Cognition: Variation in the Middle Stone Age/Middle Palaeolithic. *J. Archaeol. Method Theory* DOI: 10.1007/s10816-020-09502-6 (2021).

46. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216, DOI: 10.1038/s41586-018-0571-7 (2018).

47. Theofanopoulou, C. *et al.* Self-domestication in Homo sapiens: Insights from comparative genomics. *PLOS ONE* **12**, e0185306, DOI: 10.1371/journal.pone.0185306 (2017).

48. Godinho, R. M., Spikins, P. & O'Higgins, P. Supraorbital morphology and social dynamics in human evolution. *Nat. Ecol. & Evol.* **2**, 956–961, DOI: 10.1038/s41559-018-0528-0 (2018).

49. O'Rourke, T. & Boeckx, C. Glutamate receptors in domestication and modern human evolution. *Neurosci. & Biobehav. Rev.* **108**, 341–357, DOI: 10.1016/j.neubiorev.2019.10.004 (2020).

50. Gokhman, D. *et al.* Differential DNA methylation of vocal and facial anatomy genes in modern humans. *Nat. Commun.* **11**, 1189, DOI: 10.1038/s41467-020-15020-6 (2020).

51. Welker, F. *et al.* The dental proteome of Homo antecessor. *Nature* **580**, 235–238, DOI: 10.1038/s41586-020-2153-8 (2020).

52. Pääbo, S. The Human Condition—A Molecular Approach. *Cell* **157**, 216–226, DOI: 10.1016/j.cell.2013.12.036 (2014).

53. Potts, R. *et al.* Increased ecological resource variability during a critical transition in hominin evolution. *Sci. Adv.* **6**, eabc8975, DOI: 10.1126/sciadv.abc8975 (2020).

54. Brooks, A. S. *et al.* Long-distance stone transport and pigment use in the earliest Middle Stone Age. *Science* **360**, 90–94, DOI: 10.1126/science.aao2646 (2018).

55. Moriano, J. & Boeckx, C. Modern human changes in regulatory regions implicated in cortical development. *BMC Genomics* **21**, 304, DOI: 10.1186/s12864-020-6706-x (2020).

56. Weiss, C. V. *et al.* The cis-regulatory effects of modern human-specific variants. *bioRxiv* 2020.10.07.330761, DOI: 10.1101/2020.10.07.330761 (2020).

57. Yan, S. M. & McCoy, R. C. Archaic hominin genomics provides a window into gene expression evolution. *Curr. Opin. Genet. & Dev.* **62**, 44–49, DOI: 10.1016/j.gde.2020.05.014 (2020).

58. 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, DOI: 10.1002/ajpa.20188 (2005).

59. Reijnders, M. J. & Waterhouse, R. M. Summary Visualisations of Gene Ontology Terms with GO-Figure! *bioRxiv* 2020.12.02.408534, DOI: 10.1101/2020.12.02.408534 (2020).

60. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems* **1**, 417–425, DOI: 10.1016/j.cels.2015.12.004 (2015).

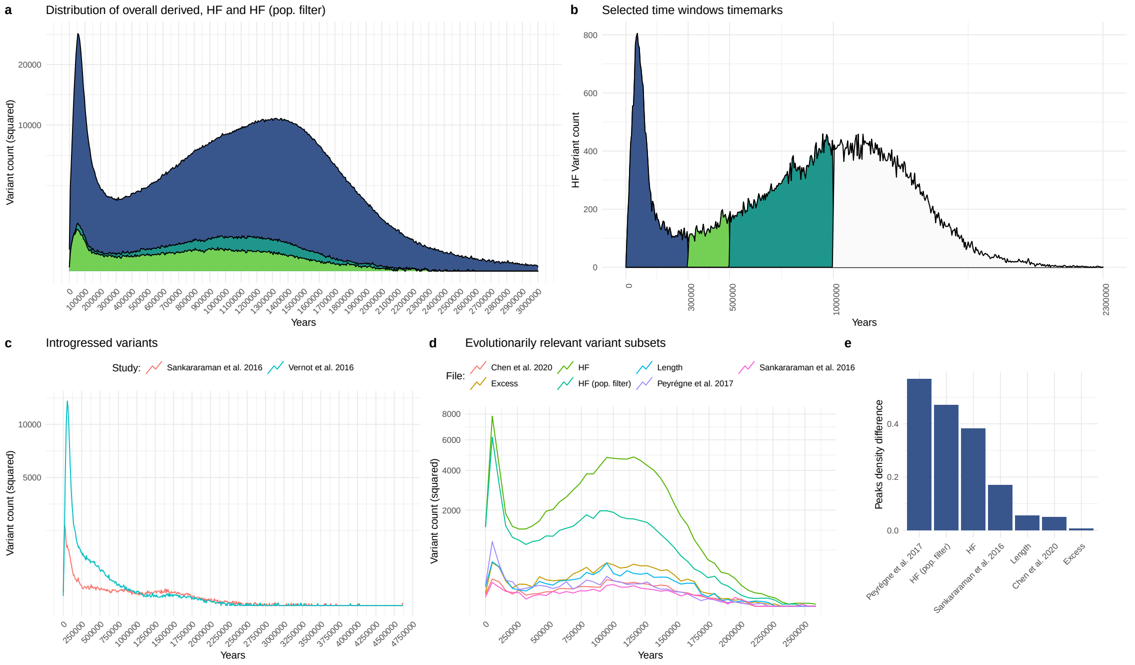


Figure 1. A: Distribution of derived *Homo sapiens* alleles over time with no frequency cutoff, in HF and the modified population-wise HF subset (see sec. 4). Trimmed at 3mya – the full distributions is shown in Fig S1 B: Selected chronological milestones used in our study, as informed by the archaeological record. C: Distribution of introgressed alleles over time, as identified by [23] and [26]. D: Plots of HF variants in datasets relevant to human evolution, including regions under positive selection [25], regions depleted of archaic introgression [23, 24] and genes showing an excess of HF variants ('excess' and 'length') [15]. Variant counts in A, C and D are squared to aid visualization. E: 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.

