Innovation in Duplication: Structural Diversity and Regulatory Control of Human Genes: *TBC1D3*

Xavi Guitart

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee

Evan Eichler, Chair

Harmit Malik

Maitreya Dunham

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

ABSTRACT

Innovation in Duplication: Structural Diversity and Regulatory Control of Human Genes: *TBC1D3*

Xavi Guitart

Chair of the Supervisory Committee:
Evan Eichler
Genome Sciences

Segmental duplications (SDs) are a major source of genomic innovation, responsible for the genomic instability that accelerates novel gene function but also causes disease. Despite their importance, SDs have historically remained difficult to study due to their high sequence identity and copy number polymorphism, making them all but impossible to resolve with short-read sequencing and assembly. Recent advances in long-read sequencing and *de novo* genome assembly have made it possible to resolve these regions at haplotype resolution, enabling systematic investigation of complex duplicated gene families. This thesis leverages these technological advances to study the evolution, structural diversity, and regulation of *TBC1D3*, a primate-specific SD gene family implicated in neuronal progenitor proliferation, cortical expansion, and cancer.

TBC1D3 is a young and highly duplicated gene family dispersed across chromosome 17, with the majority of paralogs embedded in two large SD clusters at 17q12. Prior functional studies demonstrated that *TBC1D3* promotes cellular proliferation in both cancer and neurodevelopmental contexts, yet it remained unclear how a gene family with extreme copy

number variation could contribute to tightly regulated developmental processes. In this work, I address this paradox by integrating long-read genome assemblies, comparative primate genomics, population-scale human variation, and paralog-resolved transcriptomics.

In Chapter 2, I reconstruct the evolutionary history and human diversity of *TBC1D3* using haplotype-resolved assemblies from 69 human haplotypes and 11 nonhuman primate species. I show that *TBC1D3* independently expanded in at least five primate lineages and that humans experienced a recent expansion approximately 2–3 million years ago. Human haplotypes exhibit extraordinary structural diversity, differing by up to ~1 Mbp and more than 20 gene copies, making *TBC1D3* one of the most structurally variable gene families in the human genome. Despite this variability, signatures of positive selection are detected along the African ape lineage, and I show that all human-expressed copies share a derived, human-specific modification of the protein C terminus, suggesting functional divergence during recent human evolution. Using a pangenomic and phylogenetic framework, I define distinct paralog groups and demonstrate that *TBC1D3* expression is overwhelmingly restricted to a single paralog group located at the telomeric end of cluster 2.

Chapter 3 investigates the regulatory basis of this striking paralog-specific expression. I demonstrate that *TBC1D3* expression in human neural contexts is driven by a position-effect mechanism in which a fixed, copy number-constrained promoter derived from the neighboring gene *NPEPPSP1* has been duplicated and fused upstream of a specific *TBC1D3* paralog. Using comparative epigenomics, long-read transcriptomics, and neuronal differentiation models, I show that this *NPEPPSP1–TBC1D3* fusion creates a dominant regulatory architecture that restricts transcription to a single copy despite extensive underlying copy number variation. This

mechanism provides a parsimonious explanation for how *TBC1D3* expression and function may remain stable while the surrounding gene family continues to diversify structurally.

In Chapter 4, I describe experimental efforts to interrogate the functional consequences of human-specific modifications to *TBC1D3*, including the derived C-terminal extension. Although these experiments were not ultimately successful, their outcomes inform hypotheses regarding protein localization, posttranslational regulation, and context-dependent function that motivate future work.

This thesis establishes a generalizable framework for studying complex SD gene families by integrating haplotype-resolved assemblies, evolutionary analysis, and paralog-aware regulatory interrogation. The findings reveal how SDs contribute to gene family expansion, regulatory innovation, and protein evolution, allowing rapid structural diversification. More broadly, this work demonstrates how long-read genomics enables direct investigation of genomic regions that have played a disproportionate role in human evolution and disease yet have remained largely inaccessible until now.

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGEMENTS

I have wanted to be a scientist since I was a small child – I remember wearing a white lab coat and fake glasses and playing with kids slime and magnet sets. My parents have always been supportive of these endeavors, stimulating my mind and encouraging me to explore the world. I am forever grateful to them for this.

It wasn't until my freshman year physics class, however, that I really experienced the scientific method, which I fell in love with. To be honest, the utility of math, especially trigonometry and precalculus, evaded me until Mr. Pasquesi's physics class. His teachings of mechanics, electricity, and "wavicles" inspired me and directed me on the STEM path. In a similar vein, I am grateful to my college computer science professor, Dr. Gerald Roth, who introduced me to the beauty and fun of programming. His training drove me from engineering toward data science, which brought me to the world of genomics.

I would not have found myself in genomics if it were not for the guidance of Dr. Cynthia Dunbar, who included me into her lab as a post-baccalaureate fellow. I am especially thankful for Drs. Joy Wu and Fanny Xing, whose endless patience and careful instruction in molecular biology and experimental design taught me so much.

I want to especially thank Dr. Stefan Cordes. Stefan was a clinical fellow at the NIH during my fellowship. He not only helped train me in bioinformatics, he also showed me how exciting genomics can be, included me in experimental design exercises, and kept the spark of inquiry and excitement alive when I felt lost in the complexity of biology.

I thank my advisor Evan, for everything he has done for me these last six years. Evan took me on during the start of the COVID pandemic, when so much was up in the air. More than any advisor I've had, Evan gives his trainees the time and guidance without protest or hesitation. He is passionate about his work – advancing science and especially training the next generations of scientists. I am so grateful for the time he has dedicated to me. He has without question, made me a better scientist and a better person.

I thank Jessie Brunner, Alex Pollen, and members of the Pollen lab, who helped guided me during a month-long visit to learn more about the details of *in vitro* experimentation at UCSF. It was exciting to see the work firsthand and the group was so welcoming of me.

There are several members of the Eichler lab who took the time to teach me during my training that I would like to thank: Mitchell Vollger, PingHsun (Benson) Hsieh, Yafei Mao, Willliam Harvey, David Porubsky, Phil Dishuck, Luyao Ren, and DongAhn Yoo. I am especially thankful for our technicians who helped gather and sequence the data I used every day: Kendra Hoekzema, Katy Munson, Marcello Ayllon, and Kaitlyn Sun. Thanks also to the computational team that was always available to help me: Youngjun Kwon, Isaac Wong, Julie Wertz, and Nidhi Koudinya.

Tonia Brown and Zoe Poyen have been super-star lab managers, helping me every day with manuscripts, grant details, department operations, and much, much more. Thanks, you guys.

In more ways than I can describe, the entire lab has been such a positive and fun experience. From random conversations at lunch and at the coffee machine, to exciting lab outings, thanks for the company and stimulating environment: Michelle Noyes, Taylor Real, Kumara Mastrorosa, Lizzie Plender, Yang Sui, Jiadong Lin, Lingbin Ni, Mihir Trivedi, and Tara Mack.

CHAPTER 1. INTRODUCTION

1.1 EVOLUTION BY SEGMENTAL DUPLICATION

While studying bilaterian diversification during the Cambrian explosion, evolutionary biologists observed that anatomical innovation often emerges from a repeated body plan, where redundancy permits individual segments to evolve new functions without compromising overall viability (Holland, 2012). This anatomical evolutionary characteristic has its basis in an eloquent genetic analog: gene duplications introduce functional redundancy, reduce selective pressure, and create evolutionary "space" for diversification. This genetic principle was first formalized by geneticist Susumu Ohno, who recognized gene duplication as a major driver of evolutionary novelty (Ohno, 1970). Building on this foundation, Ed Lewis and others uncovered deeply conserved clusters of duplicated homeotic genes—*Hox* genes—that pattern the body plans of metazoans (Lewis, 1978). These genes exemplify how duplication, followed by divergence, can underlie the evolution of entirely new anatomical structures and functions.

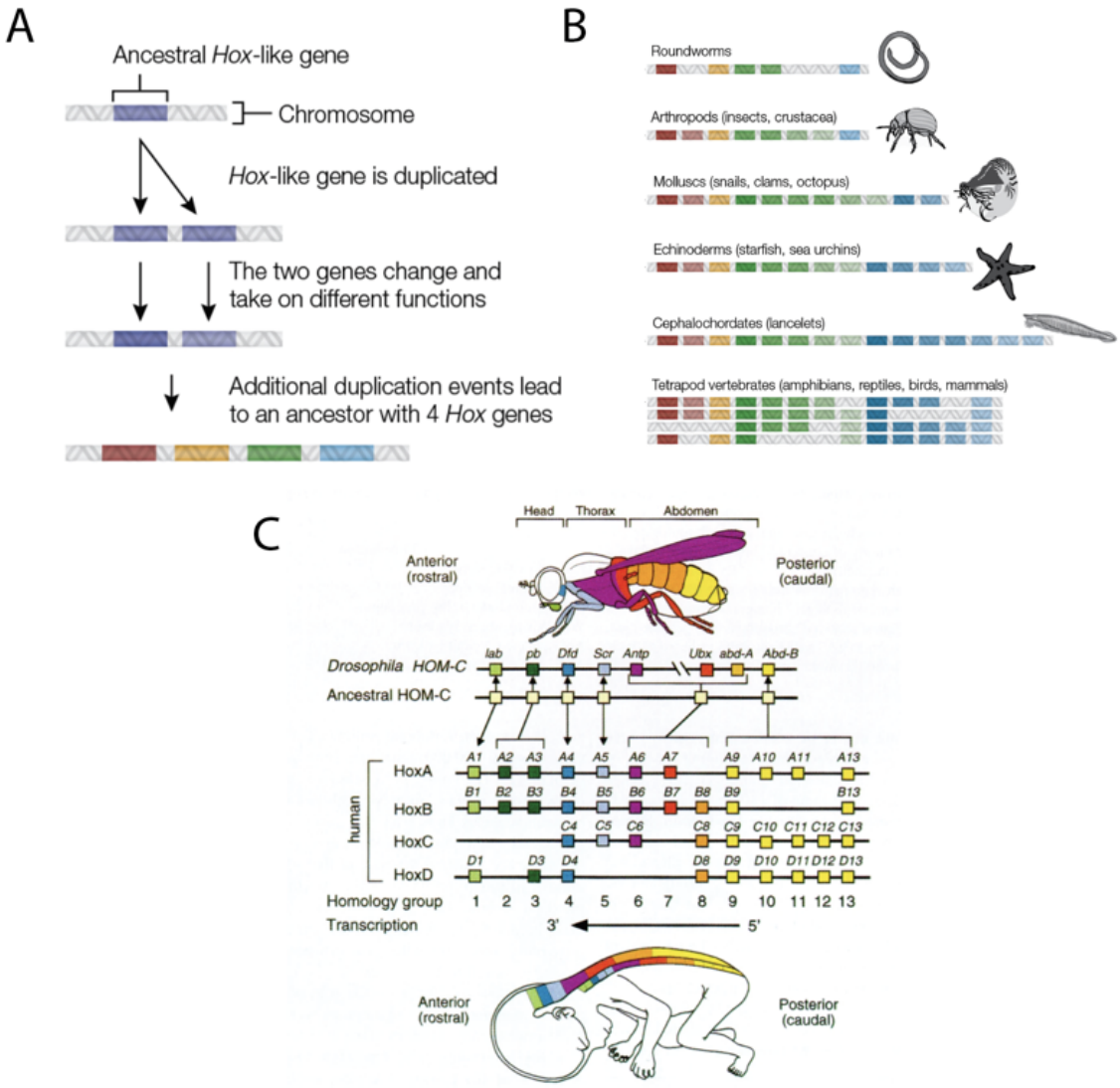


Figure 1. 1 Evolution of Hox genes through duplication

(A) Origin of *Hox* genes. The diversity of *Hox* gene families critical for development originated through the duplication and divergence of an ancestral *Hox*. **(B)** Diversity of *Hox* gene copies in animals. During development, different animal lineages duplicated and diverged *Hox* gene families, contributing to novel body plans. **(C)** Drosophila vs. human development. Orthologous *Hox* gene families (colored) duplicated and diverged in different ways leading to divergent body plans between *Drosophila* and *H. sapiens* (adapted from Chambon *et al*. 1997 and Heffer & Pick, 2013).

