MASTERARBEIT | MASTER'S THESIS

Titel | Title

Lineage-specific regulatory elements in neural crest as drivers of primate craniofacial evolution

verfasst von | submitted by

Nadezhda Romanova

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Genomics and Systems Biology
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Mihaela Pavlicev

Table of Contents

Acknowledgements

I am deeply grateful to the whole Adameyko lab (Medical University of Vienna), which played a crucial role in my development as a person eventually deserving an MSc title. I would especially like to thank my best supervisor and truly excellent researcher and bioinformatician, Nikita Vaulin, who first met me as an intern eager to learn a few more Python packages and then accompanied the entire journey to becoming a junior bioinformatician confident in coding, processing various sequencing data, and making cool plots out of it.

I would also like to thank Mona Christensen for the deep theoretical questions that always bring us back from data analysis to the biological meaning – the true value and purpose of our research. I am equally grateful for her support without hesitation, especially at moments when she somehow managed to sense my deepest concerns.

I would not have made it through this journey without Ines Rosa – another master's student who shared our "masters-only secrets" and with whom I walked the entire road from application to PhD positions shoulder to shoulder. I will definitely miss our "small subdepartment" sessions with Nikita, Mona, and Ines – and even the paper clubs when absolutely no one had the capacity to discuss the papers.

During my time in the lab, some people opened up to me a bit later, but became no less important. I would like to thank Rozalina for the multiome data from her research and, perhaps even more importantly, for the personal support that came along with it. I would also like to mention Karina Arsalanova, my music-from-the-2010s soulmate, for all our small but very supportive conversations – and for the best unintended role in one very important story.

Finally, I am deeply grateful to Igor Adameyko, who showed me a truly unique approach to being a PI. The atmosphere he created in the lab is something I have never seen before – but sincerely hope to see again (or perhaps create myself one day).

I believe I had the best university professors I could have wished for – people who not only know their subject, but are eager to reveal its beauty to their students. They were supportive, patient, and open-minded, setting a very high reference point for my future self. Still, I would like to give special thanks to a few of them.

First, my university supervisor, Mihaela Pavličev, who saw both my ups and downs during the past year but continued to believe in me and supported me during the hardest moment, when everything seemed to fall apart. Another very special person for me is Oleg Simakov, who guided me from my very first humble steps in bioinformatics. I deeply appreciate how consistently he supported all my ideas and ambitions, whatever they were. I am

also grateful to him and Igor Adameyko for the countless reference letters they wrote, which played a very important role in my academic journey.

Finally, I cannot fail to mention my amazing groupmates. With them I never felt isolated, and together we somehow managed to climb over our cultural barriers just to have fun on the other side. I also believe my knowledge would not be nearly as deep without all our conversations during seminars, lectures, and in the Mensa.

Thank you all.

Abstract

The neural crest is a vertebrate innovation that gives rise to a wide range of cell types and tissues. It plays a central role in craniofacial morphogenesis, pigmentation, and much of the peripheral nervous system – systems that together underlie the adaptability that allowed vertebrates to become a dominant animal clade. The remarkable diversity of neural crest derivatives raises fundamental questions about the regulatory logic that governs these developmental programs. Despite substantial progress in dissecting neural crest gene regulatory networks, important gaps remain: we still have a limited understanding of how variation within neural crest cis-regulatory elements translates into phenotypic diversity, and how these regulatory programs have evolved over vertebrate and mammalian evolution.

Recent advances in comparative genomics indicate that the regulatory landscape is far from static. Large-scale cross-species analyses have revealed extensive turnover of cis-regulatory elements across mammals, with primates in particular exhibiting a substantial fraction of lineage-specific regulatory regions. These observations suggest that changes in enhancer sequences and transcription factor binding sites may have played an important role in shaping primate-specific traits and complex phenotypes.

In parallel, genome-wide association studies increasingly point to a contribution of regulatory variation to neural crest–derived traits. Numerous GWAS loci associated with craniofacial morphology, map to non-coding regions active in neural crest cells, highlighting a functional connection between cis-regulatory variation and neural crest–dependent phenotypic outcomes. However, for most of these associations, the underlying regulatory mechanisms and target genes remain unresolved.

In this work, I aim to integrate single-cell multiome data with cross-species comparative analyses to characterize primate- and human-specific enhancer regions and the regulatory mechanisms they encode. By combining motif-level evolutionary analyses with enhancer–gene links and GWAS data, this study seeks to provide insight into how changes in neural crest regulatory architecture contribute to phenotypic diversity across mammals.

Zusammenfassung

Die Neuralleiste ist eine vertebratenspezifische Innovation, aus der eine Vielzahl von Zelltypen und Geweben hervorgeht. Sie spielt eine zentrale Rolle bei der kraniofazialen Morphogenese, der Pigmentierung und einem Großteil des peripheren Nervensystems – Systeme, die gemeinsam die Anpassungsfähigkeit begründen, die es den Wirbeltieren ermöglicht hat, zu einer dominanten Klade zu werden. Die bemerkenswerte Vielfalt der Neuralleistenderivate wirft grundlegende Fragen über die regulatorische Logik auf, die diesen Entwicklungsprogrammen zugrunde liegt. Trotz erheblicher Fortschritte bei der Analyse genregulatorischer Netzwerke der Neuralleiste bestehen weiterhin wichtige Forschungslücken: Unser Verständnis davon, wie Variationen in cis-regulatorischen Elementen der Neuralleiste in phänotypische Vielfalt übersetzt werden und wie sich diese regulatorischen Programme im Laufe der Wirbeltier- und Säugetierevolution entwickelt haben, ist nach wie vor begrenzt.

Jüngste Fortschritte in der vergleichenden Genomik zeigen, dass die regulatorische Landschaft keineswegs statisch ist. Groß angelegte artenübergreifende Analysen haben eine umfangreiche Fluktuation cis-regulatorischer Elemente bei Säugetieren aufgedeckt, wobei Primaten insbesondere einen erheblichen Anteil an lineage-spezifischen regulatorischen Regionen aufweisen. Diese Beobachtungen legen nahe, dass Veränderungen in Enhancer-Sequenzen und Transkriptionsfaktorbindungsstellen eine wichtige Rolle bei der Ausprägung primatentypischer Merkmale und komplexer Phänotypen gespielt haben könnten.

Parallel dazu weisen genomweite Assoziationsstudien zunehmend auf einen zentralen Beitrag regulatorischer Variation bei Merkmalen hin, die von der Neuralleiste abstammen. Zahlreiche GWAS-Loci, die mit kraniofazialer Morphologie assoziiert sind, liegen in nicht-kodierenden Regionen, die in Neuralleistenzellen aktiv sind, und unterstreichen damit einen funktionellen Zusammenhang zwischen cis-regulatorischer Variation und neuralleistenabhängigen phänotypischen Ausprägungen. Für die meisten dieser Assoziationen bleiben die zugrunde liegenden regulatorischen Mechanismen und Zielgene jedoch ungeklärt.

In dieser Arbeit verfolge ich das Ziel, Einzelzell-Multiomik-Daten mit artenübergreifenden vergleichenden Analysen zu integrieren, um primatentypische Enhancer-Regionen und die regulatorischen Mechanismen, die sie kodieren, zu charakterisieren. Durch die Kombination evolutionärer Analysen auf Ebene regulatorischer Motive mit Enhancer-Gen-Verknüpfungen und GWAS-Daten soll diese Studie Einblicke liefern, wie Veränderungen in der regulatorischen Architektur der Neuralleiste zur phänotypischen Vielfalt bei Säugetieren beitragen.

Introduction

The neural crest is a vertebrate innovation that gives rise to a wide range of derivatives. It is essential for craniofacial morphogenesis, pigmentation, and much of the peripheral nervous system – systems that together underpin the adaptability that helped vertebrates become one of the dominant animal clades. Interestingly, the neural crest cells show high plasticity, leading to developmental convergence with many other cell types. For example, cranial sensory neurons arise from both neural crest progenitors and ectodermal placodes. A similar dual origin applies to craniofacial cartilage and bone-forming cells: many craniofacial skeletal elements arise from proliferating neural crest cells, whereas other skull components are mesoderm-derived (Fig. 1A-B). As a consequence of this unique developmental pathway in vertebrates, contributing to ectodermal and mesodermal compartments, the neural crest is sometimes referred to as ectomesenchyme, comprising a "fourth germ layer" (Hall 2000).

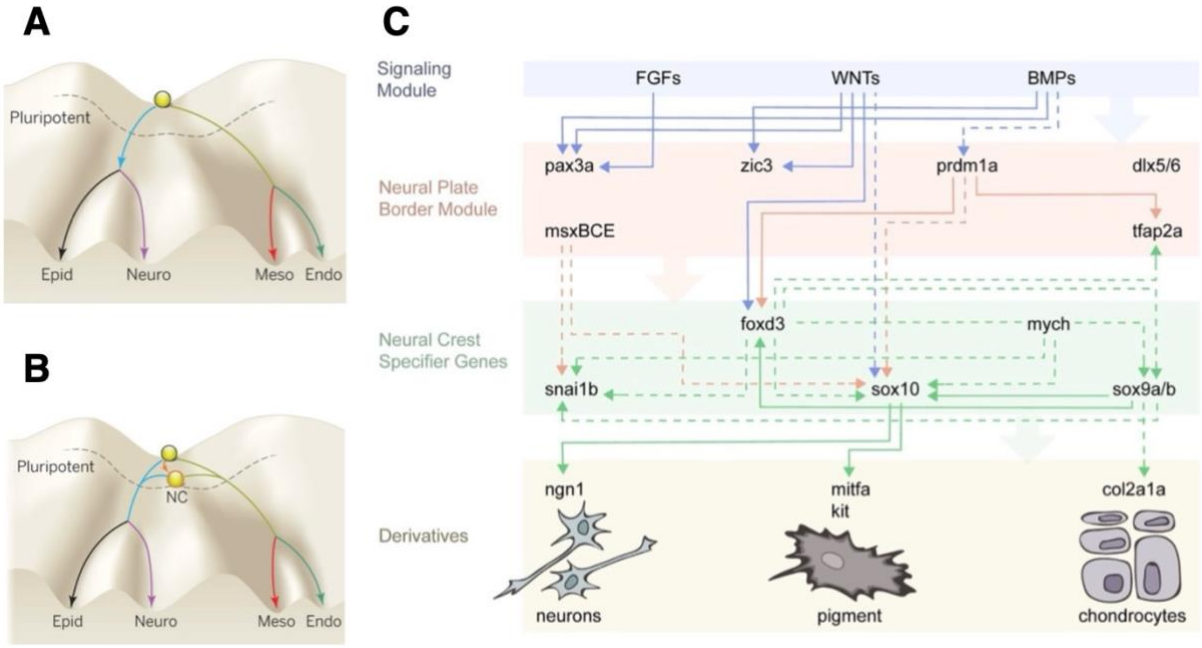


Figure 1. (**A)** Classical developmental landscape of pluripotent cells (Waddington, 1957). (**B)** Developmental landscape of early neural crest cells. Panels A–B are modified from https://doi.org/10.1126/science.aab2719. **(C)** Simplified early neural crest gene regulatory network, adapted from https://doi.org/10.1002/dvdy.122.

The peculiarity of neural crest cell fate makes it an important model for studying cell migration, differentiation, and integration in complex systems. Intuitively, this is particularly interesting because migrating cells pass through numerous chemical gradients, which act as

switches turning regulatory circuits on and off, ultimately leading to the execution of specific genetic programs. Over the past decades, this has led to a substantial body of work describing the mechanisms underlying neural crest development, including key transcription factors and signaling pathways controlling lineage specification (N. Hu et al. 2014; M. E. Bronner and LeDouarin 2012; Simões-Costa et al. 2012; Prescott et al. 2015).

However, several important gaps remain. First, our ability to functionally connect potentially significant neural crest regulatory elements to the genes they control — especially at single-cell resolution — remains limited, and we still lack an understanding of downstream regulatory effects. Second, it remains unclear how the remarkable diversity of neural crest regulatory programs emerged and how they evolved over vertebrate evolution. Third, we are still far from understanding how genetic variation in regulatory regions contributes to phenotypic outcomes, particularly in complex traits, for example, craniofacial morphology, which is largely derived from neural crest progenitors.

Taking advantage of recent advances in single-cell omics and the growing amount of available data, this thesis aims to take a further step towards understanding the emergence of regulatory elements within specific lineages (in this case, primates), with a particular focus on enhancers as the main class of regulatory elements, and how this relates to phenotypic features observed in these groups, specifically craniofacial morphology.

To address these questions, we first need to deepen our understanding of three key components of the study: neural crest regulation, enhancer biology and evolution, and craniofacial diversity.

Neural crest regulation

In embryonic development, neural crest formation can be divided into several distinct steps governed by specific sets of regulatory factors (Fig. 1C). Initiation is driven by the cumulative effects of Wnt/BMP and FGF signaling from neighboring tissues during early development (around 8–10.5 days in mouse and 4-12.5 weeks in human). A specific set of transcription factors, including Msx, Zic1, and Pax3/7, then triggers the formation of the neural plate border, which later gives rise not only to the neural crest but also to other ectodermal structures such as ectodermal placodes and Rohon–Beard cells. Further refinement of the neural plate border is achieved through coordinated input from extracellular signaling pathways and the neural plate border specifier module, which together activate a set of transcription factors collectively referred to as “neural crest specifiers”. This module includes factors such

as Snail/Slug, FoxD3, Tfap2a, Twist1, Id, Myc, and Sox9/10, all of which are expressed in pre-migratory and migratory neural crest cells. Their activation depends on continued Wnt and other signaling inputs in combination with neural plate border markers (Zic, Msx1/2, and Pax3/7), which are subsequently downregulated as cells begin to migrate. At the same time, neural crest specifiers repress the neural tube marker Sox2, thereby reinforcing neural crest identity (M. E. Bronner and LeDouarin 2012). Functionally, this transcriptional program change enables epithelial-to-mesenchymal transition (EMT), allowing neural crest cells to delaminate from the neural tube, migrate, and later differentiate into a wide range of derivatives. Later, some "neural crest specifiers" also guide differentiation into particular lineages: for example, Sox10 controls melanocyte, sensory, autonomic, and glial lineages; Twist1 is responsible for formation of early mesenchyme and Sox9 is central to the cartilage lineage.

At later stages of neural crest development, the population of migrating, stem-like neural crest cells are classically divided into several major regional subpopulations: craniofacial, trunk, vagal and sacral neural crest. These subpopulations develop in distinct microenvironmental contexts and are exposed to different combinations of transcription factors, which contribute to the remarkable diversity of neural crest derivatives (Soldatov et al. 2019). Importantly, different cell programs are initiated by distinct cis-regulatory inputs: for instance, enhancers that drive early neural crest gene expression can show strong cranial specificity, whereas other enhancers preferentially initiate expression in trunk neural crest and become active later in migrating cranial cells (Simões-Costa et al. 2012).

The craniofacial complex is one of the most diverse and rapidly evolving parts of the vertebrate body. It includes many cell types from all embryonic layers, and a broad range of neural crest derivatives is naturally integrated within this complex. Skeletogenic patterns are initiated by epithelial signal centers like Frontonasal Ectodermal Zone (FEZ) and forebrain (Kaucka et al. 2018; D. Hu and Marcucio 2009). Then, during osteo- and chondrogenesis, migratory neural crest cells can be subdivided into Hox-negative versus Hox-positive populations. The former is responsible for the normal formation of skeletal facial elements, while the latter are associated with pharyngeal arches and their major derivatives, including jaws. These programs operate in combination with transcription factors such as Sox9, Msx1, and Msx2 (Simões-Costa et al. 2012). Importantly, other neighboring tissues such as muscles, dermal cells, and adipose tissue, are engaged in mechanical and paracrine interactions with neural crest derivatives (Qian et al. 2022; Sunadome et al. 2023), which altogether get assembled in species-specific and individual unique facial shape.

Enhancers as players of evolution

A substantial fraction of the genome is involved in gene regulation, with a central role assigned to cis-regulatory elements (CREs) — DNA regions that, through interactions with transcription factors and other regulatory proteins, orchestrate context-specific gene expression in a tissue-, cell type–, and time-dependent manner. Variation in regulatory regions has been proposed as an important source of evolutionary change for decades (King and Wilson 1975), and more recent genome-wide association studies (GWAS) support this idea by showing that many trait-associated variants are located within annotated regulatory regions (Mostafavi et al. 2023; Maurano et al. 2012).

Among different types of CREs, enhancers — regulatory regions that modulate gene expression through transcription factor recruitment and interactions with promoters — stand out in the context of evolution. On the one hand, enhancers are flexible in terms of sequence variation and can accumulate mutations that subtly alter their affinity for transcription factors. This flexibility allows enhancers to produce modest quantitative changes in gene expression, enabling phenotypic shifts without causing catastrophic defects (Prabhakar et al. 2008; Chan et al. 2010; Long et al. 2016). On the other hand, enhancers and regulatory networks in general exhibit robustness that stabilizes developmental programs while preserving latent potential for adaptive change (Ogbunugafor et al. 2009; Hayden et al. 2011; Payne and Wagner 2015). Altogether, this makes enhancers a particularly interesting focus for evolutionary studies.

The functional units of enhancers are transcription factor binding sites – short DNA sequences (motifs) bound by transcription factors (TFs). The composition and arrangement of motifs within an enhancer can shape transcriptional output and, in some contexts, influence splicing patterns (Kadener et al. 2002; Kleinjan et al. 2008; Dahan et al. 2021). From an evolutionary perspective, this implies that sequence changes affecting motif content, affinity, or presence/absence can be major drivers of evolution, because they directly impact the regulation of protein-coding genes and, ultimately, the phenotype under selection.

The functional consequences of motif-level variation have been repeatedly demonstrated in invertebrates and vertebrates (Frankel et al. 2010; Crocker et al. 2015; Lim et al. 2024). A classic example is the cis-regulation of the Hox gene cluster in Drosophila, where low-affinity motif clusters were shown to play a key role in anterior–posterior patterning (Crocker et al. 2015). *In vivo* occupancy did not simply track motif strength: sequences containing higher-affinity motifs were not necessarily the most bound. This supports an

important principle of enhancer architecture – many weak sites can collectively produce strong and context-specific regulation. Such architectures provide an important mutational domain where small changes that create, weaken, or rearrange individual low-affinity motifs can fine-tune expression patterns without disrupting the entire regulatory program, thereby enabling gradual evolutionary shifts in development, physiology, or morphology.

In vertebrates, a comparable example comes from mouse reporter assays of the major Shh limb enhancer, the Zone of Polarizing Activity Regulatory Sequence (ZRS) (Lim et al. 2024). By disrupting ETS transcription factor binding sites at different positions within the enhancer, the authors demonstrated a direct link between motif-level changes, altered Shh expression, and a measurable phenotypic outcome in limb development, namely changes in digit formation. Together, these studies illustrate how relatively small sequence edits in enhancer motifs can translate into quantitative changes in gene regulation with clear developmental consequences.

Overall, motif-level variation within active enhancers can produce specific and coordinated – yet not necessarily pathological – phenotypic shifts. Because multiple traits may share elements of a common developmental origin or gene regulatory program, variation in one or several enhancers can reshape transcription factor binding and reorganize regulatory network activity, potentially contributing to concerted changes across developmentally related traits.