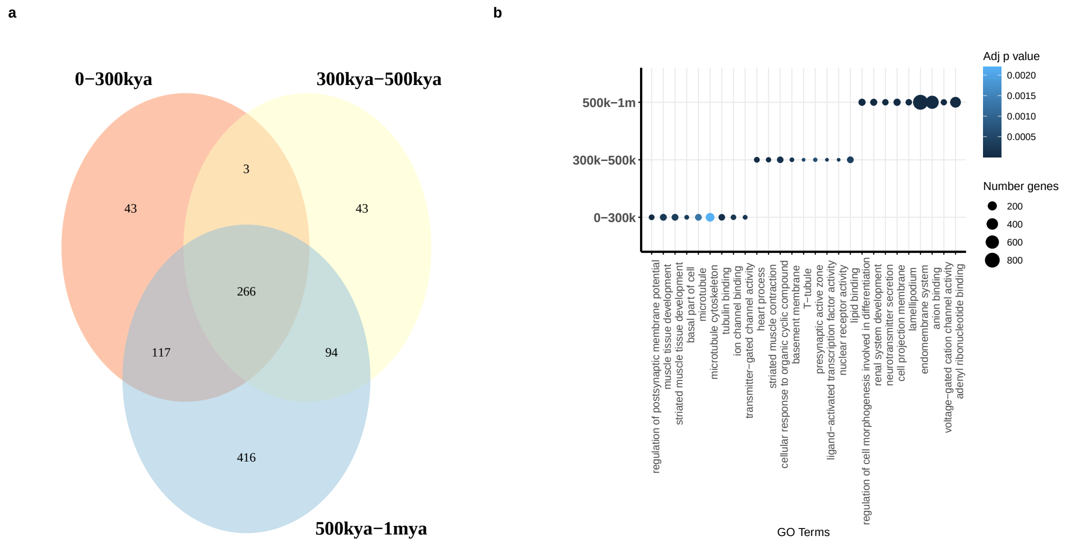


Figure 2. A: Venn diagram of GO terms associated with genes shared across time windows. B: Top GO terms per time window.

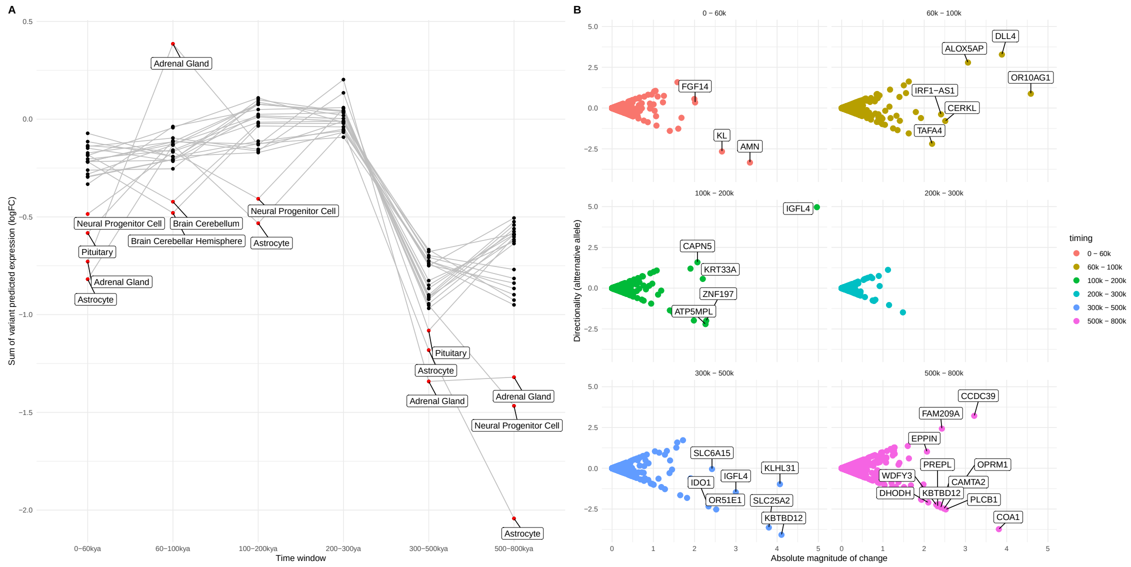


Figure 3. A: Sum of all directional mutation effects within 1kb to the TSS per time window in 22 brain regions from the ENCODE, GTEx and Road map datasets. Highlighted in red, bottom and top values labelled for illustration. Note, however, that expression values predicted are significantly different across time windows but not tissues (as detailed in sec. 2.3). B: Genes with a high sum of all directional mutation effects, and cumulative directionality of expression values.