One mechanism by which genes duplicate are segmental duplications or SDs: these are defined as duplications of DNA sequences of 1 kbp in length or greater and at least 90% or more identical to one another (IHGSC 2001). They arise by different methods, including duplicative transposition (Johnson *et al*. 2006), nonallelic homologous recombination (Ebert *et al*. 2021), and fork stalling and template switching (Lee *et al.* 2007). However, once formed, SDs provide a substrate of genomic instability and further duplication events (Marques-Bonet & Girirajan *et al.* 2009). SDs are exciting evolutionarily because they are enriched for genes, and responsible for significant adaptations—recent studies into the previously described *Hox* genes now suggests that the four *Hox* paralogs critical for mammalian body plans likely arose through segmental duplication (Ambreen *et al*. 2014). Other notable examples include SD genes contributing to size and longevity in blue whales and elephants (Bukhman *et al*. 2024, Sulak *et al.* 2016), extreme diving capacity in sperm whales (Zhang *et al.* 2025), and immunity in domesticated animals (Feng *et al*. 2017).

1.2 HUMAN EVOLUTION BY SEGMENTAL DUPLICATION

Comparative human genetics, notably the Human Genome Project and the accompanying mouse and chimpanzee genome projects, investigates the genetic differences that distinguish our species from others. This work gives us an evolutionary baseline of our genome and, more practically, provides a basis by which to reduce the sample space of our over 6-billion-base-pair genome to select regions that are conserved, unique to us, and potentially pathogenic if altered in some way. Researchers involved in these projects were initially surprised, and to some extent exasperated, at the paucity of human-specific differences at the gene level: we had essentially the same

number of genes as *Drosophila* or nematodes, and at the protein level are nearly identical to chimpanzees (International Human Genome Sequencing Consortium, 2001). Investigators offered explanations for the lack of differences—our human-specific traits may be due to a few and high-effect point mutations (Enard *et al*. 2002) or specifically by gene loss (Stedman *et al*. 2004). Perhaps most significantly, though, the community has focused on non-genic regulatory sequences as the dominant sequence responsible for human phenotypic changes (King & Wilson, 1975).

Another underappreciated aspect of human genetic divergence is divergence through SDs. Among mammals, primate genomes are conspicuous outliers in their composition and distribution of SDs, which tend to be larger and more distantly dispersed across the genome than expected by chance (Bailey *et al.* 2006). This structure represents a mixed blessing. SDs accelerate evolution. Numerous studies have identified human-specific SD-derived genes critical for our most recent and rapidly evolving biological systems, specifically immunity, metabolism, and neurodevelopment (Bailey *et al.* 2006; Pramanik *et al.* 2011; McLellan *et al.* 1997; Dennis *et al.* 2017). For neurodevelopment, numerous studies have identified human-specific duplicated genes amplifying cell proliferation of neuron progenitors (Florio *et al*. 2016; Namba *et al.* 2020*,* Dougherty *et al.* 2018), dendrite and synaptic plasticity and connectivity (Charrier *et al.* 2012; Dong *et al.* 2024), and neoteny (Florio *et al*. 2018).

Segmental duplication, however, contributes to novel phenotypes in ways besides novel genes. Duplicated sequences provide a substrate for ectopic recombination (i.e., non-allelic homologous recombination; Kim *et al*. 2008). With more distal duplications, these recombinations can and do rearrange large genomic regions or even entire chromosomes (Armengol *et al.* 2003). In these events, even a conserved, homologous protein may find itself in a new regulatory context with

novel function. *LRRC37*, for example, is an ancient and conserved gene that predates the divergence of mice and humans. However, the gene family underwent significant duplication and rearrangement in human lineage, leading to broad expression across tissues, whereas *Lrrc37* remains testis-specific in mice (Bekpen *et al.* 2012).

This rapid evolvability of SDs comes with a consequence: genetic instability. SDs mediate recurrent duplication and deletion syndromes, which typically arise *de novo* and cause debilitating phenotypes in the population. For example, SDs mediate the recurrent duplication and deletion of chr22q11.2, the greatest genetic risk factor for autism (Ousley *et al.* 2009). Other examples include 16p11.2 microduplication (Weiss *et al*. 2008), 15q13.3 microdeletion (Dibbens *et al.* 2009), and 6p22.1 associations with schizophrenia, autism, and intellectual disability (Shi *et al.* 2009). Ultimately, the adaptive benefits these dynamic regions provide may outweigh the risk of their presence, which is why common, highly heritable disorders with reduced reproductive success persist (Stefansson *et al.* 2008; Marques-Bonet & Girirajan *et al.* 2009; Itsara *et al*. 2009).

1.3 ADVANCES IN SEQUENCING AND ASSEMBLY

Despite their enrichment for rapidly evolving and disease-associated genes, SDs remain underrepresented in genomics research due to persistent challenges in mapping and assembly. Traditional platforms like Sanger sequencing and short-read next-generation sequencing (NGS) lack the read length needed to resolve near-identical paralogous sequences, limiting their ability to distinguish recent duplications. While read-depth–based copy number estimation offered some insight into structural diversity (Sudmant *et al*. 2010), it fell short of resolving sequence-level variation. Bacterial artificial chromosome (BAC) libraries provided a partial solution by isolating

individual variants for sequencing. However, the approach was technically laborious and still left many regions unresolved in the reference genome.

Recent innovations in highly accurate, long-read sequencing from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have finally made SDs tractable in genomics research. PacBio high-fidelity (HiFi) sequences reads of roughly 15 kbp with >99.9% sequence accuracy, while ONT has achieved ultra-long sequencing of reads over 100 kbp and >95% sequence accuracy (Logsdon *et al.* 2021). These read lengths and high accuracies enable a single read to span numerous sparse variants of high-identity duplicated genomic DNA. With these anchoring polymorphisms, genomes can be confidently assembled to new telomere-to-telomere (T2T) contiguity (Rautanian *et al.* 2023). Thanks to this technology, the Human Genome Project was finally completed, and the Human Pangenome Reference Consortium (HPRC; Nurk *et al*. 2022; Liao *et al*. 2023) and Human Genome Structural Variation Consortium (HGSVC; Fairley *et al*. 2020) are characterizing structural diversity at the population level.

1.4 *TBC1D3*: A DYNAMIC SD GENE FAMILY IMPLICATED IN NEURODEVELOPMENT

The reference-quality assemblies produced by the Primate T2T projects, HPRC, and HGSVC now make it possible to characterize complex gene families that were previously inaccessible due to limitations in genome assembly (Dishuck *et al.* 2025, Bolognini *et al.* 2024, Plender *et al.* 2024, Real *et al.* 2025). These resources enable us to ask foundational questions about gene family evolution (e.g., how and when did the gene arise, and which copies are under selection?), variation across human populations (e.g., which copies are polymorphic versus fixed, and which

are linked to disease?), and functional impact (e.g., which copies are transcribed, epigenetically regulated, or translated into protein?).

TBC1D3 is a medically relevant, structurally complex SD gene family. The gene family is spread across chromosome 17, but the majority of copies, and all transcribed copies in humans, are situated in two SD clusters, Cluster 1 and Cluster 2, at 17q12 (Hodzic *et al.* 2006). This gene family is young, only present in primates, and the paralogs are, on average, over 99% identical to one another. These two clusters are responsible for the recurrent 17q12 duplication and deletion (i.e., RCAD - Renal Cyst and Diabetes) syndromes. These microdeletions and duplications are phenotypically variable, but both often present with intellectual disability, developmental delay, and autism (Mitchel *et al.* 2016; Mefford *et al.* 2016).

TBC1D3 was first reported in the context of cancer—where it was observed in over 15% of prostate cancers—and reported as an oncogene (Pei *et al*. 2002). Knockdown and ectopic expression experiments of the gene family revealed *TBC1D3*'s capability of promoting cell proliferation (Frittoli *et al.* 2008; Wainszelbaum *et al.* 2008, 2012). These studies provide strong evidence that *TBC1D3* functions in the cytosol, where it amplifies both EGF and IGF pathways to increase cell metabolism and growth, but relied on expression and regulation in immortalized cell lines, including K562 and 293T cells. However, tissue-specific expression from GTEx shows that *TBC1D3* is enriched in the brain (GTEx Consortium, 2020).

In 2016 Ju *et al*. showed that, when expressed in the brain, *TBC1D3* promotes proliferation of outer radial glia, the precursors to cortical neurons. With both transgenic and transduction experiments, they showed that introducing *TBC1D3* to mouse embryos results in expansion of the frontal cortex and cortical folding. Later functional investigations by the group showed that,

in neuron progenitors, TBC1D3 localizes to the nucleus, where the protein inhibits G9a methyltransferase, a H3Kme2 histone methylation factor, to prevent chromatin remodeling and nuclear differentiation, delaying neuron differentiation and promoting extra rounds of cell division (Hou *et al*. 2021). Follow-up research by the group also found *TBC1D3* may be implicated in delay of synaptic maturation (Dong *et al.* 2024).

1.5 RESEARCH GOALS:

This thesis represents a deep exploration of *TBC1D3*, one of numerous primate-specific SD gene families. The objective is to leverage new sequencing modalities to characterize the evolution, variation, and regulation of *TBC1D3*, whose high sequence identity has, until recently, stifled thorough investigation. Understanding a gene's role in human evolution and disease requires accurate assembly, comparisons of the locus between humans and with other primates, and an understanding of the gene's regulation and function in development.

Chapter 2 begins this work by characterizing *TBC1D3*'s evolution in primates and its variation in the human population. In this chapter, I show evidence for a unique and dynamic evolution, with multiple rounds of independent expansion, and a human-specific, derived modification of the final 41 amino acids of the carboxy terminus. We also compare human diversity at *TBC1D3* and show that the gene family maintains incredible copy number and structural diversity. We complete the work with a pangenomic characterization of the individual paralog members of the gene family and show that *TBC1D3* expression may be constrained to a select few paralogs.

Next, in Chapter 3, we test the hypothesis of paralog-specific regulation, elucidating an exciting regulatory mechanism that explains tolerance for high copy number polymorphism of *TBC1D3*.

We use two separate epigenetic modalities, comparative transcriptomics, and a neuronal developmental model, to show that *TBC1D3* expression is controlled via fusion with *NPEPPSP1*, an upstream neighbor and copy number-constrained promoter.

In Chapter 4, I provide a brief summary of our attempts to investigate human-specific modifications to *TBC1D3*. We describe the experimental design, the results of the failed experiment, and hypotheses explaining why the experiment did not succeed.

Finally, in Chapter 5, I discuss the future directions, both for characterization of *TBC1D3* and other segmentally duplicated gene families in humans. This includes knock-down experiments targeting the fixed *TBC1D3* promoter and association studies of a subset of structural haplotypes with an inversion-breaking *NPEPPSP1-TBC1D3* fusion regulation. I will also discuss what resources are still missing.

As of February 2026, Chapter 2 has been published, and Chapter 3 has been submitted for peer review.