Large-scale comparative genomics now makes it possible to address enhancer-evolution questions systematically rather than case-by-case. Using the 241-way mammalian alignment generated by the Zoonomia project, a recent study mapped human candidate cis-regulatory elements (cCREs) across mammals. This approach revealed that the regulatory landscape contains both a deeply conserved core and a substantial “dynamic” component that changes over evolutionary time (Andrews et al. 2023). Crucially, the authors also identified CREs that are conserved within the primate lineage but not across mammals more broadly, pointing to potential lineage-specific remodeling of regulatory DNA. Subsequent analyses revealed that genes near primate-specific elements are enriched for functions related to environmental interaction, including odor perception and immune response, suggesting lineage-specific regulatory innovation rather than uniform drift across mammals. However, annotation of regulatory regions by proximal genes does not always reflect true regulatory interactions; therefore, this observation requires more precise validation (Fulco et al. 2019). Finally, an additional notable point of the study is that a high fraction of primate-specific

transcription factor binding sites (~90%) overlaps with classes of the evolutionary youngest transposable elements (LIME1, Alu, LTR families), suggesting a possible evolutionary scenario for regulatory network rearrangements. Overall, such primate-specific regulatory elements provide a starting point for dissecting how changes in enhancer sequence – and, at a finer scale, transcription factor binding site composition – can reshape gene regulation and contribute to complex, developmentally patterned phenotypes.

Craniofacial diversity

As was mentioned above, one of the most prominent examples of complex phenotypes largely shaped by neural crest derivatives is the craniofacial complex, specifically, facial shape. Owing to the high prevalence of abnormalities associated with facial development – such as orofacial clefts or craniosynostosis syndromes – the genetic basis of craniofacial morphology has been extensively investigated over the past decade. Several genome-wide association studies (GWAS), primarily based on European cohorts, have identified numerous loci associated with both craniofacial disorders and normal variation in facial shape (Shaffer et al. 2016a; Claes et al. 2018; Goovaerts et al. 2026). Integration of these genetic data with large-scale, high-resolution phenotypic datasets of human facial morphology has enabled the association of specific SNPs with distinct facial features. However, a major limitation of these studies is that they provide limited insight into the molecular mechanisms underlying the emergence of these phenotypic patterns. At the same time, increasing attention has been directed toward regulatory elements involved in craniofacial development, as they provide a plausible basis for both species-specific and individual variation in facial morphology.

An early attempt to bridge regulatory logic with craniofacial phenotypic outcomes was made by Prescott et al. (2015). By comparing active chromatin regions between humans and chimpanzees, the authors identified human-specific regulatory elements associated with neural crest cells. This was achieved using ChIP-seq data generated with enhancer-associated histone marks in combination with antibodies targeting neural crest–specific transcription factors. Approximately 6% of regulatory regions showed human-specific activity relative to chimpanzee, and motif analysis revealed several transcription factor binding sites with either higher frequency or stronger regulatory effects in humans. Notably, the authors identified a motif region termed a “coordinator,” characterized by both high motif frequency and strong regulatory impact. By examining nearby genes, they proposed candidate molecular pathways affected by these human-specific regulatory elements, many of which are implicated in facial

development (Prescott et al. 2015). Nevertheless, this study was limited by the use of neural crest–like cell cultures, which do not fully capture the complexity of the *in vivo* regulatory landscape. In addition, it has been already mentioned above that annotation of regulatory elements by proximity to genes has certain limitations (Fulco et al. 2019).

More recently, a large-scale study integrating over 50,000 high-resolution MRI scans with publicly available GWAS data from the UK Biobank identified 2,579 conditionally independent SNPs associated with craniofacial shape (Goovaerts et al. 2026). Importantly, a high density of SNPs was observed near key craniofacial transcription factors genes, including *TBX15, OSR1, MEIS1, RUNX2*, and *SOX9*, extending the findings of Prescott et al. by suggesting that craniofacial variation may involve not only changes in accessibility of neural crest regulatory elements, but also variation affecting the expression and regulation of core developmental transcription factors themselves. Using H3K27ac ChIP-seq data from multiple relevant cell types – including cranial neural crest cells, embryonic craniofacial tissues (4-8 weeks of embryonic development), osteoblasts, and cranial neural crest–derived chondrocytes – the authors clustered SNPs according to lineage-specific regulatory activity, thereby assigning putative effects to distinct developmental contexts and highlighting once more a strong effect on NC-related transcription factors. Furthermore, sequence-to-function deep learning models demonstrated that variation at significant loci influences TWIST1 dosage sensitivity and SOX9 responsiveness.

Taken together, current study attempts a new step toward understanding the regulatory architecture underlying craniofacial phenotypes. However, as with earlier work by Prescott et al., it does not yet fully resolve the regulatory logic linking non-coding variation to gene expression at the cell-type level and doesn't account for variation in target regulatory regions. In particular, deeper integration with high-resolution, single-cell regulatory and transcriptional data would further refine enhancer–gene assignments and provide a more direct mechanistic interpretation of neural crest–driven craniofacial variation. To conclude, these findings highlight the substantial impact of cis-regulatory variation on neural crest development and underscore the importance of further studies aimed at dissecting regulatory mechanisms underlying craniofacial phenotypic diversity.

Hypothesis and aims

The central hypothesis of this thesis is that lineage-specific craniofacial evolution is partially driven by the emergence of new enhancers associated with genes involved in face development, which alter gene expression patterns by introducing novel regulatory contexts.

Aim 1: Assess sequence conservation of craniofacial neural crest enhancers.

Using available human multiome data, I will extract enhancer regions active in craniofacial neural crest cells. These enhancers will be projected to other mammalian genomes using a combination of liftOver-like mappings and publicly available whole-genome multiple sequence alignments (Zoonomia project). Enhancers will then be classified into conservation groups (e.g., broadly conserved across mammals, primate-specific, or human-specific, based on alignability).

Aim 2: Interpret potential effects of lineage-specific enhancers by GWAS integration and gene expression analysis.

By integrating GWAS data for craniofacial traits with multiome-derived single-cell RNA-seq and ATAC-seq data, I will prioritize variants located in regulatory regions and link them to candidate target genes and cell states. Using single-cell expression profiles, I will assess whether transcriptional effects are localized to specific cell populations or distributed across multiple neural crest–derived states.

In parallel, using enhancer–gene links derived from the multiome data, I will assess whether different enhancer conservation classes are associated with distinct functional and expression patterns. First, gene set enrichment analysis will be used to identify biological pathways and neural crest–related processes linked to lineage-specific regulatory circuits. Second, single-cell expression profiles will be used to examine the expression of candidate genes across developmental stages and cell populations, allowing us to determine whether their activity is restricted to specific cell states or broadly distributed across neural crest–derived and surrounding tissues.

Aim 3: Analyze a motif composition of lineage-specific enhancers.

Using available transcription factor motif databases and motif analysis tools, I will investigate the regulatory features introduced by lineage-specific enhancers in primate and human evolution.

By applying Term Frequency-Inverse Document Frequency (TF–IDF) scoring and motif enrichment analysis, I will identify transcription factor binding motifs that are overrepresented or specific to lineage-restricted enhancers compared to the rest of the dataset. This approach will allow us to infer which regulatory network components may have been preferentially engaged or modified during the recent evolution of the lineage.

As a result, this study aims to (i) identify enhancer and motif features associated with primate and human regulatory network evolution of neural crest derivatives (Aims 1 and 3), (ii) establish initial links between trait-associated non-coding variants and their target genes (Aim 2), and (iii) identify candidate enhancer regions and transcription factors for future functional assays in these tissues (Aims 2–3). These candidate regions can focus the subsequent experimental work, which is the ultimate approach for discovering the actual connections between regulatory region variation, gene expression, and phenotypic outcome, to a feasible subset of likely targets.

Results

Identification and conservation of craniofacial neural crest enhancers

Recent studies suggest that regulatory elements continuously emerge and transform over evolutionary timescales, therefore some of them are conserved within a lineage are likely functionally relevant (Rebeiz and Tsiantis 2017; Klein et al. 2018; Andrews et al. 2023). Hence, comparative analysis of lineage-restricted conserved enhancers, combined with gene expression and phenotype-association data, may provide an opportunity to better understand the role of regulatory evolution in shaping species-specific traits. Following this reasoning, we leveraged single-cell genomic data together with publicly available multiple genome alignment.

As a starting point, we used a multiome dataset recently generated in the lab (Erickson et al., under revision). The dataset combines scRNA-seq and snRNA-seq/scATAC-seq data collected from facial dissections performed between 6 and 12.5 weeks of human embryonic development. Taking advantage of the multiomic approach, the authors identified 11,748 putative enhancer regions through peak analysis of open chromatin regions and their correlation with gene expression. The coordinates of these enhancers (according to hg38 annotation) were used as an input for the current study. For cross-species comparison, we used the 241-species alignment generated by the original Zoonomia project (Genereux et al. 2020).

We categorize enhancers based on the taxonomic range at which their sequence is preserved across the set of 22 mammalian species studied, including 8 primates (Table 4, Materials and Methods). Preservation was defined as the proportion of an enhancer sequence present in a target species relative to the human reference (hg38), based on the presence or absence of insertions and deletions (indels). This metric was prioritized over sequence similarity, as cis-regulatory regions are known to tolerate substantial nucleotide-level variation: transcription factor binding sites are short, often degenerate, and can accumulate substitutions without complete loss of function. In contrast, the gain or loss of larger sequence segments containing binding sites is more likely to disrupt regulatory activity (Osterwalder et al. 2018; Long et al. 2016; Britten R.J. 1996). Thus, sequence preservation captures the evolutionary persistence of enhancer elements while remaining robust to motif-level substitutions and high noise in inter-motif regions.

Based on resulting preservation values, enhancers were assigned to hierarchical conservation categories reflecting their phylogenetic distribution (Fig. 2):

1) Conserved – at least 18 out of 22 studied species show >90% sequence preservation;
2) Primate-specific – all primates (n = 8) show >90% preservation, while at least half of the remaining species show <10% preservation;
3) Human-specific– all species have <10% preservation;
4) Other – enhancers not meeting the above criteria.

Using this approach, we identified 2,929 conserved mammalian enhancers, 84 primate-specific enhancers, and 5 unique to humans (human-specific). The remaining 8,730 (~74%) regulatory regions displayed heterogeneous preservation patterns across species and were classified as “Other” (Supplementary Material 1).

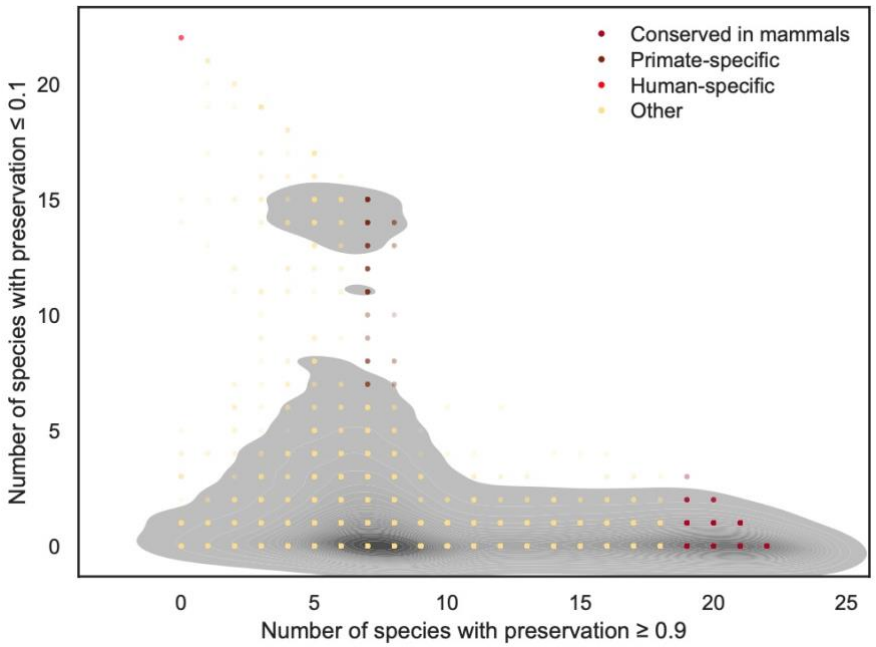


Figure 2. Scatter–density plot of categorized enhancers. A dense region around ~8 on the x-axis and ~0 on the y-axis reflects the predominance of "other" enhancers where primates still have high level of preservation compared to all other species.

Overall, these results for neural crest enhancers are roughly consistent with previously reported patterns of general cis-regulatory evolution in mammals, although quantitative differences are observed. According to Zoonomia analyses, approximately 40% of candidate regulatory elements are broadly conserved, and ~10% are primate-specific (Andrews et al. 2023); in our dataset, these proportions are 25% and 0.7%, respectively. This discrepancy may reflect either increased evolutionary constraint within the neural crest regulatory network or differences in species representation. It is also important to note that the Zoonomia dataset includes limited neural crest–specific regulatory annotations, which affects direct comparability.

GWAS variants overlap human-specific regulatory elements

Next, we integrated the categorized enhancers with 2579 known single-nucleotide polymorphisms (SNPs) associated with human craniofacial morphology from *Goovaerts et al.* (2026) as additional information linking regulatory elements to reported phenotypic variation.

Overall, 93 out of 2,579 variants fell within putative neural crest–related enhancers. We then assessed whether known SNPs overlap inferred lineage-specific enhancer regions (i.e., primate- or human-specific). We found 2 variants – rs547998 and rs479247 – to fall into

human-specific enhancers. These enhancers were associated with the *SGCZ* and *DLC1* genes according to the link-peak analysis. These SNPs are in close genomic proximity: both reside on chromosome 8 and are approximately 1.5 kb apart (Table 1). The corresponding enhancer regions are also located within less than 1 kb of each other; however, they were linked to different target genes, suggesting potentially distinct regulatory roles within a shared genomic neighborhood.

Table 1. SNPs metadata from Goovaerts et al. 2026 integrated with enhancer data

Enhancer	chr	pos	rs	p	log	maf	ref	alt	Primary anat region	Other anat region
SGCZ_en1	8	12901663	rs547998	2.0249e-12	11.6936	0.0767	T	A	Whole craniofacial region	Cranial vault
DLC1_en3	8	12903076	rs479247	1.0661e-11	10.9722	0.3976	T	A	Whole craniofacial region	Cranial vault

Overall, these observations may support the idea that craniofacial variation can be driven by newly emerged regulatory layers. However, several limitations should be considered and addressed in future studies. First, the identified relationships require experimental validation through controlled perturbation of enhancer elements, followed by assessment of gene and protein expression as well as downstream phenotypic effects. Second, it remains important to reconstruct plausible evolutionary scenarios underlying the emergence of these human-specific regions.

Lineage-specific enhancers are associated with collagen-related pathways

Using the multiome dataset (Erickson et al., under revision), and in particular enhancer–gene pairs defined by expression correlation, we next analyzed genes associated with different enhancer categories: primate-specific and human-specific. Among the 84 primate-specific regulatory regions, we filtered 96 associated protein-coding genes and 5 long non-coding RNAs (Supplementary Materials 4), indicating that some enhancers are linked to multiple expressed genes. For the 5 human-specific enhancers, we identified 5 associated genes.

Using the STRING platform (Szklarczyk et al. 2023), we performed gene set enrichment analysis on genes associated with primate-specific enhancers to assess which biological pathways may have been preferentially affected by lineage-specific regulatory changes (Supplementary materials 3).

First, several significantly enriched Reactome pathways (Ragueneau et al. 2026) were related to collagen biology (Fig. 3):

1) Collagen chain trimerization (strength = 1.03, signal = 0.48, FDR = 1.6×10^{-3});
2) Assembly of collagen fibrils and other multimeric structures (strength = 0.92, signal = 0.43, FDR= 0.0249);
3) Collagen formation (strength = 0.8, signal = 0.4, FDR = 0.0281);
4) and collagen degradation (strength = 0.89, signal = 0.36, FDR = 0.0447).

The pathways included several genes from the collagen family: *COL6A1, COL6A2, COL4A1, COL4A2, COL18A1, COL20A1*; and genes related to extracellular matrix organization, such as *PXDN* and *FBN2*. Given the central role of collagens in connective tissue structure and extracellular matrix organization, regulatory shifts in collagen-related genes could influence facial morphogenesis independently of their effects in other parts of the body, in contrast to changes in protein-coding sequences.

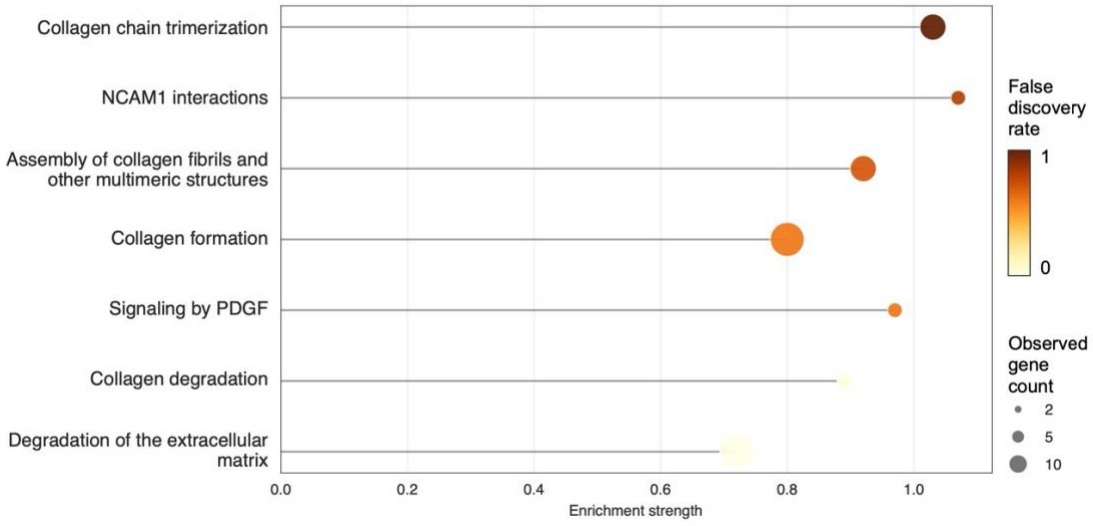


Figure 3. Dot plot of enriched Reactome pathways for the analyzed gene set. Enrichment is defined as the log10 ratio between (i) the number of proteins in the observed network annotated with a given term and (ii) the expected number of proteins annotated with that term in a random network of the same size.

Second, enrichment analysis identified STRING network clusters associated with chaperone activity (*HSPA6, HSPA1A, HSPA1B, HSPH1, DNAJB1, STIP1*). However, manual inspection of enhancer sequences linked to this pathway showed that many of these regulatory elements were close to the threshold used to define primate specificity (all except *HSPA6* and *HSPH1*). This suggests that the observed enrichment may be influenced by threshold-dependent effects rather than reflecting a robust biological signal and hence other independent evidence would be desirable before pursuing this signal further.