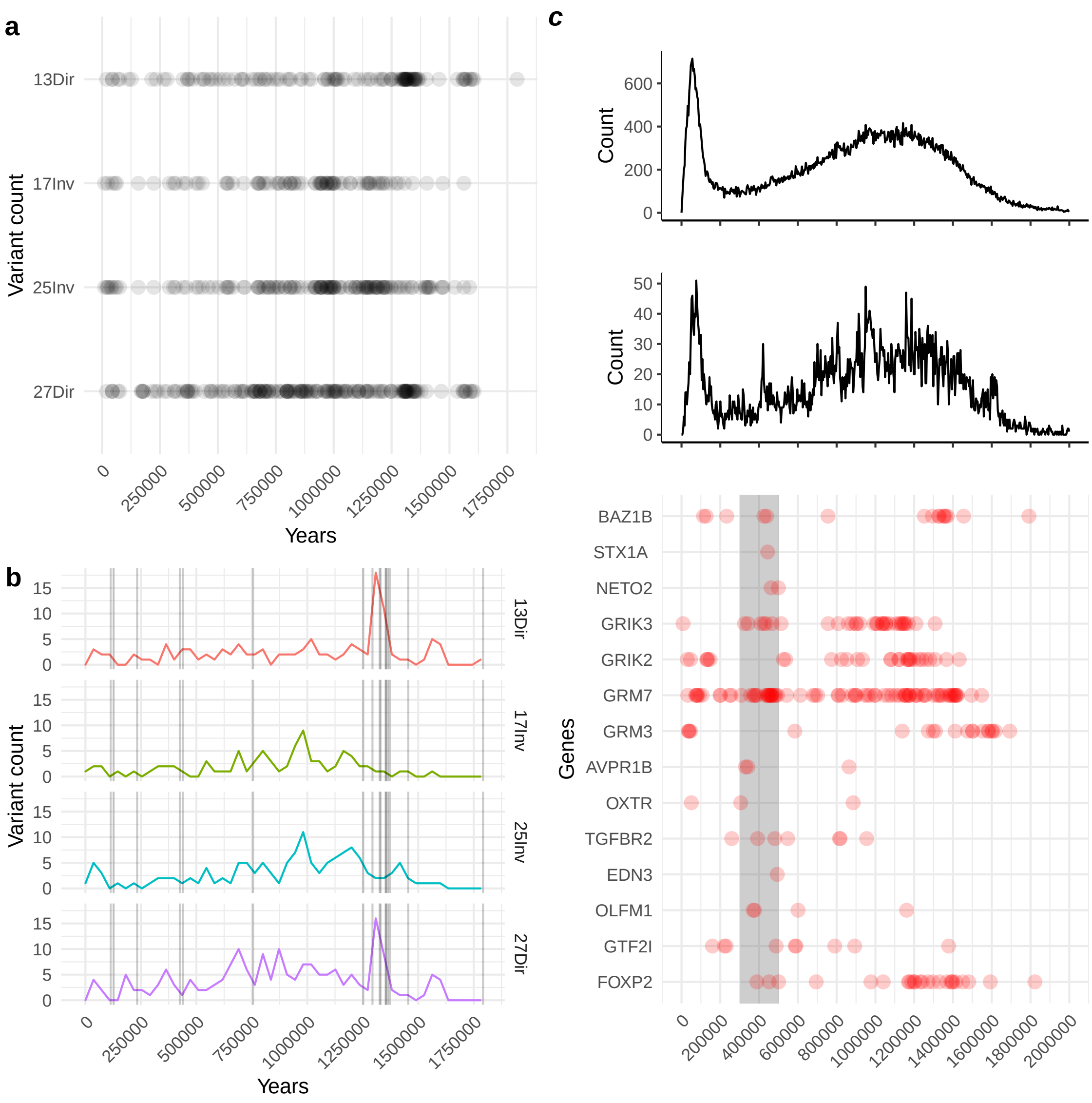


Figure 4. A: Accumulation of variants over time in genes whose expression levels are robustly correlated, directly ('Dir') or inversely ('Inv'), with BAZ1B expression, as per [18]. B: Relation of variant emergence and BAZ1B mutations (vertical black lines) per list of robustly correlated target genes. C: Distribution of HF variants (top), variants in genes showing an excess of HF mutations (middle), and date of emergence of HF variants in selected genes over time (bottom), including a highlight between 300kya and 500kya (in gray). The total number of mapped HF variants for these genes follows a linear relationship with gene length (Fig. S. 18).

Supplementary Figures
for
“Fine-grained temporal mapping of derived high-frequency variants supports the mosaic nature of the evolution of *Homo sapiens*”

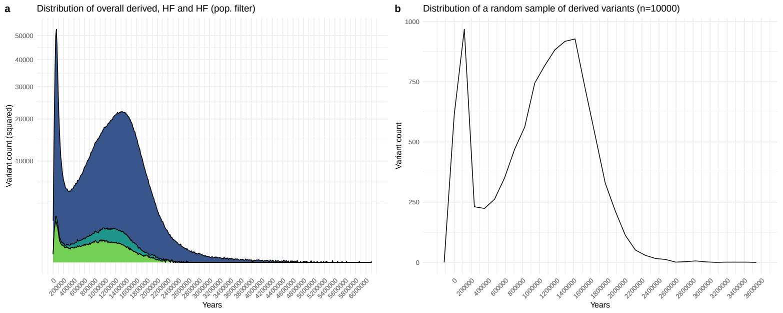


Figure 1: A: Full distribution of derived *Homo sapiens* alleles over time with no frequency cutoff, in HF and the modified population-wise HF subset (see Methods). B: Temporal distribution of a randomly selected sample of derived variants ($n = 10000$).

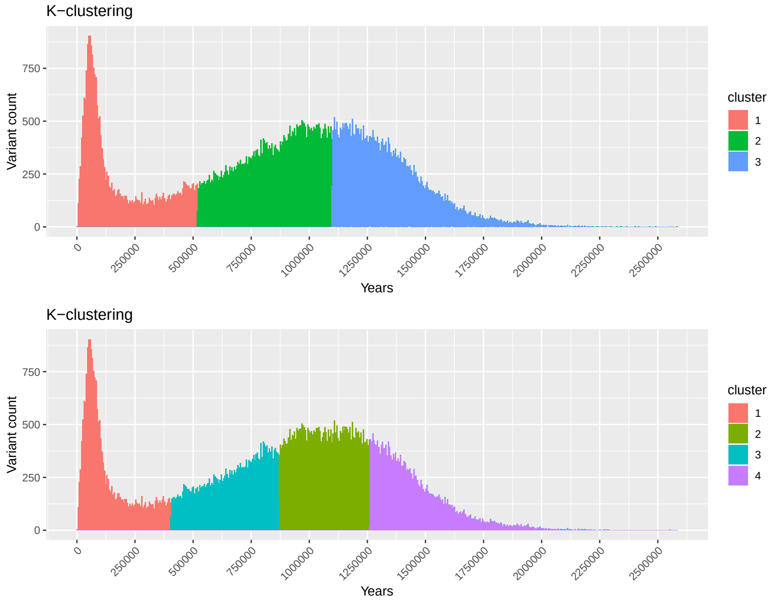


Figure 2: K-means clustering analysis of HF variant temporal distributrion.