CHAPTER 2. INDEPENDENT EXPANSION, SELECTION AND HYPERVARIABILITY OF THE *TBC1D3* GENE FAMILY IN HUMANS

Chapter 2 is adapted with minimal modification from:

Guitart, X., Porubsky, D., Yoo, D., Dougherty, M. L., Dishuck, P. C., Munson, K. M., Lewis, A. P., Hoekzema, K., Knuth, J., Chang, S., Pastinen, T., & Eichler, E. E. (2024). Independent expansion, selection, and hypervariability of the TBC1D3 gene family in humans. Genome Research, 34(11), 1798–1810. https://doi.org/10.1101/gr.279299.124

Author Contributions: X.G. and E.E.E. conceived the project; X.G. assembled genomes, performed QC analyses, and conducted the analyses relevant to all manuscript figures; D.P. identified the HG01109 human inversion; D.Y. computed nucleotide diversity of unique sequence flanking *TBC1D3* clusters; M.L.D. conducted the probe capture experiments for fetal brain Iso-Seq data; P.C.D. provided technical and scientific consultation and assisted in Iso-Seq library processing; K.M.M. and A.P.L. generated PacBio HiFi and Iso-Seq sequencing data; K.H. and J.K. generated ONT sequencing data; S.C. processed samples and DNA necessary for mouse lemur genome assembly; T.P. generated iPSC Iso-Seq libraries; X.G. and E.E.E. drafted the manuscript. All authors read and approved the final manuscript.

2.1 ABSTRACT

TBC1D3 is a primate-specific gene family that has expanded in the human lineage and has been implicated in neuronal progenitor proliferation and expansion of the frontal cortex. The gene family and its expression have been challenging to investigate because it is embedded in high-identity and highly variable segmental duplications. We sequenced and assembled the gene family using long-read sequencing data from 34 humans and 11 nonhuman primate species. Our analysis shows that this particular gene family has independently duplicated in at least five primate lineages, and the duplicated loci are enriched at sites of large-scale chromosomal rearrangements on Chromosome 17. We find that all human copy number variation maps to two distinct clusters located at Chr. 17q12 and that humans are highly structurally variable at this locus, differing by as many as 20 copies and ~1 Mbp in length depending on haplotypes. We also show evidence of positive selection, as well as a significant change in the predicted human TBC1D3 protein sequence. Lastly, we find that, despite multiple duplications, human *TBC1D3* expression is limited to a subset of copies and, most notably, from a single paralog group: *TBC1D3-CDKL*. These observations may help explain why a gene potentially important in cortical development can be so variable in the human population.

2.2 INTRODUCTION

Gene duplication followed by adaptation is one of the primary forces by which new genes emerge within species (Ohno, 1970). Many of these evolutionary events occur in segmental duplications (SDs), genomic units that are at least one kilobase pair in length and whose duplications are 90% or more identical to one another (Bailey and Eichler 2006). Many human-

specific genes reside in SDs, which often continue to vary structurally in our lineage (Bitar et al. 2019). Since the initial publication of the human and chimpanzee genomes, investigations of human-specific SD genes have found that they most often are implicated in xenobiotic recognition, metabolism, immunity, and neuronal development, playing an important role in the evolution of our species (Perry et al. 2007; Huttner et al. 2024; Dennis et al. 2012).

TBC1D3 is a primate-specific SD gene family (Paulding et al. 2003). This gene family is dispersed across the two arms of Chromosome 17, though most copies in humans map to two expansion blocks at locus Chromosome 17q12 (Fig. 2.1A). Expression data in humans from the Genome-Tissue Expression (GTEx) project reveal *TBC1D3* is modestly expressed globally, with increased expression in testis and brain tissue (The GTEx Consortium 2020). *TBC1D3* expression and function were initially observed in prostate tumor samples and originally classified as an oncogene (Hodzic et al. 2006). However, in 2016, Ju et al. showed that transgenic overexpression of *TBC1D3* in the developing mouse brain results in a proliferation of outer radial glial cells and a subsequent expansion and folding of the cortex (Ju et al. 2016).

These findings suggest that the evolution of *TBC1D3* may have contributed to human cranial expansion over the last two million years (Stringer 2016). Investigations of the sequence evolution and variation amongst humans and nonhuman primates (NHPs) would help test this hypothesis (Sabeti et al. 2006). However, the duplicated and highly identical sequences of *TBC1D3* copies make assembly impossible with standard short-read sequencing platforms. Instead, researchers have investigated copy number variation in SD genes using short-read sequencing data to understand patterns of variation (Sudmant et al. 2010). Such read-based studies have suggested extensive copy number differences among human populations. However, these experiments lack the single-base-pair resolution necessary to distinguish different

paralogous copies, structural differences among haplotypes, and which copies are likely functional or expressed. Moreover, it is unclear how a gene so variable in copy number could play such a critical role in the expansion of the frontal cortex in humans. In this study, we address these questions by leveraging long-read sequencing data generated from humans and apes to fully resolve the *TBC1D3* loci (Liao et al. 2023; Mao et al. 2024; Makova et al. 2023). The goals of this study were to reconstruct the evolutionary history of this gene family, to assess the extent of human genetic diversity, and to determine how this variation relates to changes in selection and expression of the gene family in the human lineage.

2.3 RESULTS

2.3.1 HUMAN *TBC1D3* COPY NUMBER VARIATION.

To understand *TBC1D3* organization and variation in humans, we first focused on two *TBC1D3* gene family clusters that contain the majority of *TBC1D3* paralogs, named cluster 1 and cluster 2 (Fig. 2.1B). We characterized 44 human genomes recently sequenced as part of the Human Pangenome Reference Consortium at this locus (Supplemental Table S1; Liao et al. 2023). We first assessed the integrity of each assembly by searching for sequence collapses in read depth of both Pacific Biosciences (PacBio) high-fidelity (HiFi) and Oxford Nanopore Technologies (ONT) sequencing data (Supplemental Table S2; Vollger et al. 2019; Dishuck et al. 2022; Methods). We found that 46 of the haplotypes passed quality control (QC), while 42 haplotypes failed. We attempted to re-assemble the samples that failed QC using a novel assembly algorithm that leverages both HiFi and ONT data (Verkko) (Rautiainen et al. 2023). This procedure recovered an additional 20 haplotypes where both cluster 1 and cluster 2 were fully sequenced and assembled without error (Supplemental Fig. S1). We also confirmed accurate assembly with an orthogonal sequencing platform by comparing assembly predicted copy number against Illumina read depth–based copy number estimates (Fig. 2.1A; Supplemental Fig. S2; Methods). For our investigations we required that both haplotypes of the assembly accurately resolve. In total, we validated 66 haplotypes where both *TBC1D3* clusters were fully resolved and, including three genome references, developed a total dataset of 69 human haplotypes.

Next, we estimated the copy number and organization of *TBC1D3* in clusters 1 and 2 for each human haplotype (Fig. 2.1B-D). In cluster 1, we found that *TBC1D3* varies from 1 to 14 copies, while in cluster 2 it varies from 2 to 14 copies (Supplemental Table S3). Thus, human diplotype

copy number for *TBC1D3* summing across both clusters could theoretically range from 6 to 56 based on our limited survey of human diversity. The differences in copy account for as much as 1.5 Mbp of differential size between human haplotypes. Notably, we find that *TBC1D3* copy number is significantly higher among Africans (X=34.4) when compared to non-African populations (X=25.4) (p-value = 1.7E-5). Higher African copy number is an observation that has been confirmed by Illumina WGS read-depth analysis for *TBC1D3* and seen for other recently duplicated copy number polymorphic loci (Vollger et al. 2022; Jeong et al. 2024). The basis for this is unknown but it may reflect the genetic bottleneck in the out-of-African founder populations or another manifestation of overall increased genetic diversity of African populations. For cluster 1, we find that 65% (45/69) of the haplotypes are structurally distinct. Additionally, for cluster 2, we observe similar diversity, where 68% (47/69) are structurally distinct (Supplemental Fig. S3). Based on completely assembled diploid samples, we estimate the structural heterozygosity for cluster 1 is 94%, while cluster 2 is 88%, making these two loci among some of the most structurally variable gene families in the human genome (Sudmant et al. 2010).

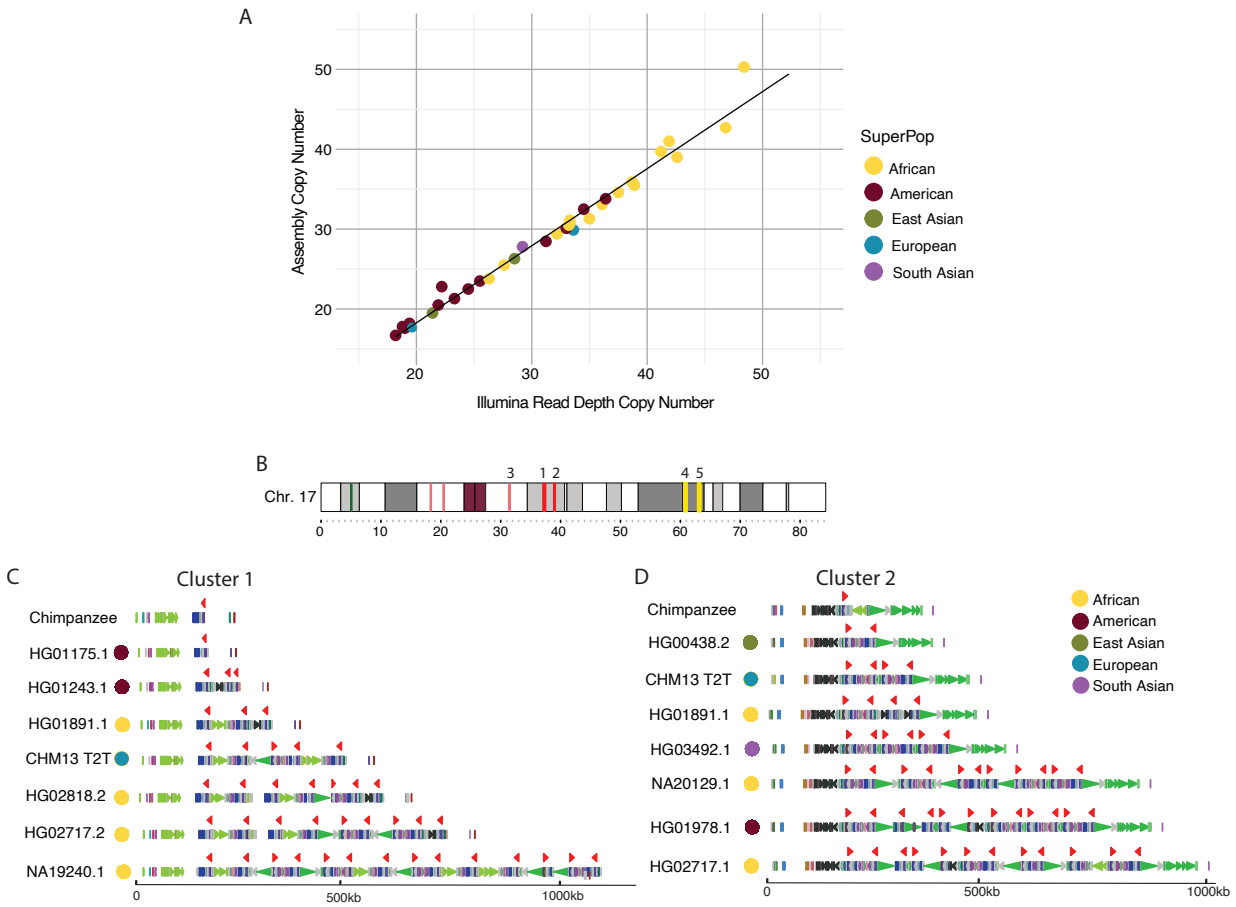


Figure 2. 1 Assembly and human variation of TBC1D3.