The remaining genes did not assemble into tightly interconnected network modules. Still, functional annotation revealed multiple genes involved in neuronal development and synaptic function (e.g., *CDH15, TENM3, AATK, PCBP3, TRIM2, CACNB2*), cytoskeletal organization and actin dynamics (*ARHGAP45, ARPC1B, EPS8, ENAH, MCF2L*), immune-related processes (*IFI16, MADCAM1, CYTH1, C17orf99*), and regulation of cell proliferation, migration, and embryonic development (*SFRP2, SOX5, SEMA6B, SKI, CYTL1, PARD6B*). The full list of genes and their functional annotations is provided in Supplementary Materials 4.

We then manually examined genes associated with human-specific enhancers (Table 2). These genes did not form a clearly defined functional unit; however, their annotated functions point to several processes potentially relevant to craniofacial regulatory programs.

Table 2. Genes associated with human-specific enhancers.

Gene	**Function**	**Human atlas description**
DLC1	GTP-binding proteins activation which participate in cytoskeletal changes	https://www.proteinatlas.org/ENSG00000164741-DLC1
KIF26B	Intercellular transport along microtubules	https://www.proteinatlas.org/ENSG00000162849-KIF26B
NLRP3	Regulation of inflammation, immune response, and apoptosis	https://www.proteinatlas.org/ENSG00000162711-NLRP3
PDZRN3	Assumed to function in differentiation of adipocytes, osteoblasts, and myoblasts, play a role in vascular morphogenesis	https://www.proteinatlas.org/ENSG00000121440-PDZRN3

SGCZ	Part of sarcoglycan complex located in membrane bridging the inner cytoskeleton and the extra-cellular matrix	https://www.proteinatlas.org/ENSG00000185053-SGCZ

In the previous section, we discussed *DLC1* and *SGCZ* as genes associated with human-specific regulatory variants. Notably, several additional genes linked to human-specific enhancers show related functional roles of its transcribed proteins. KIF26B is known for promoting cell proliferation and migration and has also been implicated in angiogenesis through the regulation of endothelial cell polarity (like DLC1). A similar role has been described for PDZRN3, a ubiquitin ligase that modulates endothelial cell extension via Wnt/PCP signaling.

Taken together, most genes associated with human-specific enhancers appear to be connected, directly or indirectly, to vascular development and endothelial cell behavior, the evolution of which requires further investigation in the context of neural crest development. The main exception is NLRP3, which is primarily involved in innate immune responses. However, it has been reported that primate genomes tend to accumulate regulatory elements associated with immune functions; therefore, the presence of *NLRP3* in this set may reflect a broader evolutionary trend.

Regulatory architecture of genes linked to lineage-specific enhancers

Genes are typically regulated by multiple enhancers that modulate their activity (Long et al. 2016; Osterwalder et al. 2018). This redundancy contributes to regulatory robustness and spatial or temporal diversification of gene expression (i.e., it differs across developmental stages and/or cell types). It raises the question of which genes tend to accumulate a large number of active enhancers (and what fraction of these are lineage-specific), and which are associated with only a few. Importantly, the number of enhancers does not directly reflect the functional importance of a gene, but instead highlights differences in regulatory architecture across the complex body plan.

Following these considerations, we aimed to characterize the regulatory complexity of genes in the dataset. First, we examined the number of enhancers associated with each gene. On average, each gene in the dataset was linked to approximately five active enhancers (Supplementary Materials 6.1). We then focused on the subset of genes associated with primate-specific enhancers.

First, we identified genes associated with a single primate-specific enhancer: *HSPA6, CERKL, NEUROD1, SLC12A7, TYMP, GABRB3, ARHGAP45, MADCAM1, ZNF460, KPNA4, PXDN*. Two of these genes are transcription factors (*NEUROD1, ZNF460*), which are typically regulated by multiple enhancers (Osterwalder et al. 2018). Notably, PXDN is involved in extracellular matrix organization, linking this group to collagen-related processes. The remaining genes have diverse functions without a clear shared biological theme (Table 3).

Table 3. Genes with a single primate-specific enhancer

Gene	Description	NCBI link
HSPA6	Heat shock chaperon protein responsible for protein folding and degradation. Usually acts only in a presence of stress factors.	https://www.ncbi.nlm.nih.gov/gene/3310
CERKL	A regulator of autophagy, acts as a cytoprotective factor, particularly in photoreceptor cells, protecting them from oxidative damage and apoptosis.	https://www.ncbi.nlm.nih.gov/gene/375298
NEUROD1	Transcription factor involved in neuronal maturation, sensory organ development.	https://www.ncbi.nlm.nih.gov/gene/4760
SLC12A7	Solute carrier expressed in inner ear and certain epithelial cells, facilitate movement of potassium and chloride ions	https://www.ncbi.nlm.nih.gov/gene/10723
TYMP	An enzyme thymidine phosphorylase, crucial for regulating nucleotide levels and maintaining mtDNA.	https://www.ncbi.nlm.nih.gov/gene/1890
GABRB3	Subunit of the GABA, neurotransmitter system in the brain.	https://www.ncbi.nlm.nih.gov/gene/2562
ARHGAP45	GTPase activator for the Rho-type GTPases (RhoGAP) domain that would be able to negatively regulate the actin cytoskeleton as well as cell spreading.	https://www.ncbi.nlm.nih.gov/gene/23526
MADCAM1	Cell adhesion molecule that binds to the integrin and L-selectin, directing lymphocyte migration to mucosal lymphoid tissues.	https://www.ncbi.nlm.nih.gov/gene/8174

ZNF460	Transcription factor from zinc finger family with no clear function. It is associated with cystathioninuria and several cancer types.	https://www.ncbi.nlm.nih.gov/gene/10794
KPNA4	A subunit of importin complex transporting NLS-containing proteins into nuclear pore complex.	https://www.ncbi.nlm.nih.gov/gene/3840
PXDN	Peroxidasin, enzyme involved in extracellular matrix formation, in particular collagen organization.	https://www.ncbi.nlm.nih.gov/gene/7837

Second, we prioritized genes associated with more than five enhancers in total (i.e., above the dataset average) and more than two primate-specific enhancers (assuming it allows us to focus on potential trend rather than coincidence). This approach aimed to identify genes that may preferentially accumulate new regulatory elements during recent evolution. Indeed, several genes exhibited this pattern (Fig. 4A). For example, AFDN, involved in cell junction organization and signaling during embryogenesis, showed that approximately 50% of its associated enhancers are lineage-specific (4 out of 8).

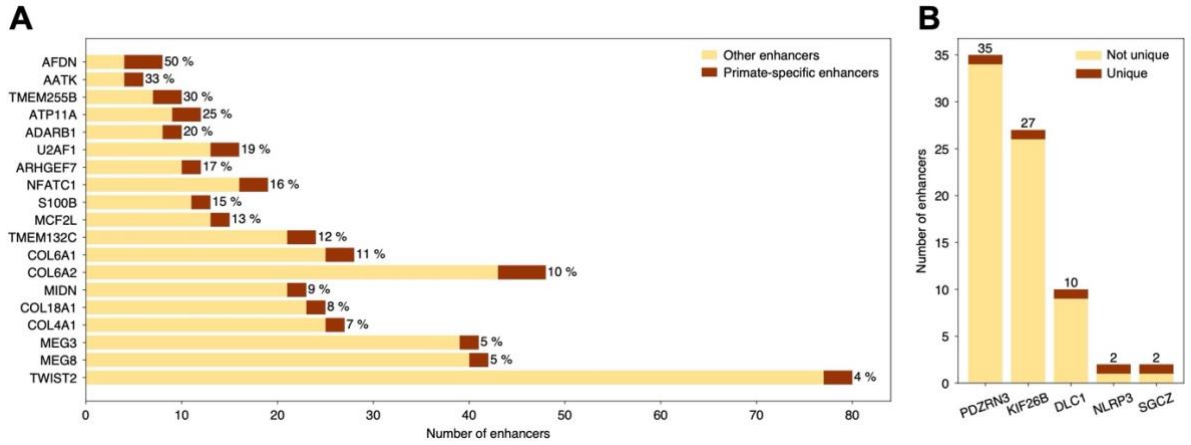


Figure 4. (**A**) Total number of enhancers per gene and the fraction of primate-specific enhancers. Precents demonstrates a fraction of primate specific enhancers. (**B**) Total number of enhancers per gene and the fraction of human-specific enhancers. Numbers above bars demonstrates a total number of enhancers.

Interestingly, many of the identified genes are linked to neural development: *AATK* promotes neuronal differentiation, *ADARB1* is involved in RNA editing and is highly expressed in neuronal tissues, *ARHGEF7* contributes to cerebellar development, and S100B regulates

glial cell formation. In addition, two genes, *NFATC1* and *TWIST2*, are key regulators of osteogenesis. Previously identified collagen-related genes also showed a relatively high number of lineage-specific enhancers, highlighting a potential expansion of expression contexts for this gene family during primate evolution. Full information on the enhancer number per gene is in Supplementary Materials 6.2.

All genes associated with human-specific enhancers had only a single specific enhancer (Fig. 4B). Nevertheless, previously discussed *SGCZ* has only two enhancers in total one of which is human-specific and contains a significant variant associated with craniofacial variation. Another gene – *DLC1* – has 10 enhancers in total, which indicates a possible robustness, but also suggests a diversity of regulatory programs across tissues. The highly immune-specific *NLRP3* gene has also only 2 enhancers, suggesting either a more universal regulatory pattern or highly localized expression. *PDZRN3* and *KIF26B* have a high number of enhancers, 35 and 27, respectively.

Expression patterns of genes linked to lineage-specific enhancers

After examining gene functions and regulatory architecture, we next asked whether genes associated with lineage-specific regulatory elements exhibit distinct expression patterns compared to other genes in single-cell RNA-seq data derived from multiome dataset (Erickson et al., 2026, under revision). The mean expression values discussed below were calculated from normalized and log-transformed raw counts (see Materials and Methods); however, in all plots, expression values are scaled from 0 to 1 separately for each gene.

First, we examined temporal-spatial expression pattern of genes associated with human-specific enhancers. Across developmental stages (Fig. 5A), expression of *SGCZ* and *KIF26B* showed a bias toward earlier embryonic stages, whereas the remaining genes exhibited either consistently low expression (e.g., *NLRP3*, likely reflecting high cell-type specificity) or not directionally distributed pattern (*DLC1*, *PDZRN3*).

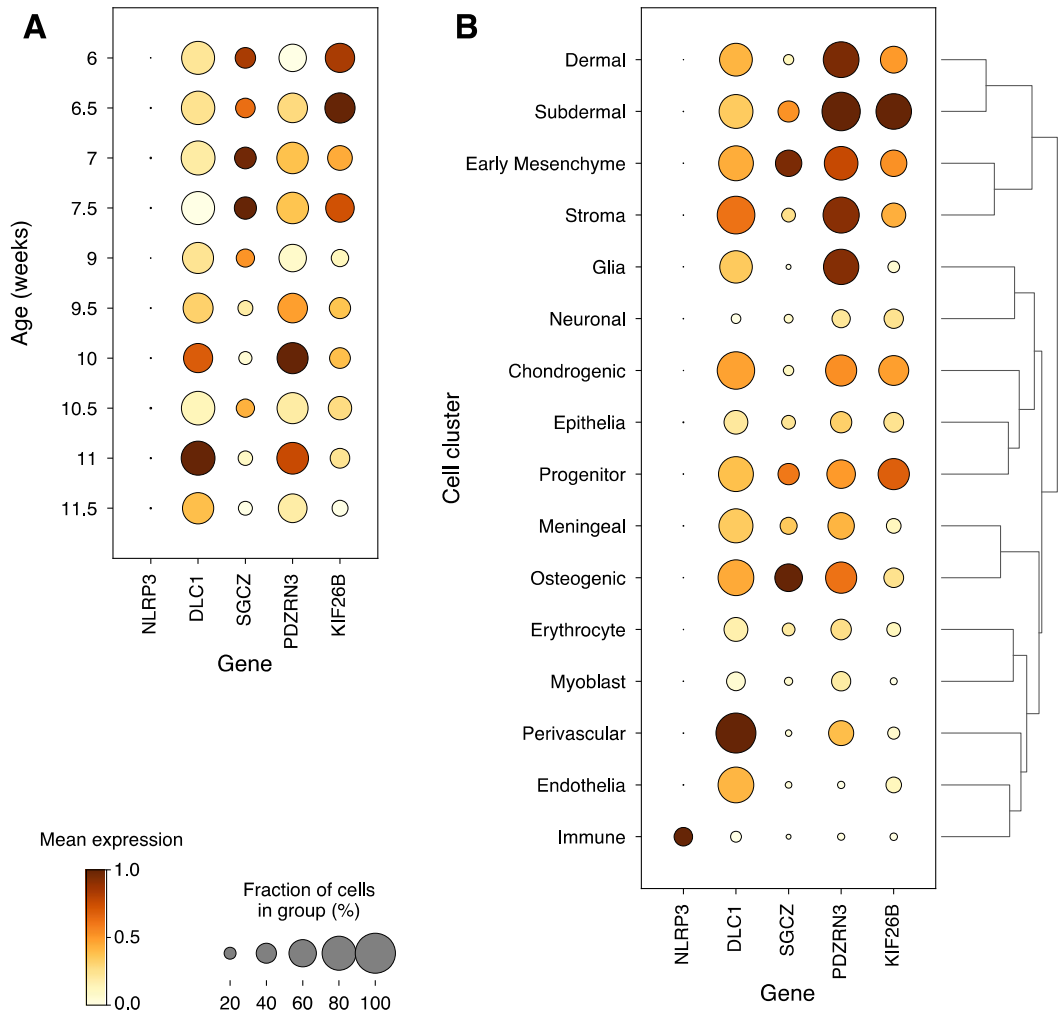


Figure 5. Dot plots showing expression patterns of genes linked to human-specific enhancers across developmental stages **(A)** and cell clusters **(B)**. Mean expression is scaled separately for each gene.

In terms of cell-type specificity (Fig. 5B), *NLRP3* displayed a strongly restricted expression pattern, confined to immune cell clusters (mean expression (ME) = 0.64, 36% of cells). *DLC1* showed robust expression across most cell types and developmental stages (ME across all clusters 1.33, sd=0.75), with the exception of neuronal (ME=0.24) and immune (ME=0.23) populations. *SGCZ* expression was more pronounced in osteogenic (ME = 1.64, 61% of cells) and mesenchymal (ME = 1.55, 57% of cells) cell types, despite its expected association with muscle-related processes and the temporal overlap with active myogenesis. The expression profile of *KIF26B* suggests a predominant role in mesenchymal derivatives. In contrast, *PDZRN3* exhibited a broader expression pattern, indicating relatively low cell-type specificity (ME across all clusters = 1.00, sd=0.58).

Next, we analyzed genes associated with primate-specific enhancers and distinct regulatory architectures (see previous section). We first focused on genes associated with a single lineage-specific enhancer to assess whether they also exhibit tissue-restricted expression patterns (Fig. 6). The most specific patterns were observed for *NEUROD1* (ME=0.28) and *CERKL* (ME=0.20), which share a single enhancer and were predominantly expressed in neuronal populations, although they were detected 18% and 14% of cells respectively within this cluster. While neuronal enrichment of *NEUROD1* was expected given its transcription factor activity, its regulatory association with *CERKL* remains unclear. *GABRB3* also showed elevated expression in neuronal and glial populations (mean expression 0.57 and 0.20 in 41% and 16% of cells respectively). Immune cells represented another distinct transcriptional context. *HSPA6*, *TYMP*, *ARHGAP45*, and *SLC12A7* showed coordinated expression primarily within immune-related clusters.

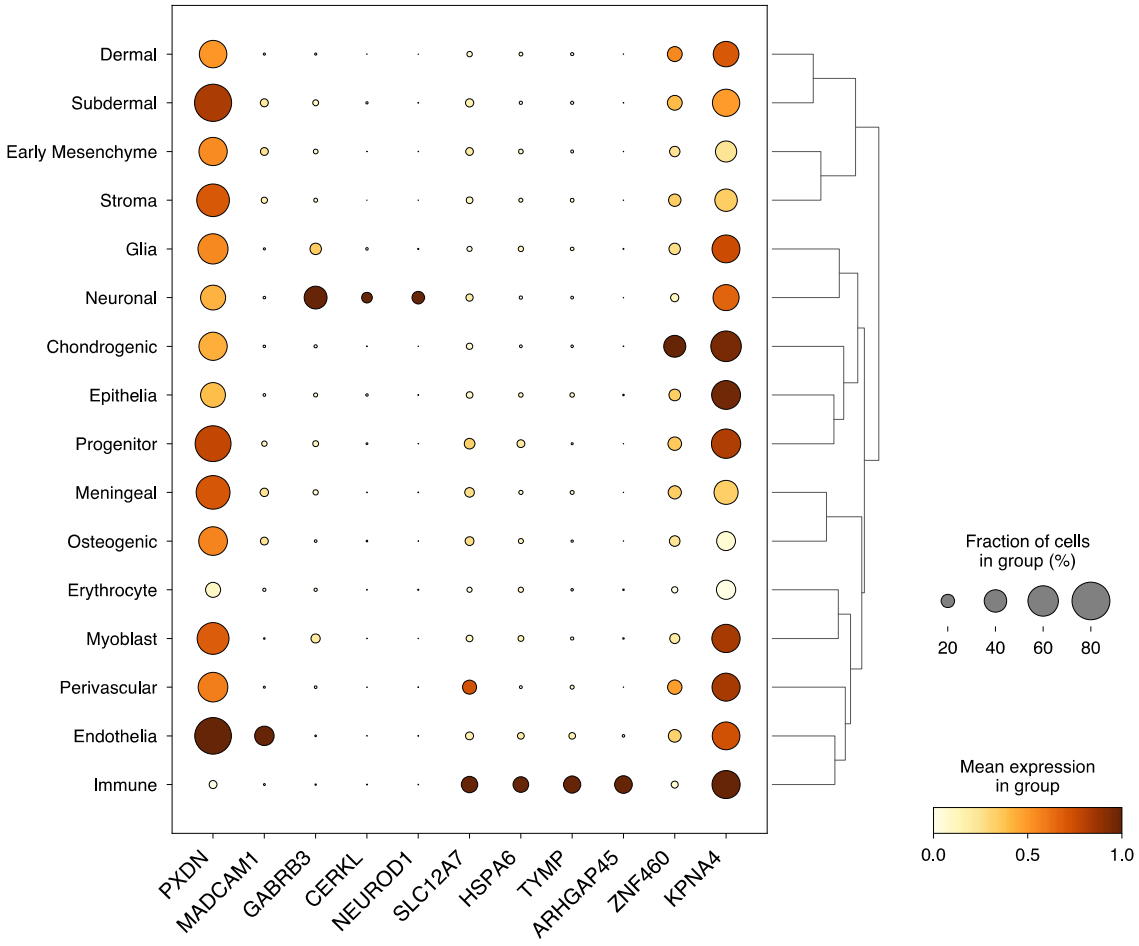


Figure 6. Dot plot showing expression patterns of genes linked to single primate-specific enhancer. Genes are clustered based on similarity of expression patterns. Mean expression is scaled separately for each gene.