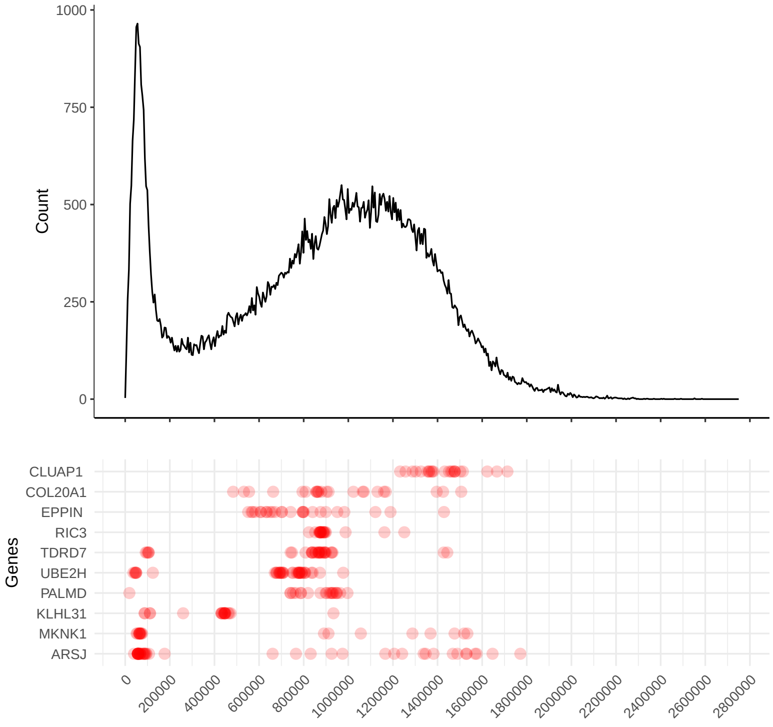


Figure 3: Temporal distribution of variants in genes depleted of archaic-specific variants, as per **?**. On top, overall distribution of the HF variant set.

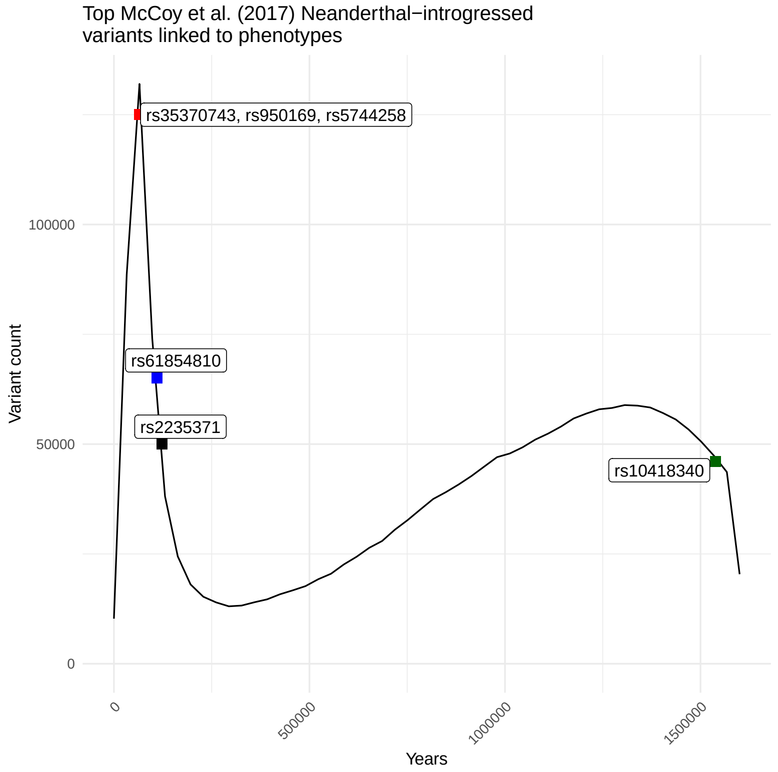


Figure 4: Temporal distribution of introgressed variants linked to phenotypes, as highlighted in Table 1 of **?**, compared to the distribution of all derived variants over time.

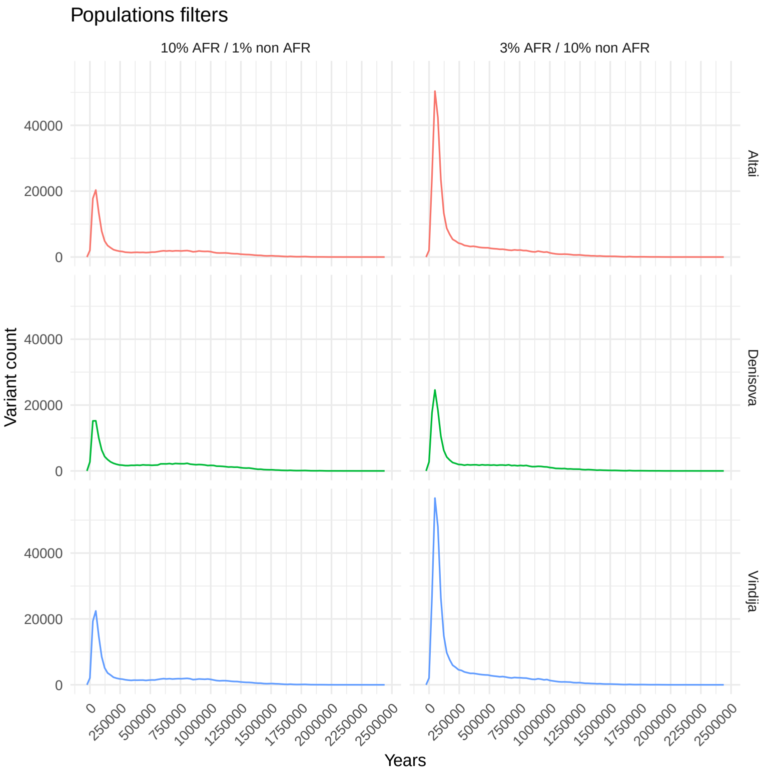


Figure 5: Temporal distribution of variants shared with each of the extinct human genomes after applying specific population frequency filters.These filter include a 10% minor allele frequency cutoff in the African metapopulation (AFR), coupled with a 1% cutoff in the rest of metapopulations, designed 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 introgresed variant genetic pool, accounting for a third of that pool to be introduced in AFR populations by back-to-Africa migrations.