(A) Assembly copy number estimate versus orthogonal Illumina sequence copy number estimate. Each point represents a sample diploid assembly, colored by superpopulation**. (B)** Reference ideogram of TBC1D3 regions. Expanded views of clusters 1 and 2 (marked in red) are illustrated in C, D. **(C,D)** Structure for chimpanzee and seven validated human haplotypes over TBC1D3 cluster 1 (C) and cluster 2 (D). TBC1D3 copies are colored as red arrows. Colored arrows below TBC1D3 illustrate segmental duplication content annotated with DupMasker (Jiang et al. 2008).

2.3.2 NONHUMAN PRIMATE (NHP) *TBC1D3* ORGANIZATION.

To better understand the evolution of the clusters, we investigated the organization of *TBC1D3* in 10 different NHP lineages (Supplemental Table S4). This included single representatives of five great ape species (bonobo, chimpanzee, gorilla, Bornean, and Sumatran orangutan), two Old World monkeys (macaque, gelada), two New World monkeys (marmoset, owl monkey), and one prosimian (mouse lemur). Eight of these genomes were previously published (Mao et al. 2024) or are part of efforts to generate telomere-to-telomere (T2T) assemblies of ape genomes (Makova et al. 2023). We generated HiFi sequence data from both the gelada and mouse lemur genomes in this study and assembled their genomes using Hifiasm (Methods).

With the exception of the mouse lemur, all NHP genomes carry multiple copies of *TBC1D3* (Supplemental Table S5). We find that *TBC1D3* is also highly copy number variable among NHPs, from two copies in marmoset to 31 copies within a single haplotype in both gelada and gibbon. We searched specifically for clustered expansions and found that most primates—human, gorilla, orangutan, macaque, and gelada—similarly contain two expanded clusters of *TBC1D3* (Fig. 2.2A). Among apes, these two clusters are orthologous to human clusters 1 and 2, separated by 1.35 Mbp of intervening sequence. Among the Old World monkeys, gelada and macaque, structural rearrangements have repositioned the two clusters such that the intervening sequence is larger and nonsyntenic. Importantly, bonobo and chimpanzee only possess 1-2 copies of *TBC1D3* at cluster 2, whereas no copies were identified at cluster 1. Thus, all humans have an increase in copy number when compared to the *Pan* lineage but are not exceptional when compared to most other NHP lineages. New World monkeys, owl monkey, and marmoset do not have *TBC1D3* organized into clusters. Instead, marmoset has two copies while owl

monkey has eight copies distributed throughout its chromosome, suggesting independent and recent expansions. Overall, we find that *TBC1D3* copy number varies from 0 to 14 copies in cluster 1 and from 1 to 17 copies in cluster 2 (Fig. 2.2B). A detailed analysis of the composition of the SDs within each primate lineage shows that the units of duplication in different species frequently differed in structure, suggesting independent duplications or gene conversion events in each lineage (Supplemental Fig. S4; Methods).

In order to estimate when the clustered *TBC1D3* copies expanded in each lineage, we constructed a maximum likelihood phylogenetic tree based on a multiple sequence alignment (MSA) generated from intronic sequence of each predicted *TBC1D3* gene copy from the various primate genomes (Fig. 2.2B; Methods). We observe complete lineage-specific stratification of the *TBC1D3* gene family members into distinct clades for human, *Pan*, gorilla, orangutan, gibbon, and owl monkey lineages. These findings strongly support recurrent duplication or gene conversion of all gene family copies in each lineage. In contrast, the gelada and rhesus macaque show both shared and lineage-specific groups, suggesting *TBC1D3* expanded before and after speciation. Using 25 and 6.5 million years ago (mya) as times of human–macaque and human–chimpanzee divergence, we estimated the timing of each lineage-specific expansion (Fig. 2.2B; Stevens et al. 2013; Dunsworth et al. 2010). In most lineages, the primate duplications occurred relatively recently. Most notably, we observe that humans experienced the most recent expansion within the apes, occurring between 2.0 and 2.6 mya.

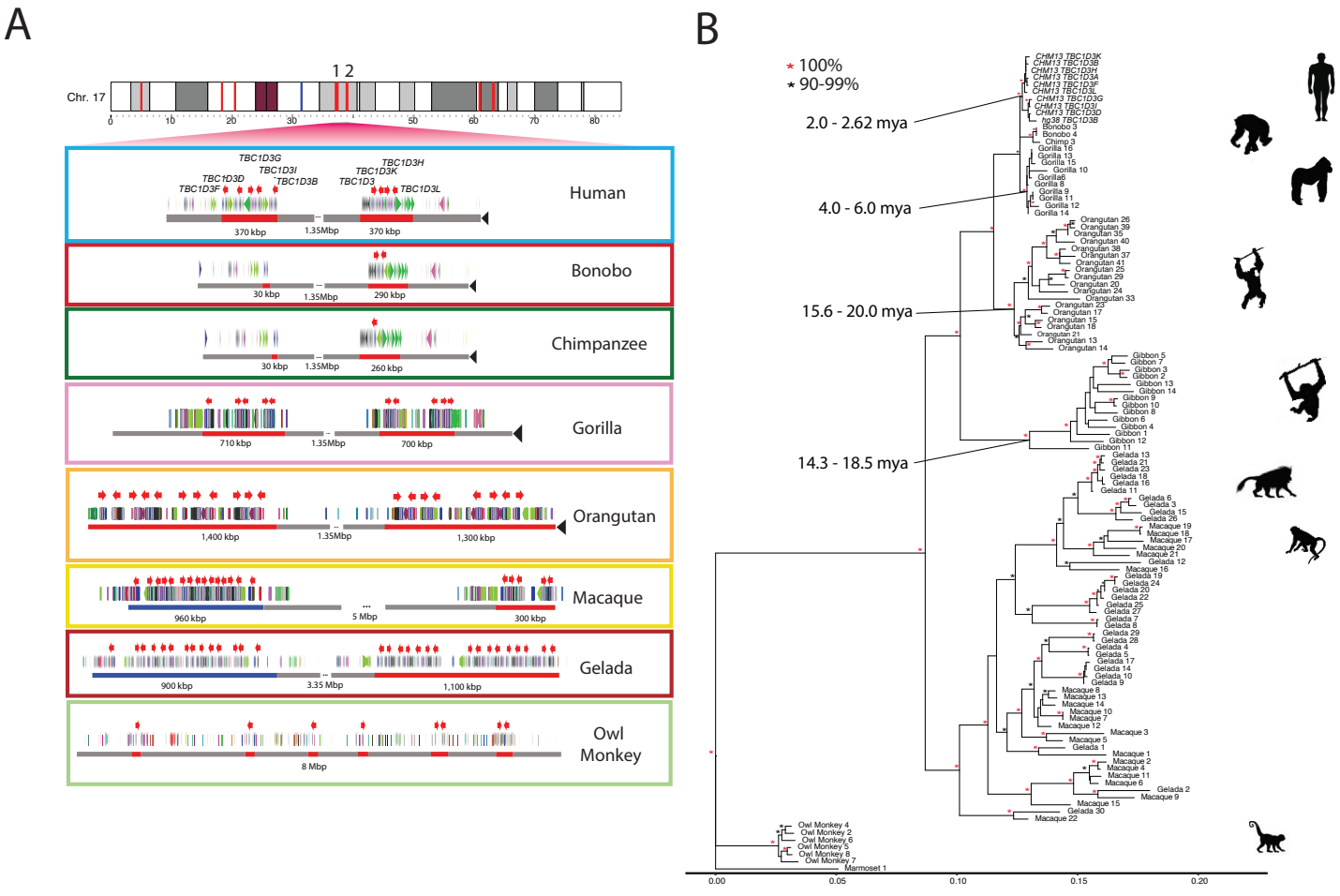


Figure 2. 2 Comparative genome structure and phylogeny of TBC1D3 gene family among primates.

(A) TBC1D3 clusters 1 and 2 structure. Orthologous TBC1D3 clusters 1 and 2 are illustrated as two clustered regions (red blocks), with flanking unique sequence in gray for the primate lineages. Old World monkey TBC1D3 expansion 1, which is nonsyntenic, is highlighted in blue. TBC1D3 paralogs (red arrows) are embedded within other segmental duplication blocks, with DupMasker annotations illustrated with colored arrows. The diverse organizational differences of each expansion, including expansion size, duplicon content, and copy number, suggest independent expansion. **(B)** TBC1D3 neutral phylogeny generated by maximum likelihood; 2300 bp of intronic sequence were aligned between all primate TBC1D3 paralogs observed in A, with the marmoset sequence used as an outgroup. The phylogeny supports the hypothesis of independent expansion with the exception of the Old World monkeys (geladas

and macaques) in which several copies duplicated before and after speciation of these two lineages (11 mya) (Liedigk et al. 2014).

2.3.3 *TBC1D3* AND LARGE-SCALE CHROMOSOMAL REARRANGEMENTS.

During our comparative analysis of NHP genomes, we noticed that chromosomal synteny frequently was disrupted at sites corresponding to interspersed TBC1D3 loci. To assess this more systematically, we selected five primate lineages for which T2T assemblies had recently been generated as part of the Primate T2T Consortium, aligned orthologous Chromosome 17s to one another, and illustrated these alignments, as well as alpha satellite and TBC1D3 loci (Methods) (Fig. 2.3A). We found that TBC1D3 consistently flanks some of the largest chromosomal rearrangements. For example, human TBC1D3P2 demarcates one end of a 12 Mbp large-scale chromosomal inversion distinguishing human and Sumatran orangutan chromosomes (see light blue alignment in Fig. 2.3A, B). In the orangutan, the corresponding breakpoint of synteny is anchored in one of the expanded TBC1D3 clusters. This structure is syntenic with the macaque, suggesting that it was the ancestral configuration, whereas the human structure, shared with gorillas and chimpanzees, was derived. Similarly, one of the fission breakpoints of Chromosome 17 resulting in gorilla Chromosomes 4 and 19 (Stankiewicz et al. 2001) maps precisely to TBC1D3 and USP6 duplications in the gorilla lineage.

To test if the association with TBC1D3 and breakpoints of synteny was significant, we developed a permutation test. We randomly selected an equivalent sequence and number of mappings throughout Chromosome 17 for these five orthologous primate chromosomes and measured the median distance of these mappings to the nearest synteny break. In more than 5000 permutation tests, we never observed a distance as low as that of true TBC1D3 mappings (Supplemental Fig.

S5). We repeated the test by limiting our samplings to SD sites on Chromosome 17. Even with this restriction, the observed distance to TBC1D3 resided in the bottom 3% of the simulated distribution (Fig. 2.3C), suggesting a nonrandom association of TBC1D3 SDs with large chromosomal rearrangements during primate evolution.

To assess the origin of TBC1D3 gene clusters, we sequenced and assembled the genome of an outgroup primate species using HiFi data generated from a mouse lemur (Microcebus murinus) and identified two sequence contigs (2.8 Mbp and 14 Mbp) spanning the region (Fig. 2.3D). Both clusters 1 and 2 appeared to be absent; however, the corresponding regions demarcate breakpoints of synteny compared with Old World monkey and ape lineages. Additionally, we aligned TBC1D3 against the entire mouse lemur assembly with BLASTN but could not identify any TBC1D3 orthologs, suggesting TBC1D3 is exclusive to the simian infraorder (Supplemental Table S6; Zhang et al. 2000). We followed up this analysis and compared human and owl monkey TBC1D3 orthologs by genomic synteny and phylogenetic approaches to identify the putative simian ancestral TBC1D3 paralog but did not find a consistent candidate (Supplemental Fig. S6).