Several genes exhibited broadly elevated expression across multiple cell types, including *PXDN* and *KPNA4*. *PXDN*, associated with extracellular matrix organization and collagen-related processes, was expressed in most cell populations except immune cells. *KPNA4*, encoding a karyopherin involved in nuclear transport, showed similarly broad expression.

Finally, we examined genes with elevated numbers of primate-specific enhancers to assess whether enhancer accumulation is associated with particular tissue contexts (Fig. 7).

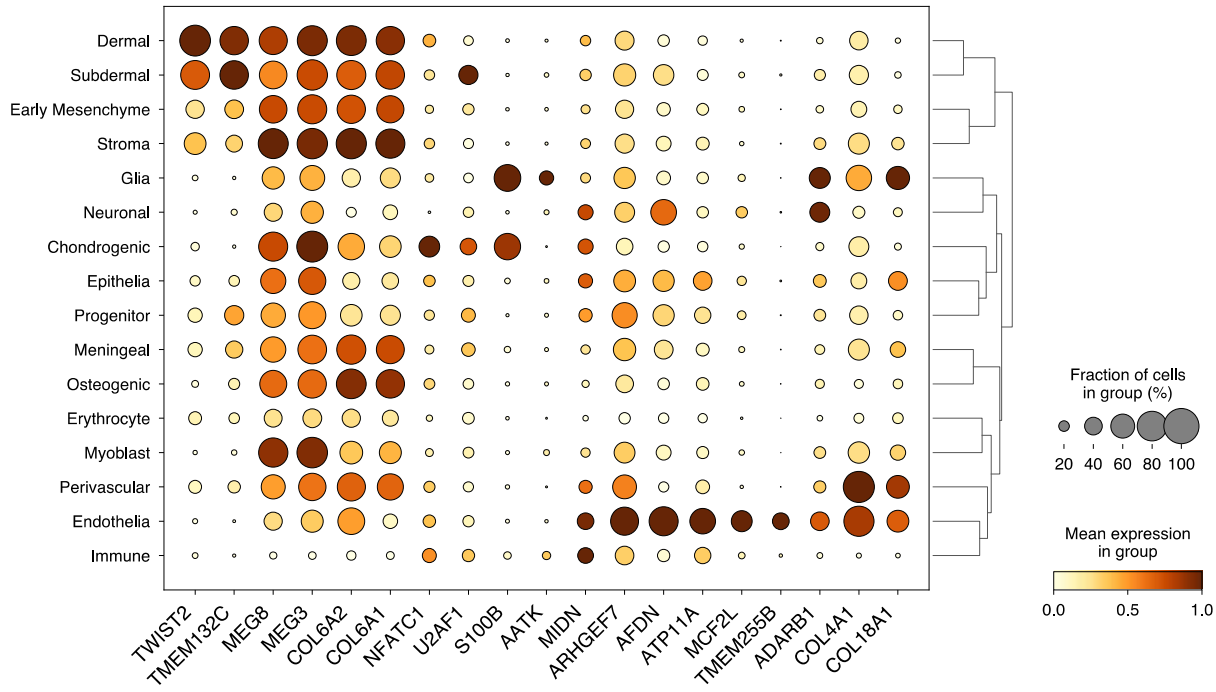


Figure 7. Dot plot showing expression patterns of genes accumulating primate-specific enhancers. Genes are clustered based on similarity of expression patterns. Mean expression is scaled separately for each gene.

Collagen genes did not exhibit a unified expression pattern. *COL6A1* and *COL6A2* were broadly expressed in neural crest derivatives, myoblasts, and perivascular cells, whereas *COL4A1* and *COL4A2* were more strongly associated with endothelia (mean expression 2.36 and 2.25 respectively, 95% of cells for both genes) and perivascular populations (ME = 2.8 in 99% cells for both genes) while maintaining moderate expression across additional cell types. *COL18A1* similarly showed enrichment in glial cells (ME = 1.10, 67% of cells) with additional strong expression in perivascular (ME = 0.94, 65% of cells) and endothelial (ME = 0.78, 61%) cell lines. In contrast, *COL20A1* displayed highly restricted expression within the glial lineage:

41% of cells demonstrated mean expression equal to 0.67, while the mean expression of all clusters together was only 0.047, sd=0.16 (collagen comparison – Supplementary Materials 5).

Several additional genes exhibited marked tissue specificity. *S100B* showed enrichment in both chondrogenic (ME = 1.64, 79% of cells) and glial (mean expression = 1.88, 81% of cells) populations. *TWIST2*, classically associated with osteogenesis, was predominantly expressed in early mesenchymal, dermal, and stromal populations rather than mature osteogenic clusters; a similar pattern was observed for *TMEM132C*. *AFDN*, the gene with the highest proportion of primate-specific enhancers, showed elevated expression in endothelial cells (ME = 1.68, 90% of cells) together with moderate neuronal expression (ME = 1.17, 76% of cells). A similar but less pronounced pattern was observed for *ADARB1*.

Other genes with high fractions of primate-specific enhancers showed lower overall expression but remained cell-type restricted. *AATK* was primarily detected in glial populations, *TMEM255B* was almost exclusively expressed in endothelial cells, and *ATP11A* also showed bias toward endothelial populations.

Interestingly, the long non-coding RNAs *MEG3* and *MEG8* exhibited relatively high expression across most cell clusters, with reduced expression in immune, endothelial, neuronal, glial, and erythrocyte populations. Together with their regulatory architecture, this raises the possibility that non-coding RNAs may contribute to lineage-specific craniofacial regulation.

Motif composition of lineage-specific enhancers

Finally, we asked what novel regulatory contributions the newly emerged enhancers may provide beyond the existing regulatory landscape of genes. The functional units of enhancers are transcription factor binding sites (TFBSs), which are short sequence motifs recognized by specific transcription factors (TFs). In this context, several scenarios can be considered: (1) newly emerged enhancers may introduce novel motifs, enabling the recruitment of new regulatory programs and potentially extending the spatial domain of the target gene; or (2) they may increase the abundance or diversity of motifs for existing TFs, thereby modulating gene expression patterns. To investigate these possibilities, we performed motif analysis in enhancers associated with genes that possess at least one lineage-specific enhancer.

Addressing the question of motif composition presented two major challenges: (1) the number of enhancers of interest (primate- and human-specific categories) was small compared to the full set of regulatory regions (84 versus 1664); and (2) enhancer sequences are inherently

noisy, while transcription factor binding sites are short (typically 8–20 bp), making it difficult to distinguish true biological signal from random sequence patterns.

For feasibility, we restricted the analysis to a specific subset of regulatory regions. We selected all enhancers associated with genes that possess at least one lineage-specific enhancer, resulting in a set of 901 enhancers, of which 84 correspond to primate-specific elements. We then counted enhancers associated with each gene in this subset to account for differences in regulatory complexity in downstream analyses. With this set of enhancers, we tried several approaches to unveil the transcription factor sites landscape in our dataset.

Motif enrichment analysis

Motif enrichment analysis is a common technique that allows the identification of overrepresented motif sequences relative to a defined or randomly generated background, using position weight matrices (PWMs) for known motifs.

We performed motif enrichment analysis under two complementary settings. First, we conducted a per-gene comparison to assess whether newly emerged enhancers harbor motifs enriched relative to other enhancers associated with the same gene. This approach was intended to identify both unique motifs and motifs whose representation increases with the appearance of lineage-specific regulatory elements. However, even for enhancer-rich genes such as *TWIST2* or members of the collagen family, no significantly enriched motifs were detected. Most likely, the number and length of sequences are insufficient to determine enrichment.

We then asked whether primate-specific enhancers share a common motif signal compared to all other 901 enhancers. This analysis assumed that partial reconstruction of shared regulatory pathways – for example, extracellular matrix remodeling – could lead to coordinated regulatory responses mediated by one or several transcription factors and, therefore, acquiring of its binding sites. In this setting, only a single motif was identified: ZNF418 (TP = 34.5%, FP = 11.5%, p-value = 2.85×10^{-5}), related to a zinc-finger transcriptional repressor. ZNF418 has been implicated in cellular homeostasis through regulation of the ubiquitin–proteasome system and suppression of MAPK signaling (Li et al. 2008; Singh et al. 2021). However, its direct relevance to neural crest–related processes is unclear.

To further assess intrinsic motif enrichment, we compared all primate-specific enhancers against a control set of shuffled sequences with similar GC-content. This analysis aimed to identify motifs that are generally overrepresented in the enhancer set, independent of comparison to other enhancers. The top enriched motifs predominantly belonged to Krüppel-

like and zinc-finger transcription factor families, which are broadly distributed and therefore provide limited specificity for functional interpretation.

TF-IDF scoring

Another approach to account for the heterogeneous structure of transcription factor binding sites within enhancers is the use of Term Frequency-Inverse Document Frequency (TF–IDF) scoring. This metric defines a "weight" for each motif in a given enhancer by considering both its frequency within the sequence (term frequency, TF) and its prevalence across all enhancers in the analyzed set (inverse document frequency, IDF). As a result, TF–IDF down-weights motifs that are broadly enriched (e.g., zinc-finger or KLF families) and highlights motifs that are relatively rare but locally overrepresented within specific enhancer regions.

Nevertheless, this approach remains sensitive to factors like the number of enhancers in the compared groups. Intuitively, larger enhancer sets increase the probability of motif occurrence and, more importantly, can inflate the contribution of abundant motif families, thereby reducing the discriminative power of the method. For that reason, we restricted the analysis to carefully selected subsets of enhancers in which differences in sequence length and group size were minimized, allowing for more balanced and interpretable comparisons.

Therefore, we analyzed 2 genes with an equal number of enhancers per category and comparable length: AFDN (with 4 primate-specific and 4 non-specific enhancers) and SGCZ (1 human-specific and 1 non-specific enhancer).

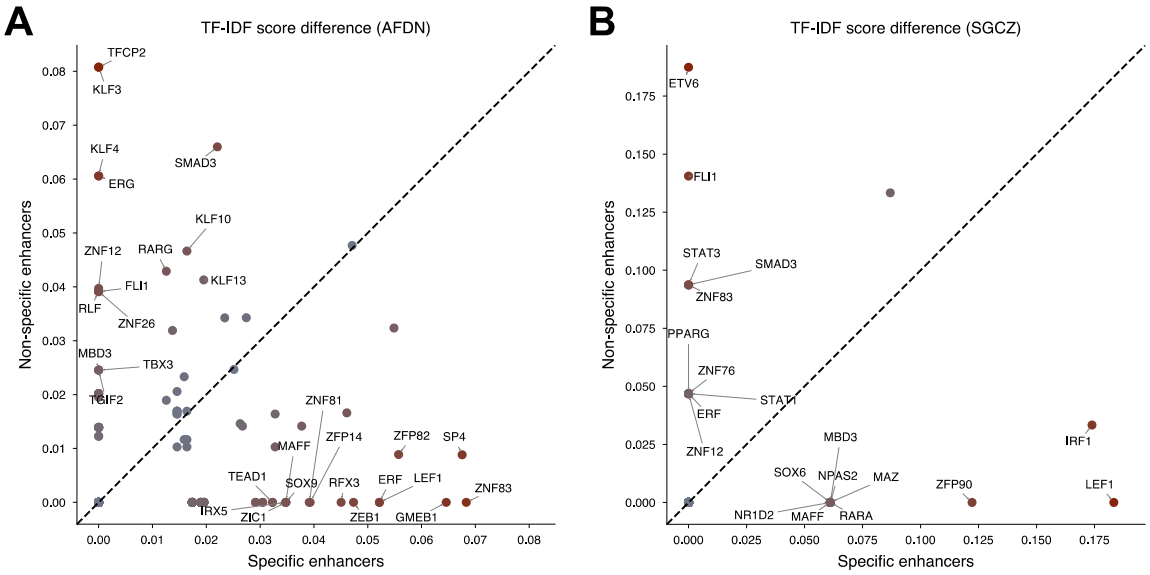


Figure 8. TF-IDF scatter plots for the regulatory architecture of two genes. X-axis – TF-IDF mean scores for a subset of specific enhancers (primate-specific or human-specific), y-axis TF-IDF mean scores for a subset of non-specific enhancers, dot color indicates the mean difference in absolute values. **A.** AFDN gene **B.** SGCZ gene.

Interestingly, for AFDN, notable transcription factor binding sites were identified in both enhancer groups (Fig.8A). In non-specific enhancers, enriched motifs included TBX3, involved in mesoderm-derived tissue development; SMAD3, a downstream effector of TGF-β/BMP signaling with roles in mesenchymal differentiation; RARG, a retinoic acid receptor broadly implicated in embryonic development; and MBD3, a chromatin remodeling factor associated with the NuRD repressor complex. The remaining enriched motifs belonged to broadly expressed transcription factor families, including KLFs, ZNFs, and ETS factors, without clear tissue-specific relevance in this context.

In primate-specific enhancers, the most relevant identified binding site was SOX9 – a master regulator of chondrogenesis and the only motif in this set with well-documented neural crest relevance. The remaining motifs included TEAD1, an effector of the Hippo signaling pathway involved in mechanosensing and tissue growth; RFX3, primarily associated with ciliary gene regulation; and GMEB1, whose role in embryonic development remains poorly characterized. Other sites belonged to broadly expressed regulatory factor families.

For SGCZ, a partially overlapping regulatory signature was observed (Fig.8B). Non-specific enhancer was enriched for SMAD3; STAT3, a JAK-STAT pathway effector with reported roles in bone formation and mesenchymal cell behavior; and RXRA, a retinoid receptor that functions as a heterodimeric partner in retinoic acid signaling. Within the human-specific enhancer, notable motifs included SOX6, a chondrogenic transcription factor acting within the SOX5/SOX6/SOX9 regulatory module; TEAD3, a Hippo pathway effector involved in tissue growth regulation; and RARA, a retinoic acid receptor. The two most enriched motifs, LEF1 and IRF1, are broadly associated with Wnt signaling and interferon-mediated immune responses, respectively, without established craniofacial or neural crest specificity.

To summarize, analysis of gene-specific regulatory architecture suggests that newly emerged enhancers may introduce additional regulatory control for target genes. However, in most cases, we observe enrichment of motifs associated with already established regulatory pathways, which may reflect modulation of expression output or shifts in temporal and spatial expression context (e.g., retinoic acid signaling and general mesenchymal programs).

Nevertheless, in some instances, lineage-specific enhancers contribute novel motif inputs (e.g., SOX-mediated chondrogenic regulation).

Discussion

Craniofacial morphology is a prominent example of a complex phenotypic trait that balances evolutionary conservation with adaptive variability in vertebrates, particularly in mammals. Core anatomical structures such as the jaws, palate, and neurocranium are preserved across species, yet display substantial variation in shape and proportion, reflecting both interspecific and, in some cases, interindividual differences (Mitsiadis et al. 2006; Tucker and Lumsden 2004; Cohen et al. 2020; Schneider 2018). In recent years, comparative and developmental studies have attempted to reconstruct the evolutionary trajectories of primate and hominid craniofacial structures, partly also to better understand the origins of congenital craniofacial disorders in humans (von Cramon-Taubadel 2014; Prescott et al. 2015).

Accumulating evidence suggests that a major source of craniofacial variation lies within the neural crest cell population (Le Douarin et al. 2004; Schneider 2018; Mitsiadis et al. 2006). During development, neural crest cells migrate and integrate environmental signals, effectively superimposing species-specific parameters such as size and shape onto conserved developmental frameworks, including axial orientation, anatomical identity, and tissue type (Fish and Schneider 2014). While these core developmental programs and neural crest gene regulatory networks are relatively well characterized, the mechanisms underlying species-specific patterning remain incompletely understood.

Recent studies in humans, great apes, and domesticated species further suggest that a substantial fraction of neural crest–derived craniofacial variation and evolution arises from changes in regulatory regions rather than protein-coding sequences (King and Wilson 1975; Prescott et al. 2015; Zanella et al. 2019; Carneiro et al. 2014; Rubio and Summers 2022). This concept, together with increasing evidence from GWA studies, linking complex traits to non-coding variation (Maurano et al. 2012; Goovaerts et al. 2026), motivated the present study on enhancer regions responsible for craniofacial development. Specifically, we asked (1) whether lineage-specific enhancers can be identified in primates and humans, (2) how these elements relate to gene expression patterns and whether their effects appear coordinated or gene-specific, and (3) whether enhancer sequence composition provides insight into their functional roles. To address these questions, we identified lineage-specific enhancers, linked them to gene

expression using craniofacial multiome data, and analyzed their sequences in the context of GWAS signals and transcription factor binding motifs.

Primate-specific enhancers shaping lineage-specific patterns

First, we identified 84 primate-specific enhancers. Given the size of the set, we asked whether particular regulatory pathways are preferentially enriched in newly emerged elements within the primate lineage. Indeed, collagen genes and collagen-associated regulators showed a notable accumulation of primate-specific enhancers.

On the one hand, this result is biologically plausible. Collagens are key components of the extracellular matrix and play an essential role in epithelial–mesenchymal interactions that guide neural crest cell migration. According to the "flypaper" model (Hanken and Thorogood 1993), collagen-rich environments can act as adhesive substrates that influence cell movement by modulating attachment and tissue penetrability. In this context, fine-tuning collagen expression may affect migration trajectories and, consequently, the size and shape of craniofacial structures. This interpretation is further supported by observations that collagen genes exhibit a high total number of enhancers, with a notable fraction being primate-specific (e.g., *COL4A1, COL18A1, COL6A1, COL6A2*, each with >20 enhancers and >7% lineage-specific ones).

However, this interpretation should be considered with caution. The collagens identified in this study are not the classical drivers of neural crest migration. Most experimental evidence highlights type I and type II collagens (e.g., COL1A1/2) as primary regulators of extracellular matrix dynamics in neural crest development (Duband and Thiery 1987; McCauley 2008; Oka et al. 2012), and mutations in these genes are strongly associated with craniofacial abnormalities (Cruz Walma and Yamada 2022). In contrast, the collagens enriched in our dataset appear to play more context-specific or less well-characterized roles. For example, *COL4A1* is associated with basement membranes and has been implicated in guiding neural crest migration, although most supporting evidence comes from avian models (Perris et al. 1993; Duband and Thiery 1987). Given the pronounced expression of *COL4A1/2* in vascular cell culture, we can propose another role of the genes in craniofacial development – regulating vascularization processes, it is indirectly involved in face shape rearrangements since it guides subsequent ossification. *COL6A1/2* genes have been linked to extracellular matrix organization during tooth development, which may relate to species-specific dentition patterns (Cruz Walma and Yamada 2022), but there is no evidence of that hypothesis to date. *COL9A3*, despite its

expression in chondrocytes, has not been directly implicated in neural crest biology, while *COL18A1* and *COL20A1* show limited or diffuse expression in the developmental stages examined here.

Taken together, these observations suggest that primate-specific regulatory changes may preferentially target non-canonical components of the extracellular matrix, potentially contributing to subtle, context-dependent modulation of craniofacial morphogenesis rather than large-scale restructuring of core developmental pathways.

Beyond the collagen-related regulatory layer, we observed more modest and gene-specific rearrangements in regulatory logic associated with primate craniofacial development. For example, we focused on genes that possess only a single lineage-specific enhancer, as these cases may reflect more direct and less buffered regulatory effects.