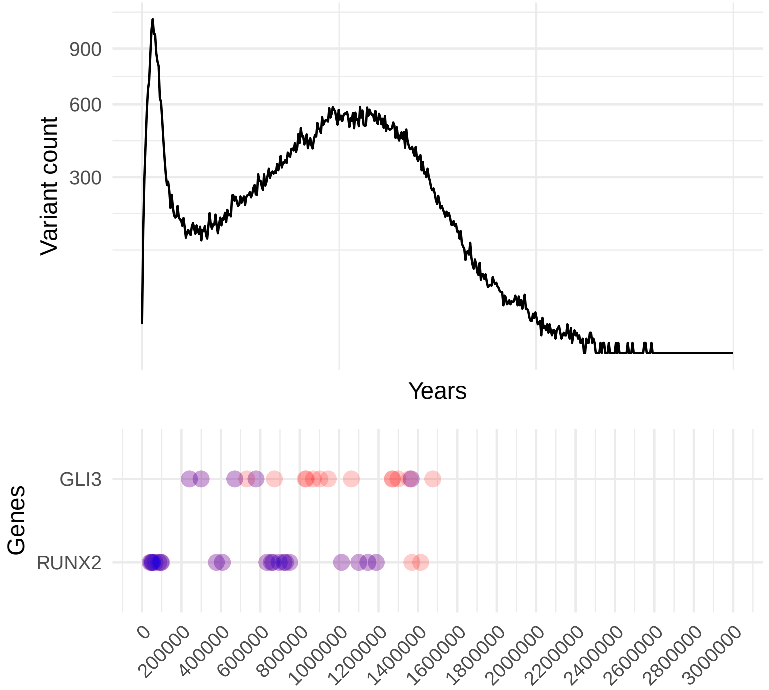


Figure 6: Temporal distribution of HF variants in two genes highlighted in early discussions of selective sweeps: *GLI3* (sweep region from **?** and *RUNX2* (sweep region from **?**). Variants in purple fall within sweep regions.

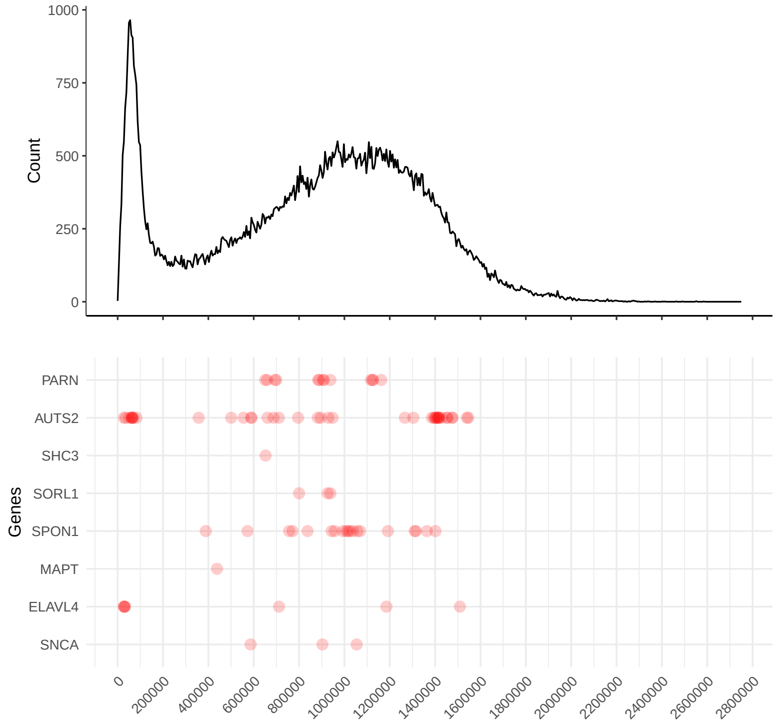


Figure 7: Temporal distribution of variants associated with genes highlighted in **?**.

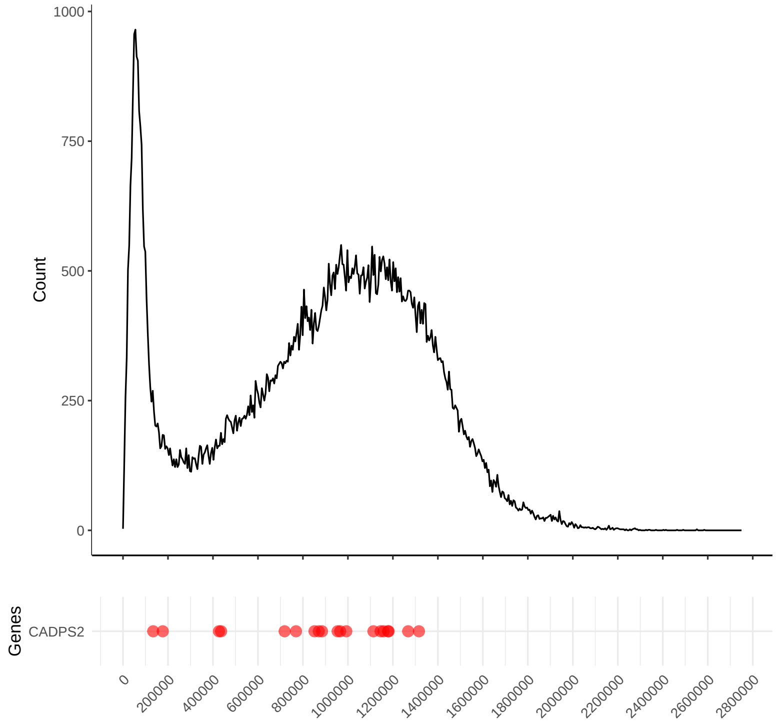


Figure 8: Temporal distribution of variants associated with *CADPS2*. The most recent variants around 200kya in particular capture the reasons this gene was highlighted in **?**: "*CADPS2* was identified in **?** as a candidate for selection . . . . The gene has been suggested to be specifically important 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".