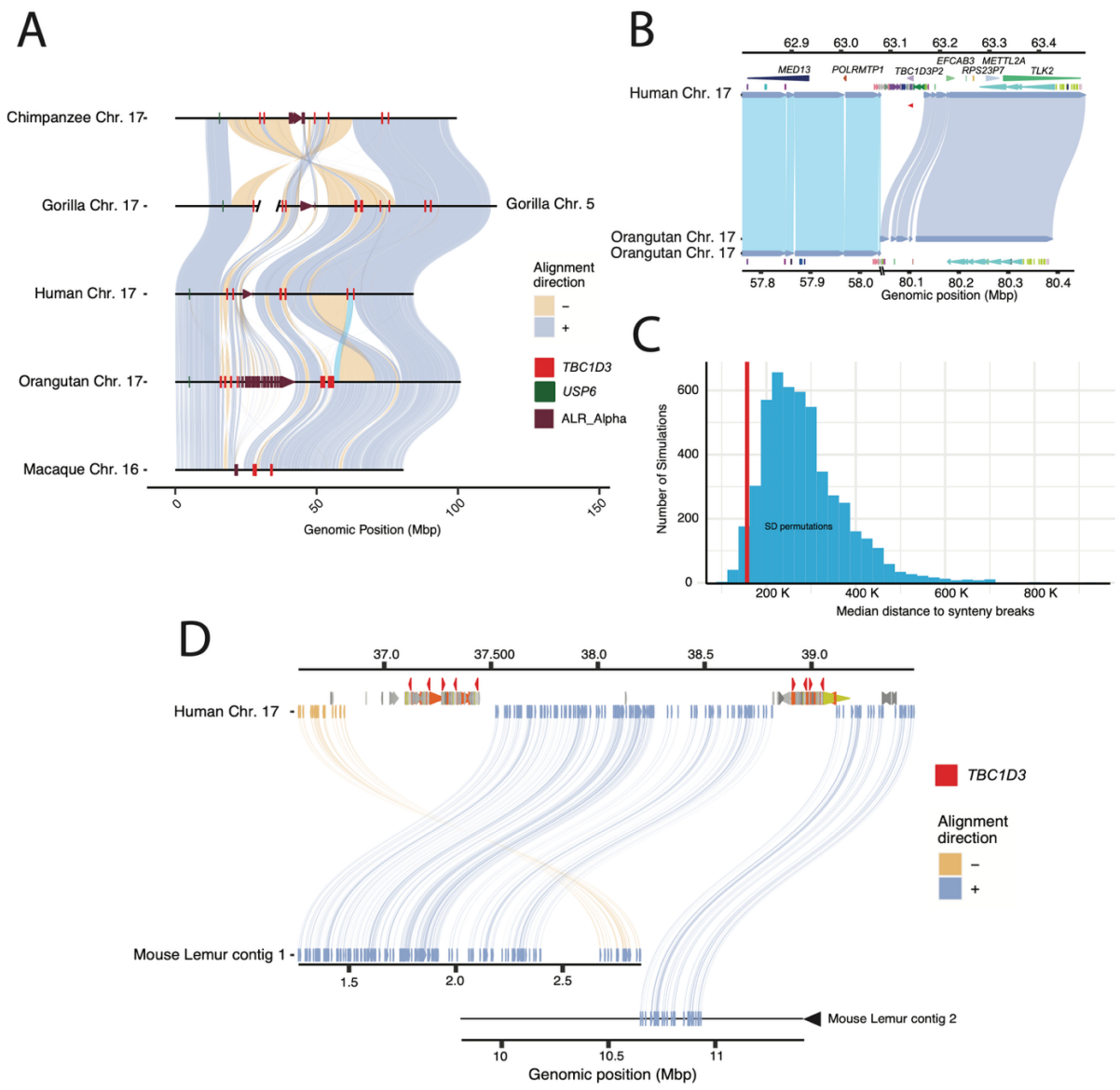


Figure 2. 3 Large-scale chromosomal rearrangements and TBC1D3 duplications.

(A) Synteny plots of orthologous Chromosome 17 in primates reveal syntenic blocks in direct (blue) and inverted (yellow) orientation. Alpha satellite sequence, TBC1D3 copies, and USP6—a hominoid fusion gene of TBC1D3—are illustrated in maroon, red, and green, respectively. TBC1D3 demarcates the boundaries of large-scale rearrangements on chromosome phylogenetic group XVII. **(B)** TBC1D3 duplication block (cluster of colored arrows) demarcates the boundary of a 12 Mbp inversion between the human and orangutan chromosomes. **(C)** Permutation test of segmental duplication proximity to synteny breaks. Five thousand permutation tests were performed, in which segmental duplication samples were

taken, and median proximity to breaks in synteny was measured. True TBC1D3 mappings fall within the lowest 3% of the permutations (red line), suggesting a nonrandom association between TBC1D3 and breakpoints in synteny. **(D)** Synteny plot showing orthologous alignments between human TBC1D3 and mouse lemur flanking genomic sequence.

2.3.4 *TBC1D3* TRANSCRIPT AND OPEN READING FRAME PREDICTION.

Gene model characterization of TBC1D3 has been particularly challenging given the high sequence identity and variable nature of the duplicated genes. This has made it difficult to distinguish genes that are expressed and potentially functional from pseudogenes. To address this limitation, we sequenced HiFi, full-length nonchimeric (FLNC) cDNA using a PacBio isoform sequencing (Iso-Seq) assay (Methods) (Dougherty et al. 2018). We generated or analyzed data from testis tissue of chimpanzees, gorillas, bonobos, and Sumatran and Bornean orangutans (Makova et al. 2024) and from pooled human fetal brain tissue (Supplemental Table S7). Additionally, we analyzed a very deep pool of about 500 million human FLNC reads recently generated from induced pluripotent stem cells (iPSCs) (Cheung et al. 2023). We mapped FLNC reads to both haplotypes of the respective species of origin genome assemblies, allowing only high-quality mappings and tracking all best map assignments versus multiple mappings among the paralogous copies for each species (Methods) (Fig. 2.4A). Although unambiguous one-to-one assignments between transcripts and specific paralogs could not always be made, the analysis revealed three important features. First, TBC1D3 is transcribed in all ape lineages with evidence of multiple paralogs expressed where there are duplications (Supplemental Fig. S7). Second, the canonical 14-exon gene model is retained across the apes, with evidence of exon exaptation and exon loss for a minority subset of transcripts in chimpanzees and Sumatran orangutans (Fig.

2.4A). Third, the predicted open reading frame (ORF) is, in general, maintained. In humans, however, both transcription and ORF maintenance are most likely to be retained among TBC1D3 copies mapping to clusters 1 and 2 in contrast to distal orphan copies (see Fig. 2.2A, human Chromosome 17 ideogram).

During our comparison of human and NHP TBC1D3 gene models, we noted that all human transcripts harbor a 43 bp deletion in the ORF absent in NHPs (Fig. 2.4B). This deletion removes the last 17 amino acid residues common to NHPs and introduces a frameshift, resulting in a 41 amino acid extension and a novel C terminus of the human TBC1D3 protein. All other NHPs lack this carboxy extension owing to a shared common stop codon. We also confirmed this human-specific difference at the level of the assembly using ProSplign (Methods) (https://www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html). Furthermore, the 43 bp deletion is restricted to TBC1D3 copies mapping to human clusters 1 and 2, in which 95% (850/896) of cluster 1 and 2 copies contain the deletion, and it is not observed among the older orphan paralogs distributed throughout human Chromosome 17 (Supplemental Fig. S8). These findings indicate that this fundamental change in the ORF is human-specific and occurred during human TBC1D3 expansion within clusters 1 and 2. We predicted the effect of this modification on the tertiary structure of TBC1D3 using AlphaFold2 but found that the novel C-terminal sequence was disordered (Supplemental Fig. S9; Jumper et al. 2021).

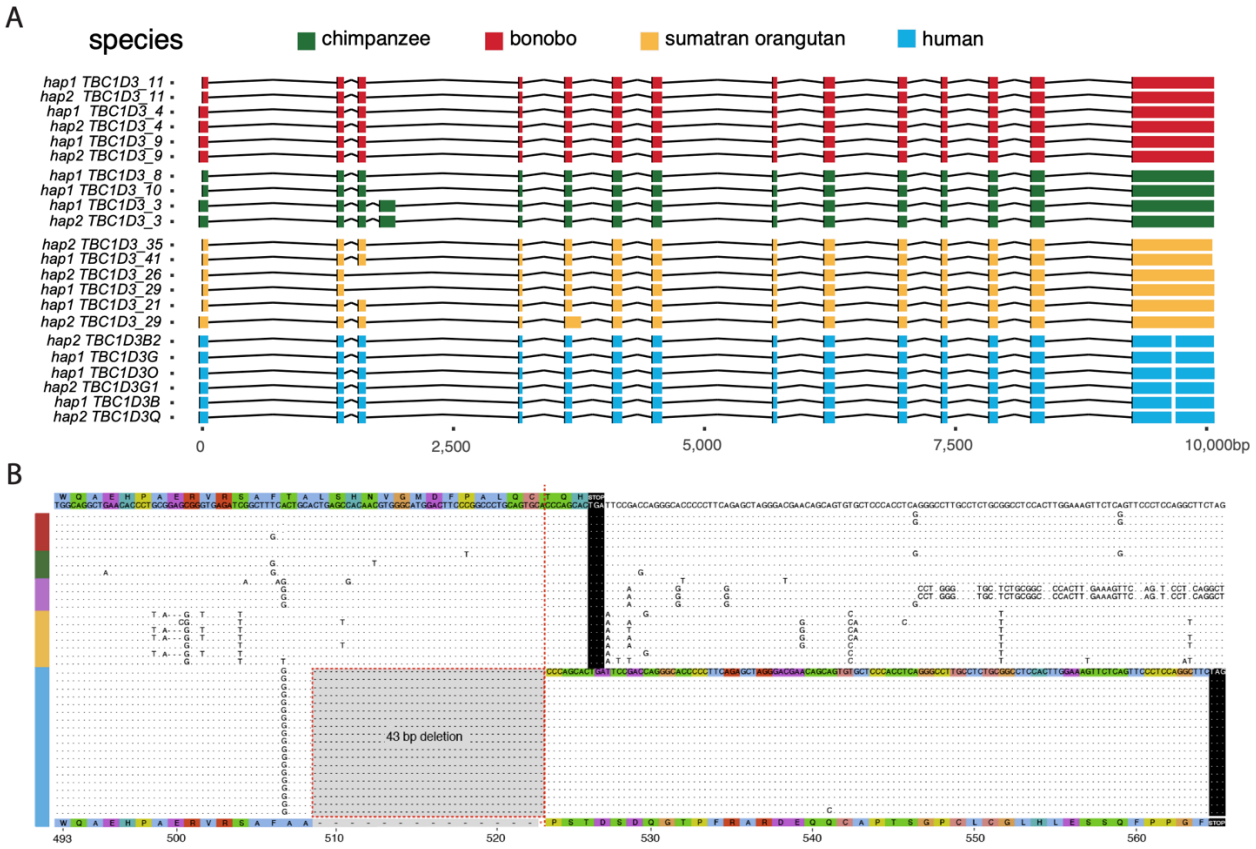


Figure 2. 4 Human-specific C-terminal modification of TBC1D3.

(A) The intron/exon structure of expressed TBC1D3 isoforms with protein-encoding ORFs. Each row constitutes a paralog-specific isoform observed based on Iso-Seq (Methods). All isoforms were mapped to human USP6 for a common reference. Exons are colored by species, with arches representing introns. **(B)** Amino acid sequence alignment of the C terminus of expressed primate TBC1D3 paralog sequences predicted from Iso-Seq full-length cDNA. All, and only, human-expressed copies contain a 43 bp deletion within the ORF of the terminal exon, resulting in a frameshift, creating an extension of 41 novel amino acids to the C terminus.