In unpublished in-lab analyses integrating GWAS and multiome data, an observation was made that neuronal populations may provide signaling cues to mesenchymal cells involved in craniofacial skeletogenesis. This hypothesis is supported by experimental evidence showing that Neurog-1 knockout disrupts facial skeletal development in mouse, although the underlying mechanisms remain unclear (Erickson et al., under revision).

Consistent with this idea, NEUROD1 - a transcription factor with a well-established role in neuronal differentiation - was identified in our dataset as being associated with a single primate-specific enhancer. A similar pattern is observed for *GABRB3*, another gene with a single enhancer, for which associations with cleft formation under environmental perturbations have been reported (D'Souza et al. 2010). Together, *NEUROD1* and *GABRB3* were expressed in neuronal cells clusters without temporal specificity in accordance to our single-cell data. *GABRB3* was also presented in glia cell population. Moreover, previously mentioned COL18A1 and COL20A1 which are not well-studied yet, showed a notable expression within glial cell population.

Thus, these observations additionally point to a possible contribution of neural signaling to craniofacial morphogenesis, where lineage-specific regulatory changes in neural genes may indirectly influence skeletal development through cross-tissue interactions.

The final gene with a potentially meaningful regulatory rearrangement in the primate lineage is *AFDN*, for which we observed that 50% of its associated enhancers (4 out of 8) are primate-specific. This gene has previously been implicated in cleft lip formation (Rebeiz et al. 2011; D'Souza et al. 2010; Awotoye et al. 2024), linking it to craniofacial development. At the same time, its association with lineage-specific enhancers suggests that its contribution may be mediated at the regulatory level rather than through changes in the protein-coding sequence

forming new patterns specific for the group. Moreover, *AFDN* was one of the genes for which we were able to perform motif content analysis across its enhancers. The results support a scenario in which general regulatory programs, associated with more common enhancers (enriched for transcription factor binding sites such as SMAD, TBX3, and RARG), are reinforced and complemented by additional, less specific motifs in primate-specific enhancers. These lineage-specific elements appear to introduce both overlapping and potentially novel regulatory inputs, including binding sites for transcription factors such as SOX9, TEAD1, and RFX3.

AFDN has so far been primarily associated with epithelial cell organization and cell–cell junctions rather than neural crest–derived lineages This is consistent with our data, where *AFDN* showed the highest mean expression in endothelial cells. However, its expression pattern was not restricted to this population: expression was also detected in epithelial, subdermal, neuronal, and progenitor cell populations, with more than 50% of cells expressing the gene. Notably, neuronal clusters also exhibited relatively high expression levels.

Therefore, a direct connection between *AFDN* and neural crest–specific processes remains uncertain. Instead, this case suggests that not all lineage-specific regulatory changes identified in this study are directly linked to neural crest migration or differentiation. Rather, this case highlights the role of additional cell types in shaping craniofacial morphology and suggests that species-specific patterning may arise from coordinated regulatory changes across multiple interacting tissues.

Human-specific enhancers shaping species-specific patterns

In parallel, we analyzed five human-specific regulatory elements that were not detected in any other species according to Zoonomia alignment. Such elements are expected to contribute to highly specific regulatory programs and may underlie both human-specific craniofacial features (e.g., frontal bone height or chin formation) and individual variation. Notably, recent work suggests that humans are the only primates exhibiting a non-neutral evolutionary rate in craniofacial regions, whereas other primates show more pronounced variation in neurocranial evolution (Gómez-Robles et al. 2025). This raises the possibility that newly emerged regulatory elements may contribute to this pattern.

Interestingly, two human-specific enhancers overlapped with previously reported SNPs associated with facial variation (Goovaerts et al. 2026). Therefore, we further investigated the genes linked to these enhancers.

The first gene, *DLC1*, is a broadly expressed regulator of embryonic development supported by a relatively complex regulatory architecture (10 associated enhancers). *DLC1* encodes a Rho GTPase-activating protein involved in cytoskeletal organization, cell migration, and angiogenesis. In embryonic models, loss of *DLC1* function leads to defects in vascular development and cell migration, as well as embryonic lethality in homozygous knockouts (Durkin et al. 2005).

In the context of neural crest biology, *DLC1* has been implicated in regulating trunk neural crest polarity via Rho signaling. While its role in craniofacial neural crest remains unexplored, its known functions suggest a plausible involvement in craniofacial morphogenesis through modulation of fibroblast behavior, cytoskeletal dynamics, and cell migration. Notably, the presence of SOX9 motifs within the human-specific DLC1 enhancer in our data further supports a potential connection to mesenchymal and chondrogenic processes. At the same time, its role in endothelial cell regulation suggests that its effects on facial morphology may be indirect, mediated through vascular or tissue integration mechanisms rather than neural crest specification itself.

The second gene, *SGCZ*, encodes a component of the sarcoglycan complex, which is part of the dystrophin–glycoprotein system linking the actin cytoskeleton to the extracellular matrix (Wheeler 2003). Although primarily associated with muscle structure, its role in embryonic development remains poorly characterized. In our dataset, *SGCZ* shows a more pronounced expression in mesenchymal and osteogenic cell populations, rather than in myoblasts, suggesting a broader or context-dependent function.

SGCZ is associated with only two enhancers within the developmental window analyzed, indicating a relatively simple regulatory architecture. Motif analysis revealed that the human-specific enhancer introduces SOX6 binding sites, potentially linking it to mesenchymal differentiation, while additional motifs such as TEAD3 and RARA suggest involvement of epithelial and growth-regulatory pathways. This combination may reflect integration of multiple regulatory inputs rather than a single dominant developmental program.

Overall, neither *DLC1* nor *SGCZ* provides a straightforward mechanistic explanation for how individual regulatory variants lead to craniofacial phenotypic variation. However, both genes highlight a broader pattern: human-specific regulatory changes may act through modulation of cell migration, tissue mechanics, and cross-tissue interactions rather than through direct alteration of core developmental identity programs. These observations emphasize the need for functional validation to disentangle the contributions of individual regulatory elements.

Conclusions and future directions

Overall, our results suggest that the evolution of regulatory regions may contribute both to coordinated shifts in morphogenetic patterns (e.g., through collagen-related regulatory changes) and to gene-specific effects (e.g., *AFDN*, *NEUROD1*, *DLC1*, and *SGCZ*). Notably, we did not identify classical key regulators of craniofacial neural crest development among lineage-specific elements. Instead, our findings point to the importance of precise downstream regulatory modulation in generating phenotypic variation without disrupting core developmental programs.

Importantly, this study also defines several directions for future research.

First, experimental validation of candidate enhancer-gene pairs is required. This includes addressing multiple related questions: how do gene knockouts compare to enhancer perturbations in their phenotypic effects? What is the degree of regulatory robustness within each gene's enhancer landscape? Do individual enhancers associated with the same gene produce different effects under knockout? Answering these questions will be essential for understanding the functional impact of regulatory variation.

Second, initially, the analysis identified an additional group of interest – conserved regulatory elements. It would be valuable to examine how these elements are distributed across genes and whether they are enriched for variants identified in GWA studies. Furthermore, their conservation could be assessed not only at the level of presence or absence but also in terms of sequence similarity. Additionally, it would be particularly informative to test whether these elements overlap with known ultra-conserved regulatory regions in vertebrate genomes (Woolfe et al. 2004).

Third, the question of human-specific craniofacial variation can be explored in greater depth. Most importantly, we need to develop a plausible scenario for the emergence of human-specific regulatory elements within a relatively short evolutionary period. Having defined these elements through cross-species comparisons, future work could then examine how they vary across present-day human populations and whether this variation correlates with phenotypic diversity. An additional promising direction would be to compare these regions with genomic data from other Homo lineages, such as Neanderthals.

In conclusion, while this study provides initial insights into the role of regulatory evolution in craniofacial development, it also highlights the complexity of these processes. In attempting to answer the initial questions, we have identified both partial explanations and new questions that may guide future research in this field.

Materials and Methods

Data preprocessing and enhancer conservation analysis

Enhancer coordinates used in this study were obtained from a previously generated and processed multiome-derived dataset (Erickson et al., under revision). This dataset integrates information on open chromatin regions identified by single-cell ATAC-seq during neural crest–relevant stages of human embryonic development (weeks 6 to 12.5), together with corresponding gene expression profiles derived from single-cell RNA-seq and single-nuclei RNA-seq. In addition, the dataset includes epigenomic region annotations (e.g., transcription start sites, enhancers, strong and weak transcription states) based on a previously published ChromHMM 25-state chromatin model (Roadmap Epigenomics Project, Supplementary Materials 2) constructed from publicly available epigenomic datasets. Using these annotations, we filtered open chromatin regions classified as enhancers (ChromHMM states 10–18), resulting in a set of 11,748 enhancer regions. This set represents putative enhancers associated with craniofacial neural crest development during human embryogenesis.

For cross-species comparisons, we used the publicly available multiple sequence alignment (MSA) of 241 mammalian genomes generated by the Zoonomia project (Genereux et al. 2020). The alignment is accessible via the UCSC Genome Browser and is provided in bigMAF format, with the human genome (hg38, GRCh38, December 2013 assembly) used as the reference. Several versions of the 241-species alignment are available; in this study, we used the original release (2020) to ensure consistency with previously reported evolutionary analyses.

Due to computational constraints, a representative subset of 22 mammalian species spanning major evolutionary lineages is selected (Table 4). While this subsampling limits fine-scale evolutionary inference, it preserves broad phylogenetic structure and enables robust comparative analysis within available computational resources. Importantly, the subset includes eight primate species representing major branches of the primate clade. Great apes – gorilla (*Gorilla gorilla*), bonobo (*Pan paniscus*), and Sumatran orangutan (*Pongo abelii*) – were included together with other Old World monkeys, represented by the proboscis monkey (*Nasalis larvatus*) and rhesus macaque (*Macaca mulatta*). New World monkeys were represented by the Geoffroy's spider monkey (*Ateles geoffroyi*) and common marmoset (*Callithrix jacchus*). We also included the blue-eyed black lemur (*Eulemur flavifrons*) as a more distant representative within the primate clade.

Enhancer regions were extracted from the MSA for all selected species and exported in both MAF and FASTA formats using custom Bash and Python scripts (Supplementary Materials 1). The scripts were built using publicly available tools and libraries: BEDTools (v. 2.31.1, Quinlan and Hall 2010), Biopython (v. 1.85, Cock et al. 2009), and bx-python (v. 0.14.0). This resulted in a set of 11,748 alignment files, each containing the human enhancer sequence as the reference, in accordance with the multiome dataset, together with the corresponding homologous regions from the selected species.

Table 4. Species sample

Order	**Species**
Carnivora	*Felis nigripes*
Carnivora	*Mungos mungo*
Carnivora	*Neomonachus schauinslandi*
Carnivora	*Ursus maritimus*
Cetartiodactyla	*Balaenoptera acutorostrata*
Cetartiodactyla	*Bison bison*
Cetartiodactyla	*Camelus ferus*
Cetartiodactyla	*Sus scrofa*
Perissodactyla	*Ceratotherium simum*
Perissodactyla	*Equus przewalskii*
Perissodactyla	*Tapirus indicus*
Primates	*Ateles geoffroyi*
Primates	*Callithrix jacchus*
Primates	*Eulemur flavifrons*
Primates	*Gorilla gorilla*
Primates	*Macaca mulatta*
Primates	*Nasalis larvatus*
Primates	*Pan paniscus*
Primates	*Pongo abelii*
Rodentia	*Marmota marmota*
Rodentia	*Mesocricetus auratus*
Rodentia	*Rattus norvegicus*

To quantify enhancer preservation, we evaluate the presence of aligned sequence for each enhancer across species relative to the human reference. Preservation is defined as the

proportion of enhancer positions covered by aligned sequence, regardless of nucleotide identity. This gap-based metric captures the evolutionary persistence of regulatory elements while minimizing the confounding effects of sequence divergence. Enhancers fully aligned without gaps were assigned 100% preservation, and the presence of gaps proportionally reduced this value. Then, enhancers were categorized in 4 groups based on its level of preservation (see Results): conserved in mammals, primate-specific, human-specific and other.

For defining the primate-specific group, a relatively soft threshold was selected after testing multiple parameter combinations. The final criterion - all primates (n = 8) show >90% preservation, while at least half of the remaining species show <10% preservation - provided an optimal balance between capturing a sufficient number of enhancers with strong conservation within primates and ensuring a clear preservation bias toward the clade. Importantly, this classification does not imply *de novo* emergence during primate evolution, but rather lineage-specific sequence preservation.

Following the presence/absence approach, it is also important to emphasize its potential limitations. Since sequence alignment algorithms rely on statistical estimation of regional alignability, they inherently produce a fraction of false-positive alignments and may be particularly sensitive to misaligned regions. Therefore, some of the described insertions or deletions may reflect alignment artifacts rather than true biological rearrangements. At the same time, these two phenomena are not necessarily independent, as genuine sequence rearrangements reduce regional similarity and consequently decrease the probability of correct alignment.

In parallel, sequence variability is assessed by comparing aligned nucleotides in each species to the human reference on a position-by-position basis. A similarity score is calculated for each enhancer–species pair, producing a second matrix that explicitly captures sequence divergence.

Scoring is based on a penalty matrix in which each type of mismatch is assigned a negative value: –1 for mismatches between chemically similar nucleobases (purine–purine or pyrimidine–pyrimidine), –2 for mismatches between different classes of nucleobases (purine–pyrimidine), and –3 for each gap. Thus, the more similar two sequences are, the higher (i.e., less negative) their similarity score. This approach follows standard principles commonly used in sequence alignment algorithms (Chao et al. 2022).

GWAS data integration

The original multiome dataset included GWAS-derived variants collected from previously published studies (e.g. Shaffer et al. 2016b; White et al. 2021; Claes et al. 2018; Liu et al. 2021)). In early 2026, a new study was released that substantially expanded the catalog of SNPs associated with craniofacial traits and neural crest–related biology (Goovaerts et al. 2026). We integrate the newly reported variants into our analysis.

Genomic coordinates of SNPs are intersected with enhancer regions based on coordinate overlap. As the initial study used the hg19 human genome assembly, all SNP coordinates are converted to hg38 using liftOver v.482 (Hinrichs 2006). Following coordinate integration, SNP-associated metadata are integrated for all matched regions (Supplementary Materials 1).

Functional enrichment analysis

Gene expression analysis is followed by functional enrichment testing using the STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins). STRING integrates experimentally validated and predicted protein–protein interactions and combines information from curated pathway databases, co-expression data, and computational predictions.

For each gene set, enrichment analysis is performed to identify overrepresented biological processes and pathways using Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome annotations. As a background, we use all genes detected within or in strong association with open chromatin regions according to the given dataset (n=6095). The false discovery rate indicates the significance of the enrichment, and the p-values for each category are subjected to multiple-testing correction using the Benjamini-Hochberg procedure. In addition, STRING network clusters is used to evaluate functional connectivity among proteins encoded by the input genes by reconstructing interaction networks.

Single-cell RNA-seq expression analysis

The processed single-cell RNA-seq dataset derived from multiome gene expression data enabled the identification of enhancer-associated effects in a tissue- and cell type–specific context (Fig. 9). These data was already preprocessed (including quality control and normalisation), clustered and annotated (Erickson et al, under revision). All enhancer analysis-

related data processing and visualization were performed in the framework of the thesis using the Python packages anndata (v. 0.12.2, Virshup et al. 2024)) and scanpy (v. 1.11.4, Wolf et al. 2018)), along with supporting libraries including pandas (v. 2.3.1, McKinney 2010), numpy (v. 2.2.6, Harris et al. 2020), scipy (v. 1.16.1, Virtanen et al. 2020), and matplotlib (v. 3.10.5, Hunter 2007) (Supplementary Materials 1).

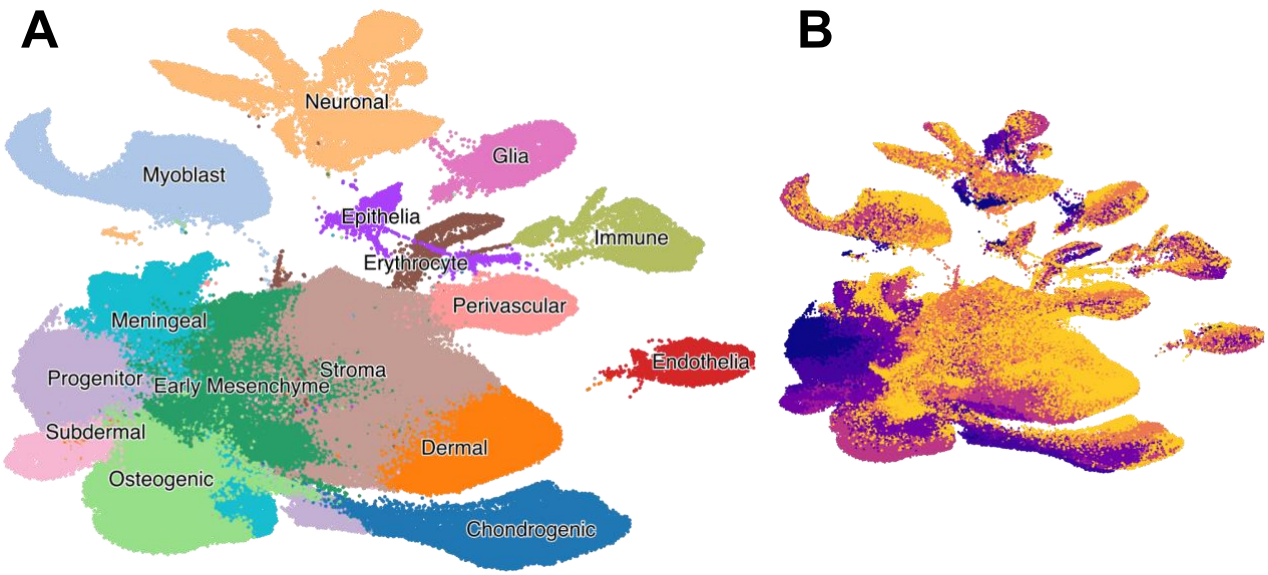


Figure 9. (A) UMAP embedding of multiome data colored by cell cluster. **(B)** UMAP embedding of multiome data colored by developmental stage (Erickson et al, under revision).

To assess expression differences between genes associated with lineage-specific enhancers, we calculated mean expression values using normalized and log-transformed counts from the original dataset. To identify the top-expressing cluster for each gene, mean expression was calculated across cell clusters using non-zero expression values only and was accompanied by the fraction of cells expressing the gene within each cluster. Same was applied for examination of expression per developmental stage.

To evaluate cell-type specificity, we combined visual inspection of expression patterns generated with the *sc.pl.dotplot* function with a quantitative enrichment metric, defined as the ratio between expression in the top-expressing cluster and the mean expression across all clusters.

Transcription factor binding sites (motifs) analysis

Due to the complexity and limitations of the input data, we apply multiple complementary approaches to characterize the regulatory landscape of the target enhancers. To improve interpretability and reduce potential biases, we restrict the analysis to a subset of enhancers associated with genes that possess at least one lineage-specific enhancer. This resulted in a set of 901 enhancers, of which 84 were classified as primate-specific. Human-unique enhancers are not included in this analysis due to their limited number and are instead examined separately and selectively (see SGCZ TF-IDF scoring in Results).