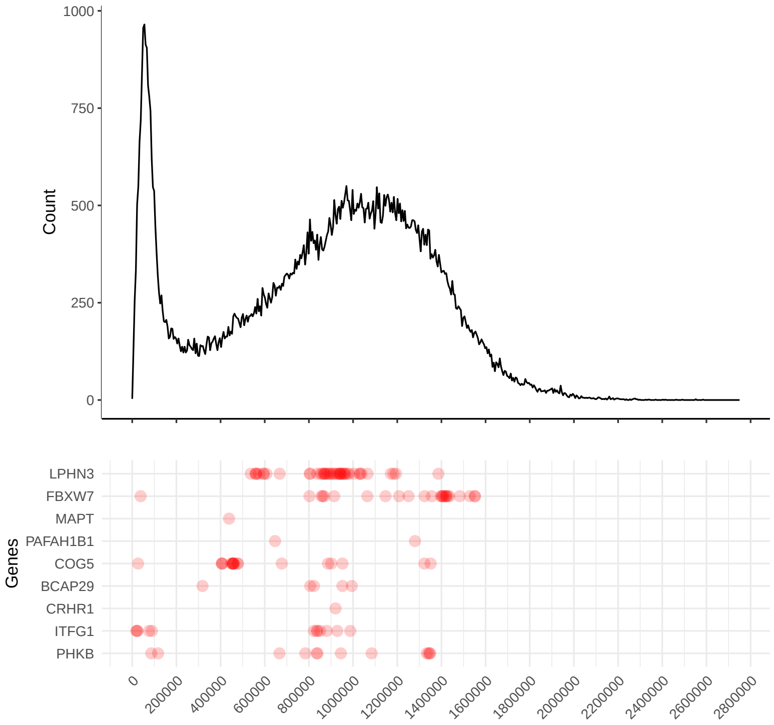


Figure 9: Temporal distribution of variants in genes found in putative positively-selected genetic windows before early *Homo sapiens* population divergence, as per **?**. Genes belonging to putative positively selected regions were retrieved from Supplementary Data, section 12 of **?**.

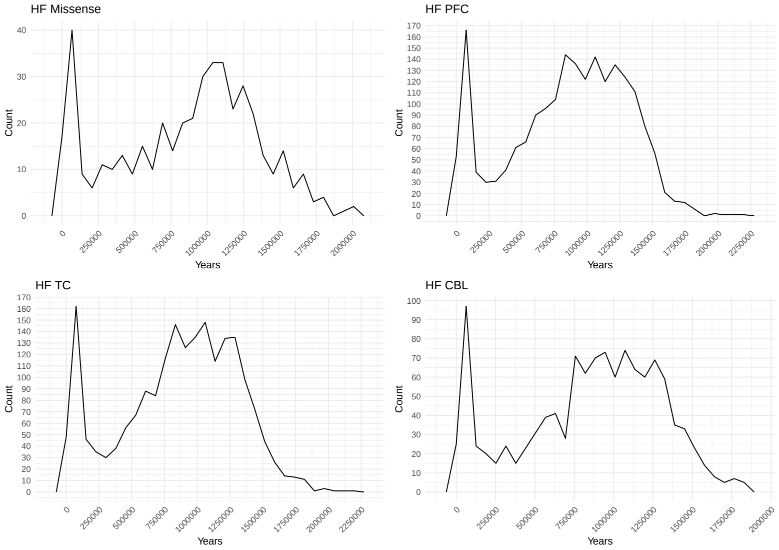


Figure 10: Temporal distribution of high-frequency missense and regulatory variants. Missense variants derived from **?**; enhancer annotations for the prefrontal, temporal and cerebellar cortices were retrieved from **?**. The difference between the two total maximum counts in the left to the right 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.

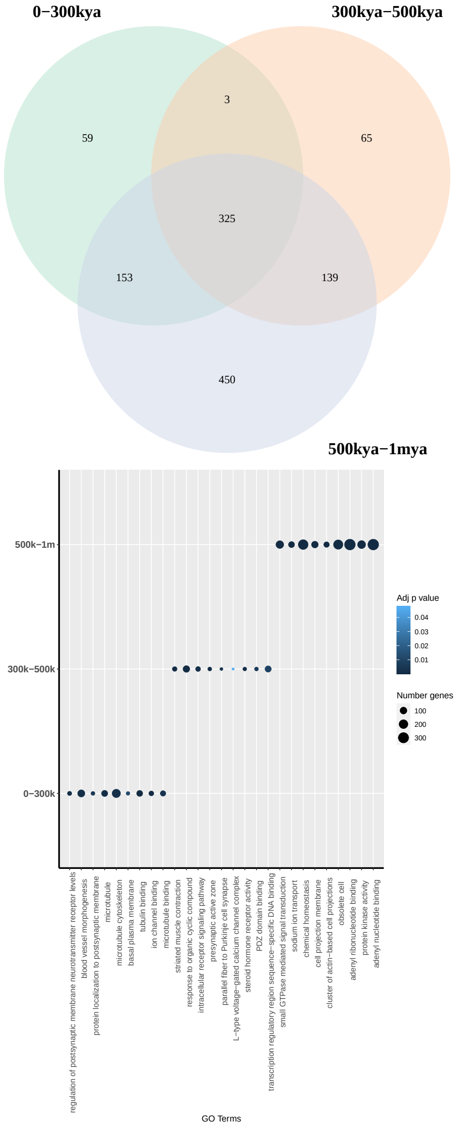
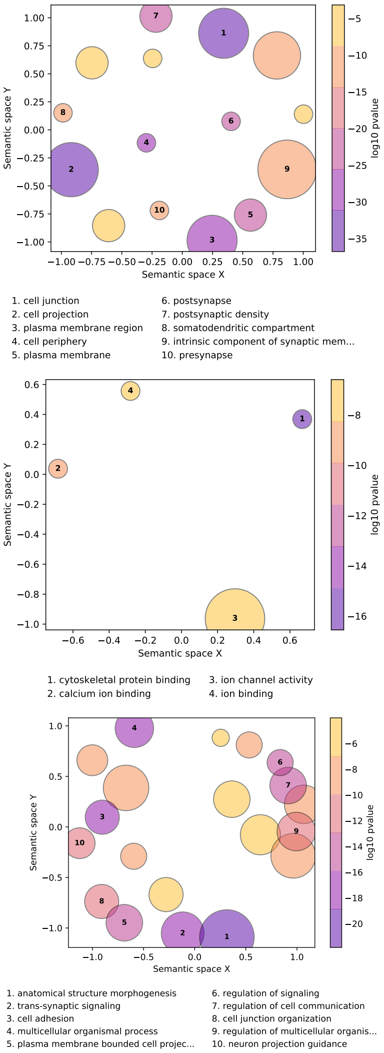


Figure 11: GO terms results when thresholding by an adjusted p-value of 0.05. Venn diagram (top) shows number of unique and shared GO terms across periods. Dot plot (bottom) highlights the top 3 GO terms by significance for each period.