2.3.5 AFRICAN APE POSITIVE SELECTION.

Using the full-length transcript isoforms that were generated and mapped to the complete genome assemblies from each primate (Fig. 2.5A), we constructed two MSAs using intronic sequence and codon-aligned exonic regions. First, we explored branches putatively under positive selection using a free-ratios model (Methods) (Yang 2007). We identified three branches and tested these for a significant excess of amino acid replacements using the codon MSA in an adaptive branch-site random effects likelihood test (absREL; Methods) (Supplemental Table S8; Smith et al. 2015). After multiple test correction, we found strong statistical support for positive selection in one of the three branches, within the ancestral branch leading to African ape cluster 1 and cluster 2 TBC1D3 copies ($P = 0.01$; Methods) (Fig. 2.5B). This positive selection is detected only for TBC1D3 copies mapping to clusters 1 and 2 and not among orphan copies or other ape clusters distributed along Chromosome 17. Furthermore, this selection occurred after divergence from orangutans and after an African ape–specific translocation of TBC1D3 paralogs to Chromosome 17q23 (Fig. 2.3A). Orangutan copies expressed from clusters 1 and 2 do not show signatures of positive selection, nor do expressed chimpanzee/bonobo copies mapping distally to clusters 3 and 4 (yellow). Focusing on African ape copies mapping to clusters 1 and 2, we tested for site-specific signatures of positive selection on amino acid residues with a branch-site model (Methods) (Yang 2007). Using a Bayesian posterior probability cutoff of 0.9, we identified six sites of positive selection, with the strongest signals mapping within the TBC/Rab GTPase-activating protein (GAP) domain, as well as two residues proximal to the C terminus of TBC1D3 (Fig. 2.5C). These signals of positive selection cannot be explained by gene conversion (Supplemental Fig. S10).

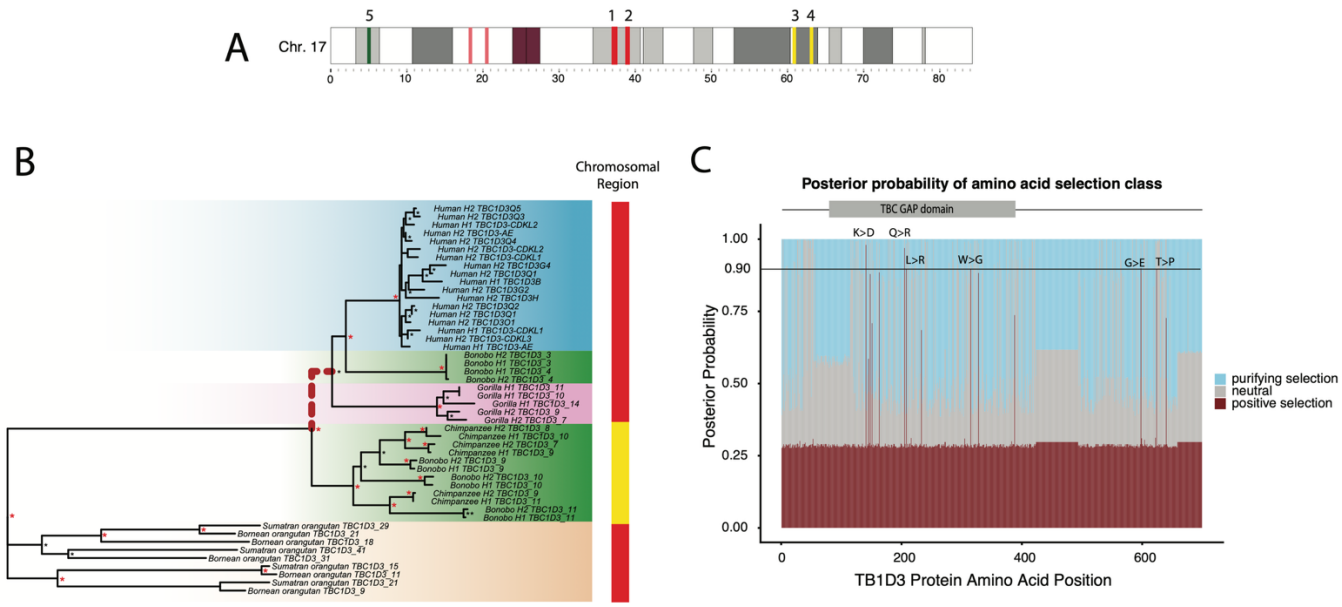


Figure 2. 5 Positive selection of the *TBC1D3* gene family.

(A) Chromosome 17 ideogram marking TBC1D3 expansion clusters (red) and distal loci (yellow) expressed in chimpanzee and bonobo. **(B)** Branch site test of selection for expressed TBC1D3 paralogs. A maximum likelihood phylogeny corresponding to the introns of expressed TBC1D3 paralogs used to visualize relation of expressed copies. The red dashed branch illustrates the ancestral branch identified under positive selection with absREL (P-value = 0.01). Colored bars on the right of the phylogeny illustrate the location of origin of each TBC1D3 copy as illustrated in A, red indicating paralogs from clusters 1 and 2 and yellow marking expressed paralogs from distal q-arm expansions 3 and 4. **(C)** Sites under selection along TBC1D3. A branch site model was conducted using the codon alignment of the same TBC1D3 expressed isoforms, with the branch leading to African ape cluster 1 and cluster 2 TBC1D3 copies as the foreground and all other branches as the background. Posterior probabilities for positive, neutral, and purifying selection are illustrated in red, gray, and blue, respectively, with red indicating sites under selection in the foreground branches (omega = 52.6). Six sites were observed with strong evidence of positive selection (141K > D; 205Q > R; 208L > R; 315W > G; 598G > E; 624T > P).

2.3.6 PANGENOMIC CHARACTERIZATION AND TRANSCRIPTION OF HUMAN *TBC1D3* COPIES.

Given the extraordinary copy number variation among human copies mapping to clusters 1 and 2, we applied a pangenomic approach to organize and characterize human paralogs. We initially constructed pangenome graphs with minigraph from the sequence-resolved human haplotypes. However, few paralogs were grouped as common or shared but, instead, the majority of TBC1D3 copies were represented as isolated nodes with single-haplotype support (Supplemental Fig. S11; Li et al. 2020). As a result, we applied a phylogenetic approach that organized TBC1D3 copies into groups in which genetic distance exceeded the expected level of intra-allelic variation (Methods). We defined 11 distinct phylogenetic groups (Fig. 2.6A) and named them based on TBC1D3 paralogs already present in the human reference genome (GRCh38) (Supplemental Fig. S12). In some cases, multiple distinct paralogs were placed into the same phylogenetic group if paralogous variation was less than the expected extent of allelic variation (e.g., TBC1D3-AE or TBC1D3-CDKL). We identified four novel phylogenetic groups representing paralogous copies not present in the human reference genome assembly: TBC1D3M, TBC1D3N, TBC1D3O, and TBC1D3Q. Most phylogenetic groups are distributed across human continental population groups and are specific to either cluster 1 or 2. TBC1D3F, however, is exclusive to Amerindians and maps to cluster 2, yet has greater homology with cluster 1 TBC1D3 members. A detailed examination of the genomic organization of one of these Amerindian haplotypes, HG01109 H2, reveals that the entire 1.35 Mbp region bracketed by clusters 1 and 2 has been inverted, suggesting that inversion, as well as gene conversion, may be playing a role in relocating TBC1D3 paralogs between clusters 1 and 2 (Fig. 2.6B).

Using this phylogenetic group classification of cluster 1 and 2 members, we revisited expression of the TBC1D3 gene family in humans, taking advantage of the deep Iso-Seq data sets that had been generated from both iPSCs and fetal brain (Supplemental Table S7). We mapped FLNC reads from both sources to the phylogenetic pangenome groups and identified the best primary paralog mapping for each read (Methods). We find that the majority of TBC1D3 expression—91% in iPSCs and 96% in fetal brain—originates from cluster 2–specific paralogs. Furthermore, the majority of this sequence—89% in fetal brain and 69% in iPSCs—maps to a single phylogenetic group: TBC1D3-CDKL. This enriched paralog expression is consistent, even when normalized by median TBC1D3 paralog copy (Fig. 2.6C). It is noteworthy that for 67 of the 69 assembled haplotypes, this expressed TBC1D3 paralog is the last copy in cluster 2 and, furthermore, is oriented such that the unique sequence flanking this telomeric end of the cluster is directly upstream to its transcription start site. A genome-wide analysis identified that the 20 kbp of this unique sequence falls within the lower 5% for pairwise nucleotide diversity and may reflect either a selective sweep or regulatory sequence under strong purifying selection (Supplemental Table S9). This paralog expression exclusivity may explain why a gene family predicted to be critical to cortical expansion may be so variable in copy number and structure among humans.

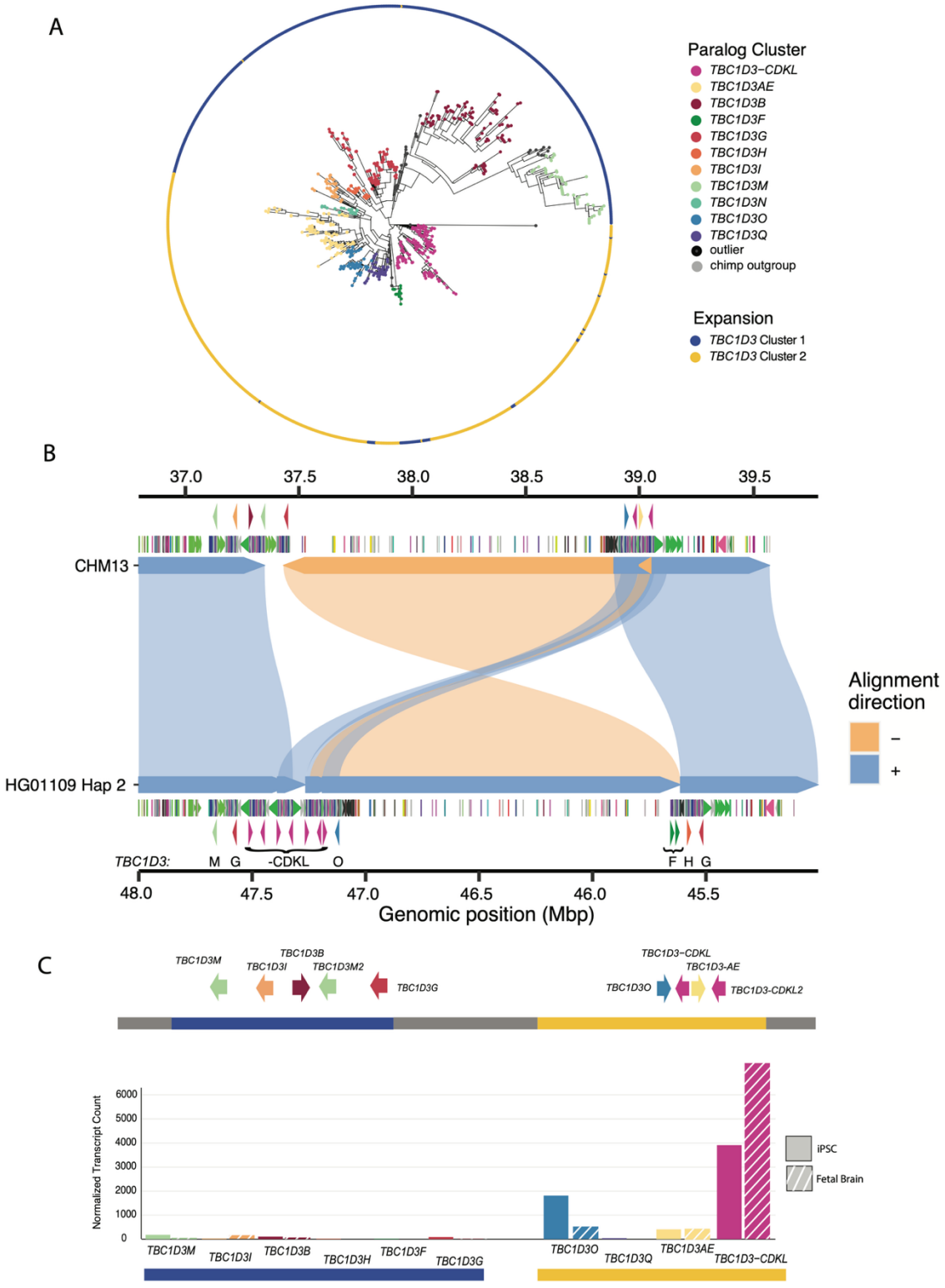


Figure 2. 6 Pangenomic characterization and expression of *TBC1D3* in humans.