Motif enrichment

The first approach aims to identify motifs enriched across a substantial fraction of lineage-specific enhancers compared to the rest of the dataset, based directly on sequence composition. For this purpose, motif enrichment analysis is performed using MEME AME (McLeay and Bailey 2010).

Motif enrichment analysis identifies overrepresented transcription factor binding sites (TFBSs) in a set of sequences by comparing them to a defined background. The analysis relies on curated TFBS databases and either shuffled sequences (preserving GC content) or user-defined background sets.

The output consists of a list of enriched motifs, including the proportion of sequences containing the motif in the query set (TP%) and in the background set (FP%), along with associated statistical measures such as adjusted p-values and E-values. Motifs with a large difference between TP% and FP% are considered enriched, indicating that they occur more frequently than expected given the sequence context.

We apply this approach under several conditions using high-confidence human motif collection HOCOMOCO v.14 (Vorontsov et al. 2024) as a motif database in all cases:

1) Gene-level enrichment: lineage-specific enhancers versus non–lineage-specific enhancers associated with the same gene;
2) Global enrichment: all lineage-specific enhancers as the query set versus all other enhancers linked to the same genes as background;
3) Control enrichment: lineage-specific enhancers as the query set versus GC-matched shuffled sequences as background.

Resulting enrichment tables are further processed and filtered using custom python scripts (Supplementary Materials 1).

TF-IDF scoring

The second approach - TF-IDF scoring - focuses on motif occurrence patterns rather than raw sequence enrichment. This method accounts for the heterogeneous structure of enhancers and reduces biases arising from enhancer length and the inherent overrepresentation of certain motif families.

To compute TF–IDF scores, transcription factor binding motifs are first mapped onto enhancer sequences. This required (i) position weight matrices (PWMs) representing nucleotide frequencies at each position of known motifs, and (ii) a computational tool to scan sequences for statistically significant motif occurrences.

We use HOCOMOCO v.14 motif collection, and scanned human enhancer sequences using MEME FIMO (Grant et al. 2011). Motif searches are performed using sequence-specific background models, with p-value thresholds adjusted for enhancer length to reduce biases related to sequence composition and size (Supplementary Materials 1). FIMO output is further filtered using a stringent adjusted p-value threshold (q-value < 0.01).

To restrict the analysis to biologically relevant transcription factors, we retain only motifs corresponding to TFs expressed in the multiome dataset. TF expression is estimated from scRNA-seq data, and only TFs expressed in more than 10% of cells with a log-normalized expression value greater than 0.1 were included. This threshold is chosen to exclude low-confidence expression signals.

Following motif mapping, we apply a TF–IDF scoring scheme to quantify the relative contribution of each motif to individual enhancers. In this framework, each enhancer is treated as a "document" and each motif as a "term." Term frequency reflects the number of occurrences of a motif within a given enhancer relative to the total number of motif occurrences in that enhancer ($tf_{x,y}$), whereas inverse document frequency ($\log N / df_x$) down-weights motifs that are frequent across many enhancers.

$$\text{TF-IDF}_{x,y} = \text{tf}_{x,y} \cdot \log\left(\frac{N}{\text{df}_x}\right)$$

As a result, TF–IDF highlights motifs that are both frequent within a specific enhancer and relatively rare across the enhancer set, thereby emphasizing candidates more likely to contribute to enhancer-specific regulatory functions rather than general binding activity.

This procedure produces a matrix with motifs as rows and enhancers as columns, where each entry represents the TF–IDF score of a motif in a given enhancer (Supplementary Materials 1). In the reported results, TF–IDF scores are calculated within particular gene-specific enhancer sets (AFDN, SGCZ).

References

1) Andrews, Gregory, Kaili Fan, Henry E. Pratt, et al. 2023. “Mammalian Evolution of Human Cis-Regulatory Elements and Transcription Factor Binding Sites.” *Science* 380 (6643): eabn7930. https://doi.org/10.1126/science.abn7930.

2) Awotoye, Waheed, Peter A. Mossey, Jacqueline B. Hetmanski, et al. 2024. “Damaging Mutations in AFDN Contribute to Risk of Nonsyndromic Cleft Lip with or without Cleft Palate.” *The Cleft Palate-Craniofacial Journal : Official Publication of the American Cleft Palate-Craniofacial Association* 61 (4): 697–705. https://doi.org/10.1177/10556656221135926.

3) Britten R.J., Britten R. J. 1996. “DNA Sequence Insertion and Evolutionary Variation in Gene Regulation. | PNAS.” https://www.pnas.org/doi/abs/10.1073/pnas.93.18.9374.

4) Bronner, Marianne E., and Nicole M. LeDouarin. 2012. “Development and Evolution of the Neural Crest: An Overview.” *Developmental Biology*, Neural Crest, vol. 366 (1): 2–9. https://doi.org/10.1016/j.ydbio.2011.12.042.

5) Carneiro, Miguel, Carl-Johan Rubin, Federica Di Palma, et al. 2014. “Rabbit Genome Analysis Reveals a Polygenic Basis for Phenotypic Change during Domestication.” *Science* 345 (6200): 1074–79. https://doi.org/10.1126/science.1253714.

6) Chan, Yingguang Frank, Melissa E. Marks, Felicity C. Jones, et al. 2010. “Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a Pitx1 Enhancer.” *Science* 327 (5963): 302–5. https://doi.org/10.1126/science.1182213.

7) Chao, Jiannan, Furong Tang, and Lei Xu. 2022. “Developments in Algorithms for Sequence Alignment: A Review.” *Biomolecules* 12 (4): 546. https://doi.org/10.3390/biom12040546.

8) Claes, Peter, Jasmien Roosenboom, Julie D. White, et al. 2018. “Genome-Wide Mapping of Global-to-Local Genetic Effects on Human Facial Shape.” *Nature Genetics* 50 (3): 414–23. https://doi.org/10.1038/s41588-018-0057-4.

9) Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics* 25 (11): 1422–23. https://doi.org/10.1093/bioinformatics/btp163.

10) Cohen, Malkiel A., Shupei Zhang, Satyaki Sengupta, et al. 2020. “Formation of Human Neuroblastoma in Mouse-Human Neural Crest Chimeras.” *Cell Stem Cell* 26 (4): 579-592.e6. https://doi.org/10.1016/j.stem.2020.02.001.

11) Cramon-Taubadel, Noreen von. 2014. “The Microevolution of Modern Human Cranial Variation: Implications for Hominin and Primate Evolution.” *Annals of Human Biology* 41 (4): 323–35. https://doi.org/10.3109/03014460.2014.911350.

12) Crocker, Justin, Namiko Abe, Lucrezia Rinaldi, et al. 2015. “Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness.” *Cell* 160 (1): 191–203. https://doi.org/10.1016/j.cell.2014.11.041.

13) Cruz Walma, D. A., and K. M. Yamada. 2022. “Extracellular Matrix in Human Craniofacial Development.” *Journal of Dental Research* 101 (5): 495–504. https://doi.org/10.1177/00220345211052982.

14) Dahan, Sara, Aveksha Sharma, Klil Cohen, et al. 2021. “VEGFA’s Distal Enhancer Regulates Its Alternative Splicing in CML.” *NAR Cancer* 3 (3): zcab029. https://doi.org/10.1093/narcan/zcab029.

15) D’Souza, Rena N., L.-Bruno Ruest, Robert J. Hinton, and Kathy K. H. Svoboda. 2010. “Development of the Craniofacial Complex.” In *Bone and Development*, edited by Felix Bronner, Mary C. Farach-Carson, and H. I. (Trudy) Roach. Springer London. https://doi.org/10.1007/978-1-84882-822-3_10.

16) Duband, Jean-Loup, and Jean Paul Thiery. 1987. “Distribution of Laminin and Collagens during Avian Neural Crest Development.” *Development* 101 (3): 461–78. https://doi.org/10.1242/dev.101.3.461.

17) Durkin, Marian E., Miriam R. Avner, Chang-Goo Huh, Bao-Zhu Yuan, Snorri S. Thorgeirsson, and Nicholas C. Popescu. 2005. “DLC-1, a Rho GTPase-activating Protein with Tumor Suppressor Function, Is Essential for Embryonic Development.” *FEBS Letters* 579 (5): 1191–96. https://doi.org/10.1016/j.febslet.2004.12.090.

18) Fish, Jennifer L., and Richard A. Schneider. 2014. “Assessing Species-Specific Contributions To Craniofacial Development Using Quail-Duck Chimeras.” *Journal of Visualized Experiments : JoVE*, no. 87 (May): 51534. https://doi.org/10.3791/51534.

19) Frankel, Nicolás, Gregory K. Davis, Diego Vargas, Shu Wang, François Payre, and David L. Stern. 2010. “Phenotypic Robustness Conferred by Apparently Redundant Transcriptional Enhancers.” *Nature* 466 (7305): 490–93. https://doi.org/10.1038/nature09158.

20) Fulco, Charles P., Joseph Nasser, Thouis R. Jones, et al. 2019. “Activity-by-Contact Model of Enhancer–Promoter Regulation from Thousands of CRISPR Perturbations.” *Nature Genetics* 51 (12): 1664–69. https://doi.org/10.1038/s41588-019-0538-0.

21) Genereux, Diane P., Aitor Serres, Joel Armstrong, et al. 2020. “A Comparative Genomics Multitool for Scientific Discovery and Conservation.” *Nature* 587 (7833): 240–45. https://doi.org/10.1038/s41586-020-2876-6.

22) Gómez-Robles, Aida, Amy Drennan, Maricci Basa, and Alfie Gleeson. 2025. “Accelerated Evolution Increased Craniofacial Divergence between Humans and Great Apes.” *Proceedings of the Royal Society B: Biological Sciences* 292 (2057): 20251507. https://doi.org/10.1098/rspb.2025.1507.

23) Goovaerts, Seppe, Jay Devine, Nina Claessens, et al. 2026. “The Craniofacial Shape of Modern Humans Embodies Genomic Signatures of Evolution, Diversity, and Clinical Conditions.” *bioRxiv: The Preprint Server for Biology*, January 3, 2025.12.29.696848. https://doi.org/10.64898/2025.12.29.696848.

24) Grant, Charles E., Timothy L. Bailey, and William Stafford Noble. 2011. “FIMO: Scanning for Occurrences of a given Motif.” *Bioinformatics* 27 (7): 1017–18. https://doi.org/10.1093/bioinformatics/btr064.

25) Hall, Brian. 2000. “The Neural Crest as a Fourth Germ Layer and Vertebrates as Quadroblastic Not Triploblastic.” *Evolution & Development* 2 (February): 3–5. https://doi.org/10.1046/j.1525-142x.2000.00032.x.

26) Hanken, James, and Peter Thorogood. 1993. “Evolution and Development of the Vertebrate Skull: The Role of Pattern Formation.” *Trends in Ecology & Evolution* 8 (1): 9–15. https://doi.org/10.1016/0169-5347(93)90124-8.

27) Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, et al. 2020. “Array Programming with NumPy.” *Nature* 585 (7825): 357–62. https://doi.org/10.1038/s41586-020-2649-2.

28) Hayden, Eric J., Evandro Ferrada, and Andreas Wagner. 2011. “Cryptic Genetic Variation Promotes Rapid Evolutionary Adaptation in an RNA Enzyme.” *Nature* 474 (7349): 92–95. https://doi.org/10.1038/nature10083.

29) Hinrichs, A. S. 2006. “The UCSC Genome Browser Database: Update 2006.” *Nucleic Acids Research* 34 (90001): D590–98. https://doi.org/10.1093/nar/gkj144.

30) Hu, Diane, and Ralph S. Marcucio. 2009. “Unique Organization of the Frontonasal Ectodermal Zone in Birds and Mammals.” *Developmental Biology* 325 (1): 200–210. https://doi.org/10.1016/j.ydbio.2008.10.026.

31) Hu, Na, Pablo H. Strobl-Mazzulla, and Marianne E. Bronner. 2014. “Epigenetic Regulation in Neural Crest Development.” *Developmental Biology* 396 (2): 159–68. https://doi.org/10.1016/j.ydbio.2014.09.034.

32) Hunter, John D. 2007. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

33) Kadener, Sebastián, Juan Pablo Fededa, Michael Rosbash, and Alberto R. Kornblihtt. 2002. “Regulation of Alternative Splicing by a Transcriptional Enhancer through RNA Pol II Elongation.” *Proceedings of the National Academy of Sciences* 99 (12): 8185–90. https://doi.org/10.1073/pnas.122246099.

34) Kaucka, Marketa, Julian Petersen, Marketa Tesarova, et al. 2018. “Signals from the Brain and Olfactory Epithelium Control Shaping of the Mammalian Nasal Capsule Cartilage.” *eLife* 7 (June): e34465. https://doi.org/10.7554/eLife.34465.

35) King, Mary-Claire, and A. C. Wilson. 1975. “Evolution at Two Levels in Humans and Chimpanzees.” *Science* 188 (4184): 107–16. https://doi.org/10.1126/science.1090005.

36) Klein, Jason C., Aidan Keith, Vikram Agarwal, Timothy Durham, and Jay Shendure. 2018. “Functional Characterization of Enhancer Evolution in the Primate Lineage.” *Genome Biology* 19 (1): 99. https://doi.org/10.1186/s13059-018-1473-6.

37) Kleinjan, Dirk A., Ruth M. Bancewicz, Philippe Gautier, et al. 2008. “Subfunctionalization of Duplicated Zebrafish Pax6 Genes by Cis-Regulatory Divergence.” *PLOS Genetics* 4 (2): e29. https://doi.org/10.1371/journal.pgen.0040029.

38) Le Douarin, Nicole M., Sophie Creuzet, Gérard Couly, and Elisabeth Dupin. 2004. "Neural Crest Cell Plasticity and Its Limits." *Development* 131 (19): 4637–50. https://doi.org/10.1242/dev.01350.

39) Li, Yongqing, Dan Yang, Yan Bai, et al. 2008. "ZNF418, a Novel Human KRAB/C2H2 Zinc Finger Protein, Suppresses MAPK Signaling Pathway." *Molecular and Cellular Biochemistry* 310 (1): 141–51. https://doi.org/10.1007/s11010-007-9674-4.

40) Lim, Fabian, Joe J. Solvason, Genevieve E. Ryan, et al. 2024. "Affinity-Optimizing Enhancer Variants Disrupt Development." *Nature* 626 (7997): 151–59. https://doi.org/10.1038/s41586-023-06922-8.

41) Liu, Chenxing, Myoung Keun Lee, Sahin Naqvi, et al. 2021. "Genome Scans of Facial Features in East Africans and Cross-Population Comparisons Reveal Novel Associations." *PLOS Genetics* 17 (8): e1009695. https://doi.org/10.1371/journal.pgen.1009695.

42) Long, Hannah K., Sara L. Prescott, and Joanna Wysocka. 2016. "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution." *Cell* 167 (5): 1170–87. https://doi.org/10.1016/j.cell.2016.09.018.

43) Maurano, Matthew T., Richard Humbert, Eric Rynes, et al. 2012. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA." *Science* 337 (6099): 1190–95. https://doi.org/10.1126/science.1222794.

44) McCauley, David W. 2008. "SoxE, Type II Collagen, and Evolution of the Chondrogenic Neural Crest." *Zoological Science* 25 (10): 982–89. https://doi.org/10.2108/zsj.25.982.

45) McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." 56–61. https://doi.org/10.25080/Majora-92bf1922-00a.

46) McLeay, Robert C., and Timothy L. Bailey. 2010. "Motif Enrichment Analysis: A Unified Framework and an Evaluation on ChIP Data." *BMC Bioinformatics* 11 (1): 165. https://doi.org/10.1186/1471-2105-11-165.

47) Mitsiadis, Thimios A., Javier Caton, and Martyn Cobourne. 2006. "Waking-up the Sleeping Beauty: Recovery of the Ancestral Bird Odontogenic Program." *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 306B (3): 227–33. https://doi.org/10.1002/jez.b.21094.

48) Mostafavi, Hakhamanesh, Jeffrey P. Spence, Sahin Naqvi, and Jonathan K. Pritchard. 2023. "Systematic Differences in Discovery of Genetic Effects on Gene Expression and Complex Traits." *Nature Genetics* 55 (11): 1866–75. https://doi.org/10.1038/s41588-023-01529-1.

49) Ogbunugafor, C. B., R. C. McBride, and P. E. Turner. 2009. "Predicting Virus Evolution: The Relationship between Genetic Robustness and Evolvability of Thermotolerance." *Cold Spring Harbor Symposia on Quantitative Biology* 74 (January): 109–18. https://doi.org/10.1101/sqb.2009.74.023.

50) Oka, Kyoko, Masaki J. Honda, Eichi Tsuruga, Yuji Hatakeyama, Keitaro Isokawa, and Yoshihiko Sawa. 2012. “Roles of Collagen and Periostin Expression by Cranial Neural Crest Cells during Soft Palate Development.” *Journal of Histochemistry & Cytochemistry* 60 (1): 57–68. https://doi.org/10.1369/0022155411427059.

51) Osterwalder, Marco, Iros Barozzi, Virginie Tissières, et al. 2018. “Enhancer Redundancy Provides Phenotypic Robustness in Mammalian Development.” *Nature* 554 (7691): 239–43. https://doi.org/10.1038/nature25461.

52) Payne, Joshua L., and Andreas Wagner. 2015. “Mechanisms of Mutational Robustness in Transcriptional Regulation.” *Frontiers in Genetics* 6 (October). https://doi.org/10.3389/fgene.2015.00322.

53) Perris, Roberto, Josef Syfrig, Mats Paulsson, and Marianne Bronner-Fraser. 1993. “Molecular Mechanisms of Neural Crest Cell Attachment and Migration on Types I and IV Collagen.” *Journal of Cell Science* 106 (4): 1357–68. https://doi.org/10.1242/jcs.106.4.1357.

54) Prabhakar, Shyam, Axel Visel, Jennifer A. Akiyama, et al. 2008. “Human-Specific Gain of Function in a Developmental Enhancer.” *Science* 321 (5894): 1346–50. https://doi.org/10.1126/science.1159974.

55) Prescott, Sara L., Rajini Srinivasan, Maria Carolina Marchetto, et al. 2015. “Enhancer Divergence and Cis-Regulatory Evolution in the Human and Chimp Neural Crest.” *Cell* 163 (1): 68–83. https://doi.org/10.1016/j.cell.2015.08.036.

56) Qian, Wei, Manfei Zhang, Kaiwen Wan, et al. 2022. “Genetic Evidence for Facial Variation Being a Composite Phenotype of Cranial Variation and Facial Soft Tissue Thickness.” *Journal of Genetics and Genomics* 49 (10): 934–42. https://doi.org/10.1016/j.jgg.2022.02.020.

57) Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

58) Ragueneau, Eliot, Chuqiao Gong, Pierre Sinquin, et al. 2026. “The Reactome Knowledgebase 2026.” *Nucleic Acids Research* 54 (D1): D673–81. https://doi.org/10.1093/nar/gkaf1223.