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

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

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

Figure 12: GO term reduction of shared terms across time windows (center of Venn diagram in Fig. 2A)

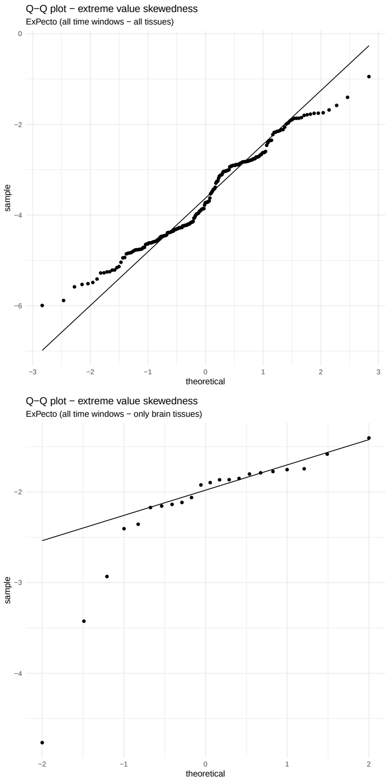


Figure 13: Quantile-quantile plots of predicted expression values skewness for all the tissue models included in **?** and the selected brain-related regions.

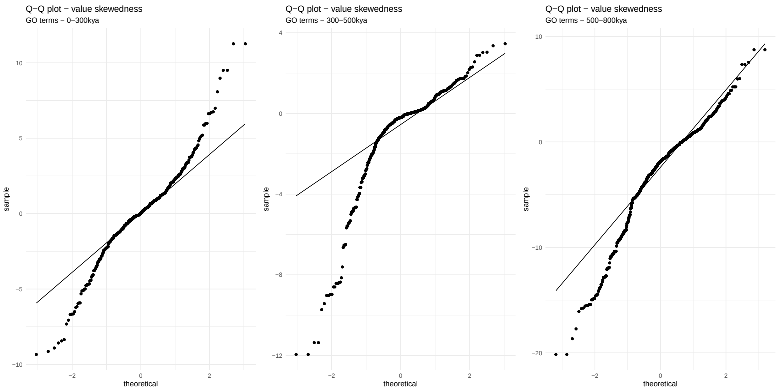


Figure 14: Quantile-quantile plots of predicted expression values of variants associated with GO-enriched genes skewness, divided in three time periods (0-300kya, 300-500kya and 500-800kya). Applied to brain tissue expression only.

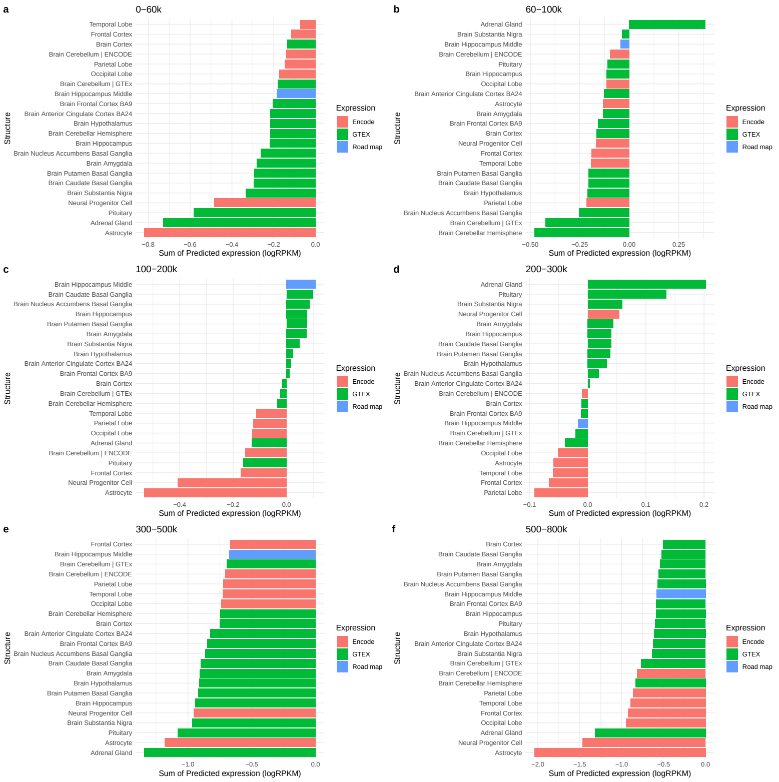


Figure 15: Cumulative predicted expression values by brain-related structure and time window (in years). Color legend indicates data source of prediction models.

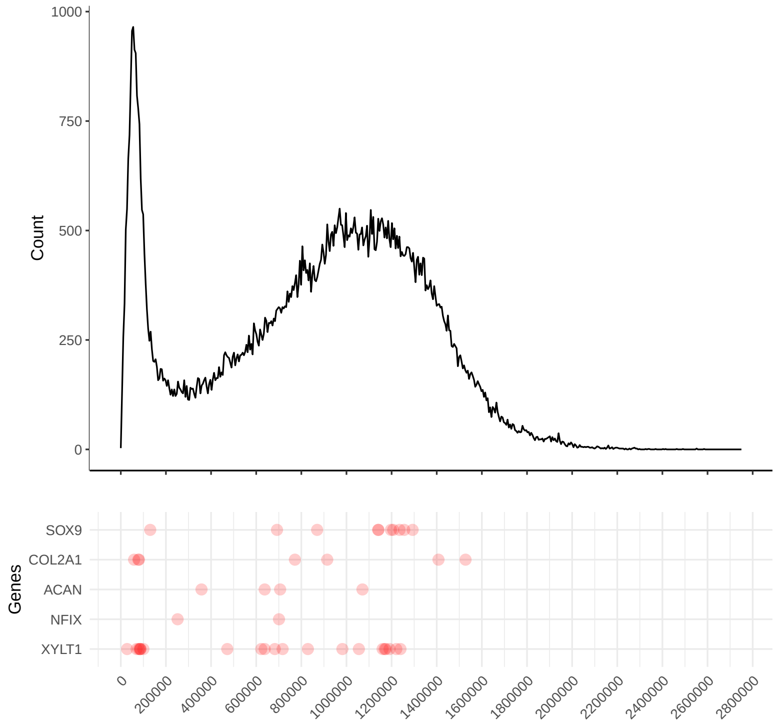


Figure 16: Temporal distribution of variants in genes exhibiting differential methylation profiles in **?** linked to modern human facial traits.