(A) Maximum likelihood phylogeny of all validated TBC1D3 cluster 1 and 2 paralogs in humans, outgrouped to chimpanzee TBC1D3. Individual cluster paralogs were identified by limiting intra-cluster variation to a 1.5× allelic variation observed in SD sequence. This resulted in a gene family of 11 common paralogs. **(B)** Inversion haplotype of HG01109 hap2 (bottom) aligned to CHM13 (top). **(C)** Visual illustration of CHM13 clusters 1 and 2 with new paralog characterization, as well as expression of these paralogs across iPSCs and fetal brain Iso-Seq libraries, normalized to median haplotype paralog copy number.

2.4 DISCUSSION

Long-read sequencing and advances in *de novo* genome assembly have enabled comprehensive characterization of complex, duplicated loci (Liao et al. 2023). Here, we investigated the evolution and transcription of TBC1D3, a "hominoid-specific" gene family functionally implicated in the proliferation of neuronal progenitors and cortical expansion and folding of the human brain (Paulding et al. 2003; Sudmant et al. 2010; Ju et al. 2016; Hou et al. 2021). Using Hifiasm and Verkko, we successfully assembled and validated 69 human haplotypes from three references (GRCh38, CHM1, T2T-CHM13) and 33 human samples across TBC1D3 clusters 1 and 2 (Cheng et al. 2021; Rautiainen et al. 2023). We find that the human TBC1D3 gene family is among the most copy number-variable gene families, with >60% of human haplotypes containing a unique structural configuration at each cluster with an overall structural heterozygosity estimated at 90%. The TBC1D3 copy number at each cluster ranges from one to 14, which we phylogenetically reduced into 11 common TBC1D3 paralog groups—four of which were novel and not represented in either the GRCh38 or T2T-CHM13 human references (Fig. 2.6A; Supplemental Fig. S12).

At first glance, this incredible genetic variation of TBC1D3 conflicts with the proposed critical function in brain cortical expansion. Leveraging a deep long-read Iso-Seq data set from two developmental contexts (iPSCs and fetal brain), we distinguished paralog expression and found that TBC1D3 paralogs mapping to cluster 2, most notably TBC1D3-CDKL, account for ~90% of assigned transcripts. We hypothesize that this restricted pattern of expression may explain how such high copy number variation is tolerated, because only one or two copies, located at the telomeric end of TBC1D3 cluster 2, are exclusively expressed. This model of regulation is reminiscent of the green opsin gene family on Chromosome X, in which a single locus control region promotes expression of the most proximal green opsin paralog and downstream duplicates are transcriptionally silent (Hayashi et al. 1999). In this model, many of the other TBC1D3 paralogs are either inactive pseudogenes or "genes-in-waiting" with the potential to become the primary gene if their position within the cluster changes. Future studies investigating TBC1D3 regulation and expression, with methods such as Fiber-seq as well as matched RNA-seq and WGS samples to correlate copy number and expression, will help elucidate the regulatory landscape of the TBC1D3 gene family (Stergachis et al. 2020).

TBC1D3 is just one example of approximately two dozen core duplicons, originally defined as focal points of sequence overrepresented in SD repeat graphs (Jiang et al. 2007; Marques-Bonet and Eichler 2009; Dennis et al. 2017). Several core duplicons have been associated with recurrent and independent duplications in primates, chromosomal rearrangements among apes, large-scale inversion polymorphisms in humans, and developmental disorders (Johnson et al. 2006; Zody et al. 2006a, b; Antonacci et al. 2010; Mohajeri et al. 2016; Nuttle et al. 2016; Maggiolini et al. 2019; Porubsky et al. 2022; Mao et al. 2024). TBC1D3 is no exception. First,

we found evidence of five separate lineage-specific expansions in the different primate lineages and observed that TBC1D3 expanded specifically in humans ~2.5 mya when the genus Homo transitioned from Australopithecus, coinciding with the onset of frontal cortical expansions in Homo habilis (Spoor et al. 2015). We found a 2.2 Mbp inversion between TBC1D3 clusters in one Amerindian haplotype, consistent with ongoing nonallelic homologous recombination between inverted TBC1D3 gene clusters, which may provide a substrate for the recurrent 17q12 microdeletion syndrome associated with renal cyst and diabetes syndrome (RCAD) (Mefford et al. 2007). Finally, we found a suite of changes in the TBC1D3 protein sequence, including positively selected amino acid changes among African apes and a significantly transformed C terminus exclusive to humans. Unlike other African apes, all human TBC1D3 copies that we have detected as expressed harbor this modified C terminus, suggesting it may have been a key event underlying the potential neofunctionalization of the gene family in our lineage.

Functional investigations have suggested different biochemical roles for the TBC1D3 protein at the cellular level, all of which increase cell proliferation. Two functions occur in the cytosol, where TBC1D3 antagonizes ubiquitination and degradation of EGFR and IRS1 receptors, driving cell proliferation in cell culture (Wainszelbaum et al. 2008; 2012). The third, in contrast, proposes that TBC1D3 is shuttled to the nucleus in neuron progenitor cells, where it antagonizes EHMT2 methyltransferase and, as a result, epigenetically inhibits neural progenitor differentiation (Hou et al. 2021). Our work suggests that the extensive expansion of this gene family in humans has had limited dosage effect owing to the preferential expression/regulation of the distal cluster 2 copy. Instead, we propose that the human-specific modified C terminus plays a critical role in these adaptive functions by potentially directing novel posttranslational

modifications or altering the localization and trafficking of TBC1D3 proteins (Sharma and Schiller 2019). It will be important to compare the structure and function of human and NHP TBC1D3 proteins to determine if neofunctionalization has indeed occurred as a result of these changes in the human lineage. The power of long-read sequencing to resolve structural variation, expression, and regulation of complex gene families such as TBC1D3 makes these fundamental questions addressable.

2.5 METHODS

2.5.1 LONG-READ SEQUENCE AND ASSEMBLY

The majority of genomes used in this study were sequenced previously as part of other assembly efforts to generate phased genomes or T2T genomes and are publicly available under NCBI BioProject (https://www.ncbi.nlm.nih.gov/bioproject/) accession numbers PRJNA941350, PRJNA877605, PRJNA941358, PRJNA916732, PRJNA916733, PRJNA916735, PRJNA916734, PRJNA916736, and PRJNA916737 (Liao et al. 2023; Makova et al. 2024; Mao et al. 2024). For species, coverage, and project details, see Supplemental Table S10. This study focused only on analyzing sequence contigs that contained copies of TBC1D3 paralogs, and we evaluated each contig for gaps and contiguity (see Assembly validation section below). Most human genomes were originally assembled using Hifiasm (version 0.15.2), but TBC1D3-containing contigs that failed QC were reassembled with Verkko (versions 1.0, 1.1, 1.2, and 1.4) using a combination of both HiFi and ONT sequence. In general, haplotypes were phased using parental k-mer information when available, or Hi-C chromatin capture data (Auton et al. 2015; Kronenberg et al. 2021). For the Chromosome 17 comparison, it was observed that the macaque

orthologous chromosome was fragmented and was subsequently scaffolded using RagTag (version 2.1.0) with the Mmul10 reference as the scaffold (Hughes et al. 2012; Alonge et al. 2022). In this study, we generated assemblies for only two species: gelada (Theropithecus gelada) and mouse lemur (M. murinus). High-molecular-weight DNA was prepared from peripheral blood of a male gelada (DRT_2020_14_TGE) and from skin fibroblasts of a female mouse lemur (Inina_MMUR). HiFi sequence data (50×, 30×) were generated using the Sequel II platform, and assemblies were generated with Hifiasm (Supplemental Table S10).

2.5.2 ASSEMBLY VALIDATION

2.5.2.1 ILLUMINA COPY NUMBER VALIDATION

Sample assemblies were first validated using diploid assembly TBC1D3 copy number estimates to Illumina sequence copy number estimates, an orthogonal sequencing approach (Supplemental Fig. S2). Sample genome haplotypes were merged and k-merized into 32 bp k-mers using Meryl (version 1.3) (Rhie et al. 2020). In parallel, sample Illumina sequence libraries were similarly k-merized into 32 bp with Meryl. Next, k-mer libraries were aligned to the T2T-CHM13 reference genome using FastCN, allowing for up to two mismatches between the k-mer and assembly alignments (Pendleton et al. 2018; Nurk et al. 2022). We estimated the copy number of TBC1D3 by taking the average copy number over one TBC1D3 paralog, TBC1D3L, and compared these estimates against one another in a scatter plot (Fig. 2.1A; Supplemental Fig. S2).

2.5.2.2 SELF-READ MAPPING VALIDATION

We also applied NucFreq (Vollger et al. 2019) to assess the integrity of each TBC1D3 assembly. Each sample's respective HiFi sequencing libraries were trio phased using Canu (version 2.1.1) (Koren et al. 2017) and mapped back onto their respective *de novo* assemblies. To qualitatively

validate assembly, we plotted the sequence depth of both the primary and secondary bases of reads aligned over the TBC1D3 expansions (Supplemental Fig. S1). First, we removed samples with obvious gaps over the TBC1D3 expansion 1 and 2 loci, which could be identified if the locus was broken across multiple contigs or if the assemblies had a lack of HiFi sequence support over a given region. Next, we identified assemblies with collapses over the TBC1D3 expansion 1 and 2 regions by looking at secondary base read depth. HiFi sequencing is 99.9% accurate, with occasional low-frequency false base calls. Our expectation is that this frequency can be observed over a given region as the secondary base, remaining well below 1% frequency. Any haplotypes with a noticeable increase in secondary base frequency over particular stretches were marked as collapsed. Usually, these samples included a spike in primary base coverage as well as over the collapsed region. Additionally, Hifiasm samples were validated with GAVISUNK (Dishuck et al. 2023). Phased ONT reads were mapped over each sample's respective assemblies, and singly unique nucleotide k-mer anchors were marked. We expect, for correct assemblies, that every region of the assembly will be supported by at least one ONT sequence, which is not used during Hifiasm assembly. Any locations with a gap in ONT assemblies were marked as not validated.

2.5.3 REPEAT AND GENE MAPPING ANNOTATION

We defined repeat content in the genome using Tandem Repeat Finder (TRF) (version 4.09; Benson et al. 1999) for simple tandem repeats, RepeatMasker (version 4.1.2-p1; http://www.repeatmasker.org) for common transposon and retrotransposon elements, and DupMasker to define duplicons associated with human SDs (Jiang et al. 2008). TBC1D3 loci were identified in the GRCh38 reference genome based on RefSeq annotations and mapped to

other assemblies using minimap2 (version 2.24), using the asm20 standardized setting and allowing for up to 1000 secondary alignments (Li 2018). These mappings were filtered to contain at least 6 kbp of sequence over half the length of the canonical TBC1D3 gene model. For more distantly related lineages, including the New World monkeys, we mapped TBC1D3 sequence using BLAT (version 3.5), allowing a maximum intron length of 5 kbp, half the TBC1D3 gene model length, and a minscore of 100. These relatively loose mapping constraints identified many candidate TBC1D3 paralogs, more than expected by either Illumina- or assembly-based TBC1D3 copy number estimates, that were subsequently filtered based on expression, divergence, or minimum length match.