59) Rebeiz, Mark, Nick Jikomes, Victoria A. Kassner, and Sean B. Carroll. 2011. “Evolutionary Origin of a Novel Gene Expression Pattern through Co-Option of the Latent Activities of Existing Regulatory Sequences.” *Proceedings of the National Academy of Sciences* 108 (25): 10036–43. https://doi.org/10.1073/pnas.1105937108.

60) Rebeiz, Mark, and Miltos Tsiantis. 2017. “Enhancer Evolution and the Origins of Morphological Novelty.” *Current Opinion in Genetics & Development*, Developmental mechanisms, patterning and evolution, vol. 45 (August): 115–23. https://doi.org/10.1016/j.gde.2017.04.006.

61) Rubio, Andrew O., and Kyle Summers. 2022. “Neural Crest Cell Genes and the Domestication Syndrome: A Comparative Analysis of Selection.” *PLOS ONE* 17 (2): e0263830. https://doi.org/10.1371/journal.pone.0263830.

62) Schneider, Richard A. 2018. “Neural Crest and the Origin of Species-specific Pattern.” *Genesis (New York, N.y. : 2000)* 56 (6–7): e23219. https://doi.org/10.1002/dvg.23219.

63) Shaffer, John R., Ekaterina Orlova, Myoung Keun Lee, et al. 2016a. “Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology.” *PLOS Genetics* 12 (8): e1006149. https://doi.org/10.1371/journal.pgen.1006149.

64) Shaffer, John R., Ekaterina Orlova, Myoung Keun Lee, et al. 2016b. “Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology.” *PLOS Genetics* 12 (8): e1006149. https://doi.org/10.1371/journal.pgen.1006149.

65) Simões-Costa, Marcos S., Sonja J. McKeown, Joanne Tan-Cabugao, Tatjana Sauka-Spengler, and Marianne E. Bronner. 2012. “Dynamic and Differential Regulation of Stem Cell Factor FoxD3 in the Neural Crest Is Encrypted in the Genome.” *PLOS Genetics* 8 (12): e1003142. https://doi.org/10.1371/journal.pgen.1003142.

66) Singh, Sonia R., Moritz Meyer-Jens, Erda Alizoti, et al. 2021. “A High-Throughput Screening Identifies ZNF418 as a Novel Regulator of the Ubiquitin-Proteasome System and Autophagy-Lysosomal Pathway.” *Autophagy* 17 (10): 3124–39. https://doi.org/10.1080/15548627.2020.1856493.

67) Soldatov, Ruslan, Marketa Kaucka, Maria Eleni Kastriti, et al. 2019. “Spatiotemporal Structure of Cell Fate Decisions in Murine Neural Crest.” *Science* 364 (6444): eaas9536. https://doi.org/10.1126/science.aas9536.

68) Sunadome, Kazunori, Alek G. Erickson, Delf Kah, et al. 2023. “Directionality of Developing Skeletal Muscles Is Set by Mechanical Forces.” *Nature Communications* 14 (1): 3060. https://doi.org/10.1038/s41467-023-38647-7.

69) Szklarczyk, Damian, Rebecca Kirsch, Mikaela Koutrouli, et al. 2023. “The STRING Database in 2023: Protein–Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest.” *Nucleic Acids Research* 51 (D1): D638–46. https://doi.org/10.1093/nar/gkac1000.

70) Tucker, Abigail S., and Andrew Lumsden. 2004. “Neural Crest Cells Provide Species-Specific Patterning Information in the Developing Branchial Skeleton.” *Evolution & Development* 6 (1): 32–40. https://doi.org/10.1111/j.1525-142X.2004.04004.x.

71) Virshup, Isaac, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. 2024. “Anndata: Access and Store Annotated Datamatrices.” *Journal of Open Source Software* 9 (101): 4371. https://doi.org/10.21105/joss.04371.

72) Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, et al. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 261–72. https://doi.org/10.1038/s41592-019-0686-2.

73) Vorontsov, Ilya E., Irina A. Eliseeva, Arsenii Zinkevich, et al. 2024. “HOCOMOCO in 2024: A Rebuild of the Curated Collection of Binding Models for Human and

Mouse Transcription Factors." *Nucleic Acids Research* 52 (D1): D154–63. https://doi.org/10.1093/nar/gkad1077.

74) Wheeler, M. 2003. "Sarcoglycans in Vascular Smooth and Striated Muscle." *Trends in Cardiovascular Medicine* 13 (6): 238–43. https://doi.org/10.1016/S1050-1738(03)00101-4.

75) White, Julie D., Karlijne Indencleef, Sahin Naqvi, et al. 2021. "Insights into the Genetic Architecture of the Human Face." *Nature Genetics* 53 (1): 45–53. https://doi.org/10.1038/s41588-020-00741-7.

76) Wolf, F. Alexander. 2018. "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis | Genome Biology | Springer Nature Link." February 6. https://link.springer.com/article/10.1186/s13059-017-1382-0.

77) Woolfe, Adam, Martin Goodson, Debbie K. Goode, et al. 2004. "Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development." *PLOS Biology* 3 (1): e7. https://doi.org/10.1371/journal.pbio.0030007.

78) Zanella, Matteo, Alessandro Vitriolo, Alejandro Andirko, et al. 2019. "Dosage Analysis of the 7q11.23 Williams Region Identifies BAZ1B as a Major Human Gene Patterning the Modern Human Face and Underlying Self-Domestication." *Science Advances* 5 (12): eaaw7908. https://doi.org/10.1126/sciadv.aaw7908.

Supplementary materials

Supplement 1. Scripts and notebooks

GitHub repository - https://github.com/Romanova-N/MSc_Craniofacial_NC_enhancers

Scripts		
1	1_1_MAF_per_enh.sh + 1_1_maf_subset.awk	Together, these scripts take genomic coordinates and list of species as an input and dissect the regions from bigMAF multiple sequence alignment of Zoonomia Project. Awk script is included in bash script.
2	1_2_add_maf_header.sh	Fix of MAF output files, just adding a comment line in the beginning of each MAF file
3	1_2_MAF2FASTA_per_enh.py	Format conversion from MAF to FASTA
4	5_2A_Motif_enrichment.sh	Script utilizing MENE AME command line tool to find motifs in query sequences. This version finds motif enrichment per gene

		comparing its specific enhancers with the rest.
5	5_3A_Global_control.sh	Similar as 5_2A, but all specific enhancer sequences are tested against a shuffled sequences as a background.
6	5_3A_Global_motif_enrichment.sh	Similar as 5_2A, but all specific enhancer sequences are tested against all non-specific, but related to the same genes.
7	5_1_Motif_search_HOCOMOC.smk/ 5_1_Motif_search_JASPAR.smk	Snakemake workflow for motif identification via MEME FIMO with different motif databases: HOCOMOCO or JASPAR.
8	5_1_Motif_cleanup.ipynb	Short notebook to get all TFs expressed in scRNA-seq dataset.
9	5_1_Motif_cleanup.py	It filters MEME FIMO output per each enhancer considering thresholds, TF presence in scRNA-seq and amplified TFs.
Jupyter Notebooks		
1	0_Curating_init_dataset.ipynb	It preprocesses initial multiome dataset to get enhancer information, GWAS dataset to process SNP coordinates and prepare gene set for gene enrichment background (step 2_1). It requires output files from scripts 1_*.
2	1_CREs_anndata.ipynb	One of the central notebooks where enhancers are categorized to specific/non-specific based on preservation calculation
3	2_1_STRING_visualisation.ipynb	It post-process STRING output dataset which was generated via online tool
4	2_2_CREs_and_expression.ipynb	It analyses general patterns of expression which genes associated with primate-specific enhancers demonstrate.
5	2_3_Dot_plots_per_gene.ipynb	It observe in more details per gene expression in different cell clusters and developmental stages.

6	3_1_GWAS_analysis.ipynb	It explores how integrated SNPs associated with craniofacial variation are distributed among regulatory elements.
7	4_Per_gene_enhancer_analysis.ipynb	It explores the gene regulatory architecture for genes possessing some number of lineage-specific enhancers.
8	5_1_TF_IDF_analysis.ipynb	It contains all calculations and visualisations regarding TF-IDF scoring of motif content. It requires output data from all scripts 5_1
9	5_2_Motif_enrichment_analysis.ipynb	It contains all calculations and visualisations regarding motif enrichments. It requires output data from all scripts 5_2 and 5_3.
10	MSA_visualisations	A short notebook which visualize all enhancer alignments via MSAViz.

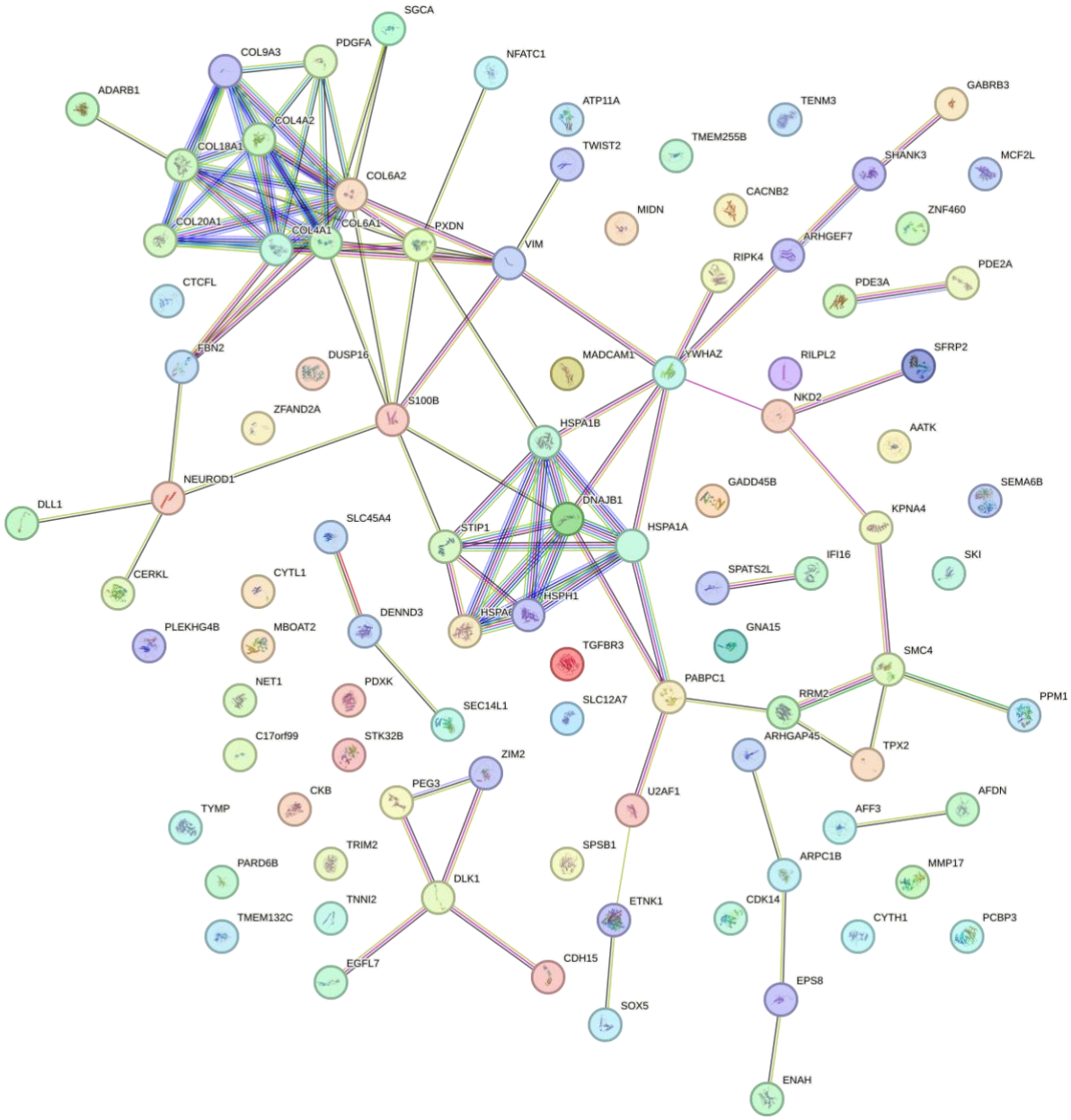
Supplement 2. 25-state ChromHMM model description

1_TssA = Active TSS; 2_PromU = Promoter Upstream TSS; 3_PromD1 = Promoter Downstream TSS 1; 4_PromD2 = Promoter Downstream TSS 2; 5_Tx5 = Transcribed -5' preferential; 6_Tx = Strong transcription; 7_Tx3 = Transcribed– 3' preferential; 8_TxWk = Weak transcription; 9_TxReg = Transcribed and regulatory (Prom/Enh); 10_TxEnh5 = Transcribed 5' preferential and Enh; 11_TxEnh3 = Transcribed 3' preferential and Enh; 12_TxEnhW = Transcribed and Weak Enhancer; 13_EnhA1 = Active Enhancer 1; 14_EnhA2 = Active Enhancer 2; 15_EnhAF = Active Enhancer Flank; 16_EnhW1 = Weak Enhancer 1; 17_EnhW2 = Weak Enhancer 2; 18_EnhAc = Primary H3K27ac possible Enhancer; 19_DNase = Primary DNase; 20_ZNF_Rpts = ZNF genes & repeats; 21_Het = Heterochromatin; 22_PromP = Poised Promoter; 23_PromBiv = Bivalent Promoter; 24_ReprPc = Repressed Polycomb; 25_Quies = Quiescent/Low.

More info: https://egg2.wustl.edu/roadmap/web_portal/imputed.html#chr_imp

Supplement 3. Full STRING output for genes associated with primate-specific enhancers

Full analysis: https://version-12-0.string-db.org/cgi/network?taskId=bSw9b0vIcEKO&sessionId=bGzMkXxn1fs4



Supplement 4. Genes with descriptions and links

Note: in the main text we don't separate group 2 and 3 and call them collectively "primate-specific enhancers".

Group 1 – Linked to human-specific enhancers
These enhancers has no aligned regions from other genomes

Gene	Function	Human atlas description

DLC1	GTP-binding proteins activation which participate in cytoskeletal changes	https://www.proteinatlas.org/ENSG00000164741-DLC1
KIF26B	Intercellular transport along microtubules	https://www.proteinatlas.org/ENSG00000162849-KIF26B
NLRP3	Regulation of inflammation, immune response, and apoptosis	https://www.proteinatlas.org/ENSG00000162711-NLRP3
PDZRN3	Assumed to function in differentiation of adipocytes, osteoblasts, and myoblasts, play a role in vascular morphogenesis	https://www.proteinatlas.org/ENSG00000121440-PDZRN3
SGCZ	Part of sarcoglycan complex located in membrane bridging the inner cytoskeleton and the extra-cellular matrix	https://www.proteinatlas.org/ENSG00000185053-SGCZ

Group 2 – Linked to primate-conservative enhancers

All primates have preservation level > 90% with up to 1 exception (either 7/8 or 8/8) and all other species have < 10% of preservation.

We interpret this group of enhancers as “emerged in primate lineage”.
“Gr” shows if gene nodes anyhow interlinked on the STRING bitmap.

N	Gr	Gene	Description	NCBI link
1	1	COL18A1	Collagen-family	https://www.ncbi.nlm.nih.gov/gene/80781
2		COL4A1	Collagen-family	https://www.ncbi.nlm.nih.gov/gene/1282
3		COL4A2	Collagen-family	https://www.ncbi.nlm.nih.gov/gene/1284
4		COL6A1	Collagen-family	https://www.ncbi.nlm.nih.gov/gene/1291
5		COL6A2	Collagen-family	https://www.ncbi.nlm.nih.gov/gene/1292
6		ADARB1	neuronal function, connected to COL18A1 performing angiogenesis	https://www.ncbi.nlm.nih.gov/gene/104
7		SGCA	muscle fiber membrane	https://www.ncbi.nlm.nih.gov/gene/6442
8		S100B	neurotrophic factor	https://www.ncbi.nlm.nih.gov/gene/6285
9		VIM	Making up cytoskeleton	https://www.ncbi.nlm.nih.gov/gene/7431

10		PXDN	enzyme in collagen formation	https://www.ncbi.nlm.nih.gov/gene/7837
11		NEUROD1	neuronal maturation, sensory organ development	https://www.ncbi.nlm.nih.gov/gene/4760
12		CERKL	regulates autophagy, acts as a cytoprotective factor, particularly in photoreceptor cells, protecting them from oxidative damage and apoptosis.	https://www.ncbi.nlm.nih.gov/gene/375298
13	2	NKD2	Has been shown promoting osteoblast and adipocytes differentiation	https://www.ncbi.nlm.nih.gov/gene/85409
14		KPNA4	seems to mediate some nuclear transport?	https://www.ncbi.nlm.nih.gov/gene/3840
15		SMC4	structural organization of chromosomes during mitosis	https://www.ncbi.nlm.nih.gov/gene/10051
16		PPM1L	apoptosis regulation	https://www.ncbi.nlm.nih.gov/gene/151742
17		RRM2	enzyme involved in cell-cycle progression	https://www.ncbi.nlm.nih.gov/gene/6241
18		TPX2	proper chromosome segregation in cell division	https://www.ncbi.nlm.nih.gov/gene/22974
19	3	SOX5	TF, neurogenesis (especially) and chondrogenesis	https://www.ncbi.nlm.nih.gov/gene/6660
20		ETNK1	enzyme essential for cell membrane synthesis	https://www.ncbi.nlm.nih.gov/gene/55500
21		U2AF1	encodes the small subunit which plays a critical role in both constitutive and enhancer-dependent RNA splicing by directly mediating interactions between the large subunit and proteins bound to the enhancers.	https://www.ncbi.nlm.nih.gov/gene/7307
22	4	AFF3	transcriptional activator, cell adhesion, mesoderm/ectoderm development, and lymphocyte development	https://www.ncbi.nlm.nih.gov/gene/3899
23		AFDN	signaling and organization of cell junctions during embryogenesis	https://www.ncbi.nlm.nih.gov/gene/4301
24	5	DENND3	endosome-to-lysosome transport, stress-induced autophagy. Mutations and deficiencies are linked to conditions like Hirschsprung disease	https://www.ncbi.nlm.nih.gov/gene/22898
25		SEC14L1	possible role of the gene product in an intracellular transport system	https://www.ncbi.nlm.nih.gov/gene/6397