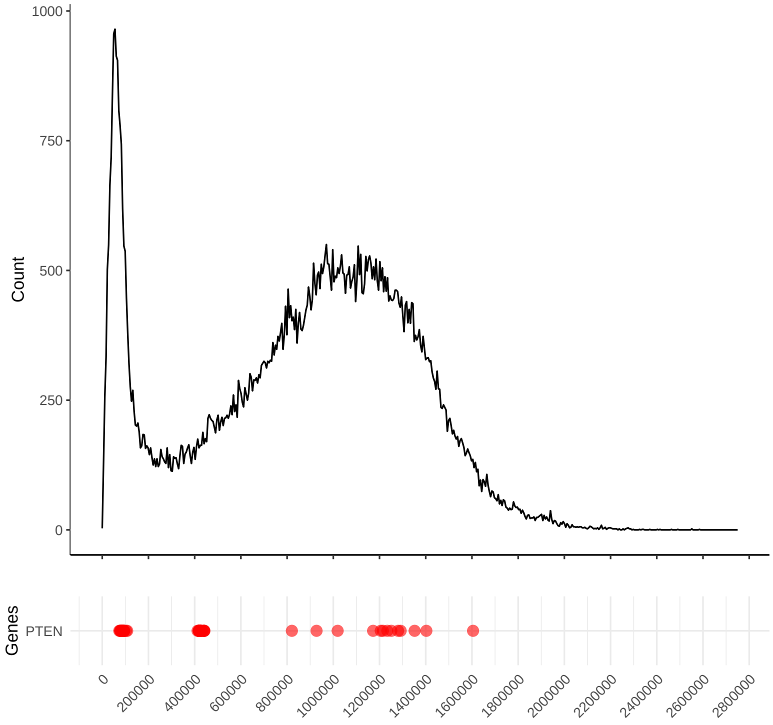


Figure 17: Temporal distribution of variants in $PTEN$, a gene highlighted in **?**. The gene displays HF variants clustering around the two periods highlighted in the main text (around 100kya and around 400kya.

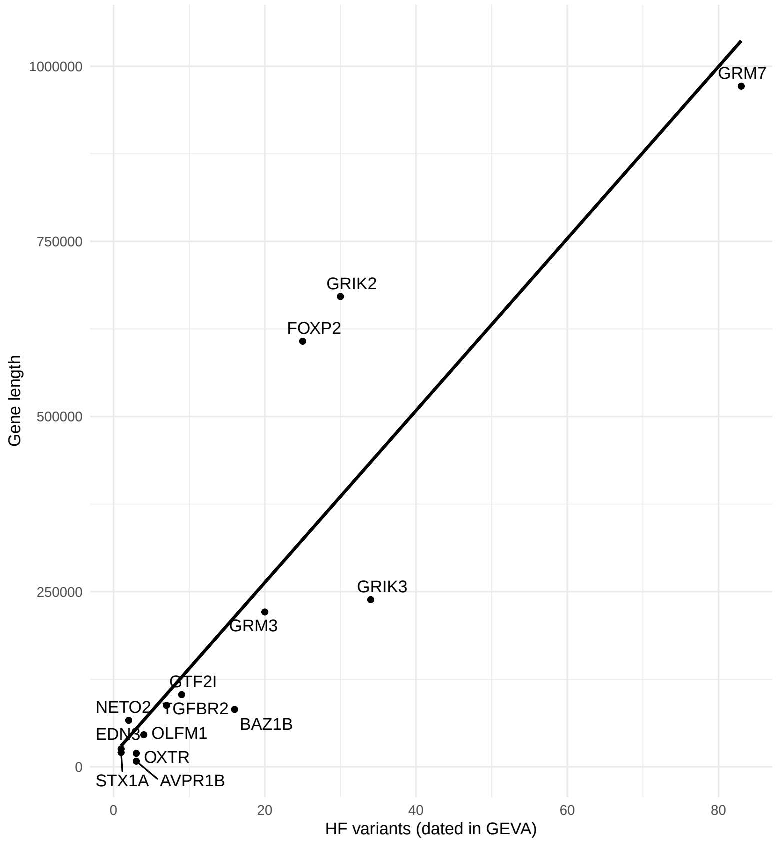


Figure 18: Relationship between gene length and HF variants mapped in GEVA, along with line of best fit.

5 | List of publications

This is a list of publications and preprints produced during my PhD. Manuscripts 3, 4 and 6 are reflected in chapters 2, 3 and 4. Manuscripts 1, 2 and 4 are not included in this thesis but are referenced and briefly explained in the introduction when relevant.

1. *Human-derived alleles in SOST and RUNX2 3'UTRs cause differential regulation in a bone cell-line model.* Juan Moriano*, Núria Martínez-Gil*, **Alejandro Andirkó**,Susana Balcells, Daniel Grinberg, Cedric Boeckx.
   BioRxiv. https://doi.org/10.1101/2021.04.21.440797 (2021)

2. *A distinct expression profile in the cerebellum and striatum for genes under selection within introgression deserts.* Raul Buisan*, Juan Moriano*, **Alejandro Andirkó**, Cedric Boeckx.
   BioRxiv. https://doi.org/10.1101/2021.03.26.437167 (2021)

3. *Fine-grained temporal mapping of derived high-frequency variants supports the mosaic nature of the evolution of Homo sapiens.* **Alejandro Andirkó**, Juan Moriano, Alessandro Vitriolo, Martin Kuhlwilm, Giuseppe Testa, Cedric Boeckx.
   BioRxiv. https://doi.org/10.1101/2021.01.22.427608 (2021)

4. *An Agent-based model of the gradual emergence of modern linguistic complexity.* Marcel Ruland, **Alejandro Andirkó**, Iza Romanowska, Cedric Boeckx.
   BioRxiv. https://doi.org/10.1101/2020.11.12.380683 (2020)

5. *Dosage analysis of the 7q11.23 Williams region identifies BAZ1B as a major human gene patterning the modern human face and underlying self-domestication.* Matteo Zanella, Alessandro Vitriolo, **Alejandro Andirkó**, Pedro T 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, Giuseppe Testa.
   Science Advances. `https://doi.org/10.1126/sciadv.aaw7908` (2019)

6. *Modern human alleles differentially regulate gene expression across brain regions: implications for brain evolution.* **Alejandro Andirkó**, Cedric Boeckx
   BioRxiv. `https://doi.org/10.1101/771816` (Last updated: 2020. Note that the current state of the preprint does not reflect the contents of Chapter 2).