26		CACNB2	calcium channels subunit in cardiac and neuronal tissue	https://www.ncbi.nlm.nih.gov/gene/783
27		CYTH1	leukocyte adhesion, migration, and the "inside-out" signaling necessary for immune response	https://www.ncbi.nlm.nih.gov/gene/9267
28		AATK	promotes neuronal differentiation, neurite extension, apoptosis	https://www.ncbi.nlm.nih.gov/gene/9625
29		TMEM255B	membrane transporter, detected in epithelial cells, no more info	https://www.ncbi.nlm.nih.gov/gene/348013
30		RILPL2	no clear function? associated with: tremor and Osteosclerotic Metaphyseal Dysplasia	https://www.ncbi.nlm.nih.gov/gene/196383
31		PLEKHG4B	enables actin cytoskeletal remodeling during epithelial cell–cell junction formation	https://doi.org/10.1242/jcs.249078
32		PDE3A	expressed in cardiac myocytes, vascular smooth muscle cells, platelets, and oocytes	https://www.ncbi.nlm.nih.gov/gene/5139
33		MCF2L	regulator of Rho/Rac signaling pathways, controlling cytoskeleton dynamics	https://www.ncbi.nlm.nih.gov/gene/23263
34		ARHGEF7	dendritic spine maturation and synapse structure	https://www.ncbi.nlm.nih.gov/gene/8874
35		TYMP	crucial for regulating nucleotide levels and maintaining mtDNA	https://www.ncbi.nlm.nih.gov/gene/1890
36		MADCAM1	binds to the integrin and L-selectin, enabling leukocyte adhesion to high endothelial venules (HEV) and directing lymphocyte migration to mucosal lymphoid tissues.	https://www.ncbi.nlm.nih.gov/gene/8174
37		DLK1	a key regulator of cell growth and differentiation	https://www.ncbi.nlm.nih.gov/gene/8788
38		GABRB3	subunit of the GABA, neurotransmitter system in the brain	https://www.ncbi.nlm.nih.gov/gene/2562
39		MIDN	mediate degradation of TFs: FOSB, EGR1, NR4A1, and IRF4	https://www.ncbi.nlm.nih.gov/gene/90007
40		MBOAT2	enzyme, catalyzes the transfer of an acyl group from an acyl-CoA to a lysophospholipid	https://www.ncbi.nlm.nih.gov/gene/129642
41		PARD6B	involved in asymmetrical cell division and cell polarization processes. Probably involved in	https://www.ncbi.nlm.nih.gov/gene/84612

			formation of epithelial tight junctions	
42		ZFAND2A	stress-induced, aiding protein degradation	https://www.ncbi.nlm.nih.gov/gene/90637
43		CKB	enzyme crucial for cellular energy homeostasis, particularly in the brain, heart, and smooth muscle	https://www.ncbi.nlm.nih.gov/gene/1152
44		TMEM132C	membrane protein, no certain info	https://www.ncbi.nlm.nih.gov/gene/92293
45		MAPK10	mediating neuronal death (apoptosis) in response to stress and is involved in controlling neuronal migration and differentiation.	https://www.ncbi.nlm.nih.gov/gene/92293
46		FGFR1OP	Centrosomal protein	https://www.ncbi.nlm.nih.gov/gene/11116
47		ATP11A	regulate membrane curvature, cell adhesion, and fusion. Somehow connected with eye/ear formation	https://www.ncbi.nlm.nih.gov/gene/23250
48		C17orf99	B cell-associated cytokine	https://www.ncbi.nlm.nih.gov/gene/100141515
49		SEMA6B	neurogenesis, axon guidance	https://www.ncbi.nlm.nih.gov/gene/10501
50		NFATC1	It is the "master regulator" of RANKL-induced osteoclast differentiation and activation.	https://www.ncbi.nlm.nih.gov/gene/4772
51		MEG3	long non-coding RNA, regulates cell proliferation, apoptosis, and angiogenesis, with its down-regulation or inactivation (often via promoter hypermethylation) observed in many cancers	https://www.ncbi.nlm.nih.gov/gene/55384
52		MEG8	long non-coding RNA, expressed from the maternal allele in skeletal muscle, and appears to be coordinately regulated with other imprinted genes in this region	https://www.ncbi.nlm.nih.gov/gene/79104

Group 3 – Linked to primate-specific enhancers
All primates have preservation level > 90% with up to 1 exception (either 7/8 or 8/8), at least half of other species have <10%, and it doesn't belong to group 2.

We interpret this group of enhancers as important for primates, but not unique for the group "Gr" shows if gene nodes anyhow interlinked on the STRING bitmap.

N	Gr	Gene	Description	NCBI link

1	1	HSPA6	Chaperon protein	https://www.ncbi.nlm.nih.gov/gene/3310
2		HSPA1B	Chaperon protein	https://www.ncbi.nlm.nih.gov/gene/3304
3		HSPA1A	Chaperon protein	https://www.ncbi.nlm.nih.gov/gene/3303
4		HSPH1	Chaperon protein	https://www.ncbi.nlm.nih.gov/gene/10808
5		STIP1	co-chaperone adaptor protein	https://www.ncbi.nlm.nih.gov/gene/10963
6		DNAJB1	a critical molecular chaperone protein that stimulates Hsp70 ATPase activity	https://www.ncbi.nlm.nih.gov/gene/3337
7		PABPC1	mRNA stability	https://www.ncbi.nlm.nih.gov/gene/26986
8		YWHAZ	neuronal development and migration 14-3-3 protein which is also have "chaperone-like" activity important in neural development	https://www.ncbi.nlm.nih.gov/gene/7534
9		RIPK4	a kinase crucial for epidermal differentiation, keratinocyte development, and barrier formation	https://www.ncbi.nlm.nih.gov/gene/54101
10	2	COL20A1	Collagen	https://www.ncbi.nlm.nih.gov/gene/57642
11		COL9A3	Collagen	https://www.ncbi.nlm.nih.gov/gene/1299
12		PDGFA	Protein presumably serving either as platelet-derived growth factors (PDGF) or vascular endothelial growth factors (VEGF)	https://www.ncbi.nlm.nih.gov/gene/5154
13	3	ARHGAP45	Contains a GTPase activator for the Rho-type GTPases (RhoGAP) domain that would be able to negatively regulate the actin cytoskeleton as well as cell spreading	https://www.ncbi.nlm.nih.gov/gene/23526
14		ARPC1B	actin polymerisation	https://www.ncbi.nlm.nih.gov/gene/10095
15		EPS8	actin-regulatory model, crucial for the development of auditory hair cells	https://www.ncbi.nlm.nih.gov/gene/2059

16		ENAH	actin-regulatory protein involved in controlling cell cytoskeleton dynamics, motility, adhesion, and morphology	https://www.ncbi.nlm.nih.gov/gene/55740
17	4	SPATS2L	Enables RNA binding activity	https://www.ncbi.nlm.nih.gov/gene/26010
18		IFI16	endothelial cells, lymphoid tissues, and myeloid cells. Acts as an innate immune sensor for DNA	https://www.ncbi.nlm.nih.gov/gene/3428
19	5	ZIM2	**ZIM2** and **PEG3** have a common promoter (but 2 different enhancers than?) and share some exons. Not well studied, but GO supports *DNA-binding transcription factor activity*	https://www.ncbi.nlm.nih.gov/gene/23619
20		PEG3		https://www.ncbi.nlm.nih.gov/gene/5178
21		SLC45A4	proton symporter activity	https://www.ncbi.nlm.nih.gov/gene/57210
22		C22orf34	?	https://www.ncbi.nlm.nih.gov/gene/348645
23		SHANK3	master scaffolding protein for synaptic function	https://www.ncbi.nlm.nih.gov/gene/85358
24		DLL1	neurogenesis, angiogenesis and notch-signalling	https://www.ncbi.nlm.nih.gov/gene/28514
25		H19	long non-coding RNA that is highly active during embryonic development but largely silent after birth	https://www.ncbi.nlm.nih.gov/gene/283120
26		LINC01572	long non-coding RNA	https://www.ncbi.nlm.nih.gov/gene/101927957
27		GNA15	involved in calcium-mediated signalling	https://www.ncbi.nlm.nih.gov/gene/2769
28		CYTL1	secreted protein involved in cartilage development, homeostasis, and chondrogenesis	https://www.ncbi.nlm.nih.gov/gene/54360
29		SFRP2	involved in cell fate decisions?	https://www.ncbi.nlm.nih.gov/gene/6423
30		CDK14	encodes a serine/threonine kinase involved in cell cycle regulation (G2/M transition) and signaling pathways	https://www.ncbi.nlm.nih.gov/gene/5218

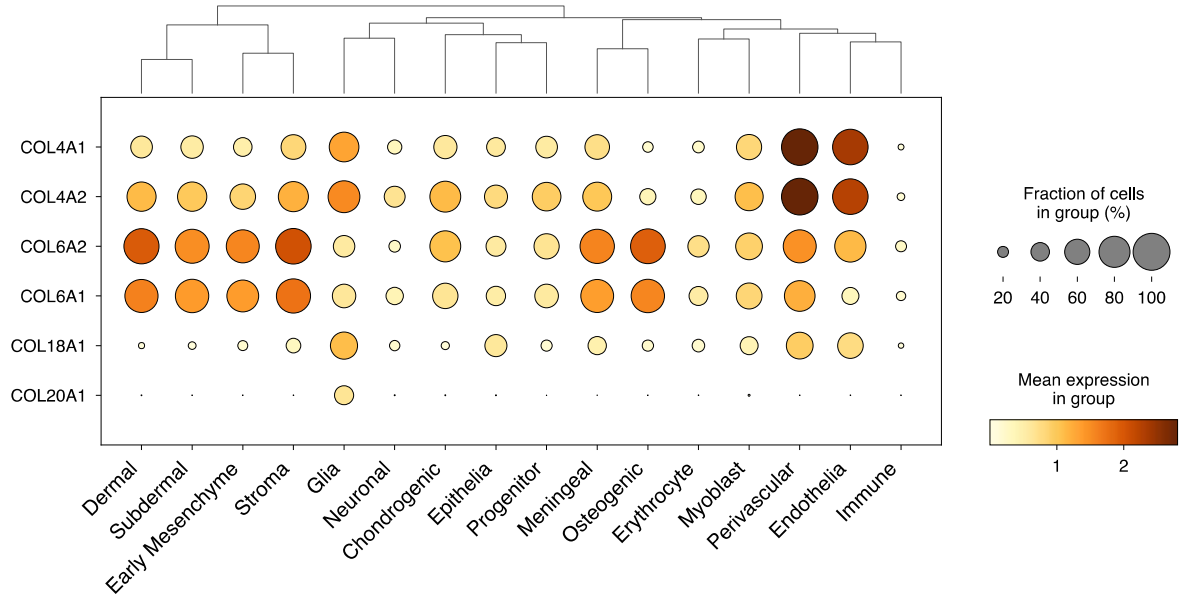
31		GADD45B	stress-induced gene, cell regulation of growth and apoptosis	https://www.ncbi.nlm.nih.gov/gene/4616
32		ZNF460	TF, no clear function, associated with cystathioninuria	https://www.ncbi.nlm.nih.gov/gene/10794
33		CDH15	involved in muscle development mainly and also neurodevelopment	https://www.ncbi.nlm.nih.gov/gene/1013
34		TRIM2	protein degradation in nervous system	https://www.ncbi.nlm.nih.gov/gene/23321
35		FBN2	microfibril formation	https://www.ncbi.nlm.nih.gov/gene/2201
36		PDXK	enzyme, phosphorylates vitamin B6	https://www.ncbi.nlm.nih.gov/gene/8566
37		TNNI2	skeletal muscle protein, may play a role in regulation of smooth muscle function.	https://www.ncbi.nlm.nih.gov/gene/7136
38		NET1	interacts with RhoA within the cell nucleus	https://www.ncbi.nlm.nih.gov/gene/10276
39		TGFBR3	cell proliferation, migration, differentiation, and apoptosis. Absence causes lethal cardiac and vascular abnormalities	https://www.ncbi.nlm.nih.gov/gene/7049
40		TENM3	crucial for neural development, particularly in establishing connectivity in the visual pathway and hippocampal networks.	https://www.ncbi.nlm.nih.gov/gene/55714
41		STK32B	role in cell signaling and proliferation. Associated with Ellis-van Creveld syndrome, an autosomal recessive skeletal dysplasia	https://www.ncbi.nlm.nih.gov/gene/55351
42		SLC12A7	expressed in inner ear and certain epithelial cells, facilitate ovement of potassium and chloride ions	https://www.ncbi.nlm.nih.gov/gene/10723
43		PCBP3	brain, retina expression, post-transcriptional effects by binding to mRNA	https://www.ncbi.nlm.nih.gov/gene/54039

44		EGFL7	vascular development by guiding endothelial cells. Antagonist of Notch signaling and inhibits smooth muscle cell migration.	https://www.ncbi.nlm.nih.gov/gene/51162
45		SKI	regulating cell growth, differentiation, and development. Associated with SDS - condition characterized by craniosynostosis (premature skull bone fusion), distinct facial features, skeletal abnormalities, and intellectual disability	https://www.ncbi.nlm.nih.gov/gene/6497
46		SPSB1	protein ubiquitination, low tissue specificity	https://www.ncbi.nlm.nih.gov/gene/80176
47		MMP17	breakdown of extracellular matrix (incl in embryonic dev), expressed at the cell surface	https://www.ncbi.nlm.nih.gov/gene/4326
48		PDE2A	enzyme, involved in signaling pathways in the brain and heart	https://www.ncbi.nlm.nih.gov/gene/5138
49		TWIST2	inhibiting osteoblast maturation and maintaining cells in a preosteoblast state	https://www.ncbi.nlm.nih.gov/gene/117581
50		LINC01237	long non-coding RNA	https://www.ncbi.nlm.nih.gov/gene/?term=LINC01237

Some genes appeared in both groups 2 and 3, which means that they have both primate conservative and primate-specific enhancers. It is:

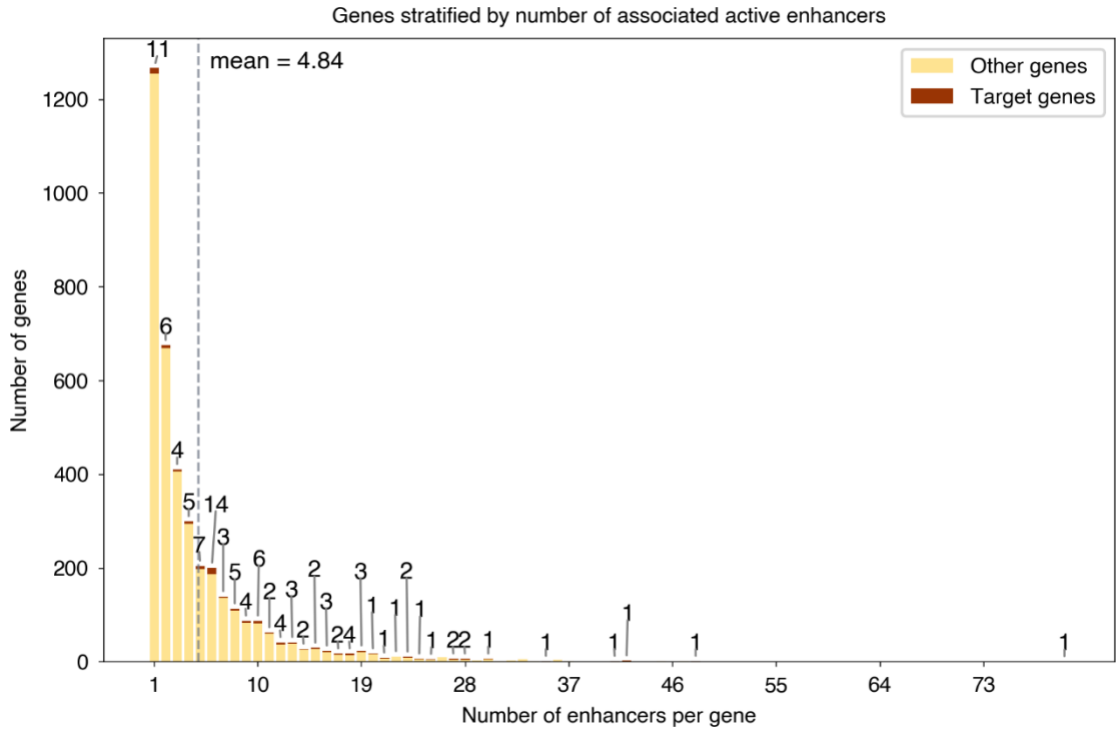
1) AATK
2) ADARB1
3) AFDN
4) ARHGEF7
5) ATP11A
6) COL18A1
7) COL4A1
8) COL6A1
9) COL6A2
10) MCF2L
11) MIDN
12) NFATC1
13) S100B
14) TMEM132C
15) TMEM255B
16) U2AF1

Supplement 5. Collagen expression in cell clusters



Supplement 6. Number of enhancers per gene

6.1. Full dataset



6.2. Genes possessing primate-specific enhancers

Gene	Total N of enhancers	Not primate-specific	Primate-specific
TWIST2	80	77	3
COL6A2	48	43	5
MEG8	42	40	2

MEG3	41	39	2
PDZRN3	35	35	0
HSPH1	30	29	1
COL6A1	28	25	3
DLK1	28	27	1
COL4A1	27	25	2
KIF26B	27	27	0
COL18A1	25	23	2
TMEM132C	24	21	3
MIDN	23	21	2
CKB	23	22	1
COL4A2	22	21	1
IFI16	21	20	1
TGFBR3	20	19	1
SOX5	19	18	1
VIM	19	18	1
NFATC1	19	16	3
HSPA1B	18	17	1
H19	18	17	1
FBN2	18	17	1
CYTH1	18	17	1
CDH15	17	16	1
COL9A3	17	16	1
U2AF1	16	13	3
SKI	16	15	1
DNAJB1	16	15	1
HSPA1A	15	14	1
MCF2L	15	13	2
EGFL7	14	13	1
PDGFA	14	13	1
TENM3	13	12	1
S100B	13	11	2
SEMA6B	13	12	1
ATP11A	12	9	3
ARHGEF7	12	10	2
AFF3	12	11	1
SEC14L1	12	11	1
MBOAT2	11	10	1
GADD45B	11	10	1
TMEM255B	10	7	3
C22orf34	10	9	1
DLC1	10	10	0
ADARB1	10	8	2
DENND3	10	9	1

SPATS2L	10	9	1
CDK14	9	8	1
PEG3	9	8	1
PABPC1	9	8	1
TRIM2	9	8	1
AFDN	8	4	4
NET1	8	7	1
RILPL2	8	7	1
C17orf99	8	7	1
SPSB1	8	7	1
SGCA	7	6	1
PARD6B	7	6	1
LINC01572	7	6	1
SFRP2	6	5	1
MAPK10	6	5	1
ETNK1	6	5	1
ZFAND2A	6	5	1
PCBP3	6	5	1
SLC45A4	6	5	1
YWHAZ	6	5	1
ENAH	6	5	1
DLL1	6	5	1
GNA15	6	5	1
TPX2	6	5	1
STK32B	6	5	1
STIP1	6	5	1
AATK	6	4	2
COL20A1	5	4	1
PDE2A	5	4	1
EPS8	5	4	1
CYTL1	5	4	1
SHANK3	5	4	1
ARPC1B	5	4	1
RIPK4	5	4	1
PPM1L	4	3	1
RRM2	4	3	1
LINC01237	4	3	1
NKD2	4	3	1
PDE3A	4	3	1
CACNB2	3	2	1
MMP17	3	2	1
PDXK	3	2	1
TNNI2	3	2	1
NLRP3	2	2	0

PLEKHG4B	2	1	1
SGCZ	2	2	0
FGFR1OP	2	1	1
ZIM2	2	1	1
SMC4	2	1	1
TYMP	1	0	1
NEUROD1	1	0	1
CERKL	1	0	1
HSPA6	1	0	1
ARHGAP45	1	0	1
KPNA4	1	0	1
SLC12A7	1	0	1
MADCAM1	1	0	1
ZNF460	1	0	1
PXDN	1	0	1
GABRB3	1	0	1