2.5.4 STRUCTURAL VARIATION AND HETEROZYGOSITY CHARACTERIZATION

Validated cluster 1 and 2 TBC1D3 haplotypes were aligned to one another in an all-by-all fashion using minimap2 (version 2.24) auto settings -x asm5, allowing up to 1 kbp of insertions in cigar strings. We labeled two haplotypes as structurally equivalent if ≥90% of their sequence could be mapped to one another in a single alignment. We repeated this exercise for all pairs of haplotypes, calculated the number of valid haplotypes with no structurally equivalent pair, and divided by the total number of validated haplotypes to determine our structural variation statistic. For structural heterozygosity, we identified all samples whose two haplotypes were not structurally equivalent and divided by the total assembled samples. Contig and chromosome alignments (e.g., Figs. 2.3 and 2.5) were visualized by SVByEye using either plotMiro for pairwise alignment, or plotAVA for all-versus-all alignments (https://github.com/daewoooo/SVbyEye). Blue alignments represent directly orientated alignments, and yellow indicates inverted alignments. For local TBC1D3 structure comparison

(Supplemental Fig. S4), we extracted primate TBC1D3 copies, along with 25 kbp of flanking sequence, from five primate lineages and mapped to one another. These copies were organized to reflect the closest alignments, by both length and identity.

2.5.5 *TBC1D3* BREAKPOINT SIMULATION

We mapped orthologous Chromosome 17 relationships and annotated TBC1D3 copies using minimap2 -x asm20. Synteny was annotated using Asynt get.synteny.blocks.multi command, with max_gap = 200,000, min_block_size = 1,000,000, and min_subblock_size = 50,000, producing a tab-delimited file marking the target and query breaks of blocks (Kim et al. 2022). For each TBC1D3 copy, we identified the nearest synteny break along the respective chromosome and then computed median distance to synteny breaks of all TBC1D3 mappings. Next, we conducted a permutation experiment. For each primate orthologous Chromosome 17, we randomly selected ~11 kbp blocks at the same quantity as the number of TBC1D3 mappings observed in the respective primate chromosome. We repeated the median distance experiment and plotted the distribution of 5000 permutations.

2.5.6 MULTIPLE SEQUENCE ALIGNMENT

Sequence was extracted from assemblies by mapping TBC1D3 sequence to full genome assemblies with minimap2 (version 2.24) and extracting the mapped reference sequence with BEDTools (version 2.29.2) (Quinlan and Hall 2010; Li 2018). MSAs were constructed with MAFFT with parameters --reorder --maxiterate 1000 --thread 16 (version 7.453) (Katoh et al. 2002). Following MSA construction, spurious alignments were pruned with trimmal (--gappyout; version 1.4) and manually trimmed. Codon alignments were generated with matched ORF and

amino acid sequence FASTA files. First, an amino acid MSA was generated with MAFFT, and then the ORF FASTA was aligned to the amino acid MSA with pal2nal (Suyama et al. 2006).

2.5.7 PHYLOGENETIC ANALYSES

Maximum likelihood phylogenies were generated with iqtree2 using model setting -m MFP, 1000 lrt replicates, and -b 1000 replicates for bootstrap (version 2.1.2). Additionally, each phylogeny generated was outgrouped to a sequence: marmoset TBC1D3 for primate phylogenetic analysis and chimpanzee TBC1D3 for human paralog clustering. Phylogenetic trees were illustrated in R with ggtree (Yu 2023). Timing estimates for individual primate expansions were conducted using BEAUTi for data input and BEAST2 for computation (Drummond et al. 2012; Bouckaert et al. 2019). We used human–macaque and human–chimpanzee divergence times of 25 and 6.5 mya, estimated by the fossil record, as benchmarks for the computation (Dunsworth 2010; Stevens et al. 2013). With these references, we calculated the 95% confidence intervals of mutation rate within sequences and then estimated species-specific expansions with this mutation rate as well as branch lengths of the primate phylogeny. For tests of positive selection, we isolated intronic sequence and exonic sequence from paralog isoforms with expression support from the human, chimpanzee, bonobo, gorilla, Sumatran orangutan, and Bornean orangutan genome assemblies.

We tested for positive selection in coding sequence using both the PAML package and absREL (Yang 2007; Smith et al. 2015). We focused on TBC1D3 paralog isoforms for which there was evidence of transcription based on Iso-Seq FLNC analysis from the human, chimpanzee, bonobo, gorilla, Sumatran orangutan, and Bornean orangutan samples. To serve as a proxy for neutral evolution, we isolated 7245 bp of intronic sequence from each expressed paralog and generated

an MSA and maximum likelihood phylogeny, with orangutan TBC1D3 copies as our outgroup. In parallel, we extracted 1884 bp of exonic sequence, predicted amino acid sequence with ORFipy, and codon-aligned exonic sequence with Pal2Nal (Suyama et al. 2006). With the intronic phylogeny and codon-aligned MSA, we identified branches undergoing accelerated evolution with a free-ratios model, in which independent dN/dS values are computed for each branch in the tree (Yang 2007). We ignored predicted dN/dS values for terminal branches, as too few changes occurred, and they were underpowered to detect selection. Among deeper branches, we identified three that were predicted to be under selection, as discussed in the text. We more stringently tested these three branches with the absREL test hosted on hyphy, which infers the optimal number of omega values and tests branches under positive selection with a likelihood ratio test statistic (Supplemental Table S8; https://stevenweaver.github.io/hyphy-site/methods/selection-methods/). After multiple test corrections, we identified one branch under positive selection. For site-level resolution, we isolated this branch in a branch-site model test and selected the amino acid residues under selection using the Bayes empirical Bayes posterior probability (Yang et al. 2005).

2.5.8 ISO-SEQ AND TRANSCRIPT ANALYSES

Primate Iso-Seq testis data were generated by Makova et al. (2024) and made available from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession numbers SRX18421140, SRX18280098, SRX18280097, SRX19199753, SRX19199753, and SRX18421141. Similarly, human iPSC Iso-Seq was previously generated by Cheung et al. (2023) and made available from the database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) under accession number phs002206.v4.p1. Fetal brain tissue was derived from 59 spontaneously aborted fetuses with sequence available from SRA under accession number SRR28199631. This sequence was enriched

for both TBC1D3 and NPIPA1, using the hybridization capture protocol described by Dougherty et al. (2018), with probes provided in Supplemental Table S11. FLNC libraries were mapped to respective species libraries with minimap2 using the parameters -ax splice --sam-hit-only --secondary = yes -p 0.5 --eqx -K 2G -G 8k -N 20. FLNC libraries were first filtered for reads ≥1000 bp in length and with sequence quality of ≥99.9%. Each library was subsequently mapped to the genome assembly corresponding to the respective species of origin using SAMtools (Danecek et al. 2021) and BEDTools. Next, we determined which TBC1D3 paralogs were likely expressed by selecting paralogs with read support with mapping quality ≥99.9% sequence identify. These reads were subsequently reduced into common isoforms with IsoSeq3 (4.0.0, PacBio; https://github.com/ylipacbio/IsoSeq3) collapse, and ORFs were predicted with Orfipy (Singh and Wurtele 2021). For primate TBC1D3 gene model comparison, isoforms with at least three independent reads of support and with the longest maintained ORF were compared. We required these reading frames to span within 100 bp of the canonical TBC1D3 start and stop as defined by RefSeq (O'Leary et al. 2016). Human FLNC reads from fetal brain and iPSCs were mapped to all validated human haplotypes. Next, we compared these primary alignments to one another and considered the cluster paralog from which they were derived. Any Iso-Seq read with primary minimap2 alignment scores of 10 or greater for a given paralog cluster relative to all other cluster mappings was retained, whereas other mappings were marked as ambiguous and ignored.

2.5.9 ANALYSIS OF CODING SEQUENCE

To validate the observed deletion of coding sequence in humans, we selected human TBC1D3L amino acid sequence and mapped this sequence to all genome assemblies with ProSplign (https://www.ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html), a tool that predicts DNA sequence representing the codons for a given protein amino acid sequence. ProSplign predicts splice junctions, as well as start and stop codons, and illustrates amino acid substitutions, frameshift mutations, and deletions in the underlying nucleotide sequence that are inconsistent

with the provided amino acid sequence. We predicted the human TBC1D3 tertiary structure using the EMBL-EBI AlphaFold2 database (Jumper et al. 2021; https://alphafold.ebi.ac.uk/). The predicted tertiary structure was illustrated using PyMOL (2.0, https://www.pymol.org).

2.5.10 HUMAN PANGENOME GRAPH CONSTRUCTION

We built a pangenome graph of TBC1D3 with minigraph (version 0.20; Li et al. 2020), with the settings -S -xggs -L 250 -r 100000 -t 16. We attempted graph construction with lower -l and -g settings as well but consistently observed that most haplotype TBC1D3 paralogs were isolated to nodes without any allelic overlap from other human haplotypes.

2.5.11 HUMAN *TBC1D3* PARALOG GROUPING

We generated a phylogeny with the whole TBC1D3 sequence for all cluster 1 and 2 copies identified in validated human assemblies, outgrouped to chimpanzee TBC1D3. We defined a heuristic cutoff based on allelic variation to define our clusters. Vollger et al. (2023) previously predicted allelic variation of 15.3 single-nucleotide variants per 10 kbp. We recursively identified clades with an intra-variation of up to 1.5 times the allelic variation identified in SDs. Additionally, we required that a given cluster have at least 10 independent paralogs of representation to be defined as a population-level paralog group.

2.6 DATA ACCESS

Gelada sequence and assembly data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers

PRJNA1081468 and PRJNA1081469. Mouse lemur sequence and assembly data generated in this study have been submitted to the NCBI BioProject database under accession numbers PRJNA1082315 and PRJNA1082316. Assembled contigs corresponding to the TBC1D3 genomic regions for both the gelada and mouse lemur are also available at Zenodo (https://doi.org/10.5281/zenodo.12808906). Gelada and mouse lemur sequencing data used for these assemblies have been submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession numbers SRR28199625–SRR28199630 and SRR28217961–SRR28217966, respectively. Fetal brain Iso-Seq data generated in this study have been submitted to the BioProject database under accession number PRJNA659539 and are available from the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRR28199631.

2.7 ACKNOWLEDGEMENTS

This work was supported, in part, by U.S. National Institutes of Health (NIH) grants HG002385, HG010169, and HG007497 to E.E.E. We thank Noah Snyder-Mackler, Kenny Chou, and the Simien Mountains Gelada Research Project for providing peripheral blood for the sequence and assembly of the T. gelada genome. We thank Mark Krasnow for access to material from the mouse lemur (M. murinus). We thank the Primate T2T Consortium, especially Kateryna Makova and Adam Phillippy, for providing us with early access to the high-quality ape genome assemblies. We thank Tonia Brown, Zoe Poyen, and Gerta Janss for manuscript proofreading and editing. We thank generous donors to Children's Mercy Kansas City and Genomic Answers for Kids program supporting the human iPSC Iso-Seq (T.P.). E.E.E. is an investigator of the

CHAPTER 3. AN *NPEPPS* SEGMENTAL DUPLICATION DRIVES POSITION EFFECT EXPRESSION OF *TBC1D3* IN THE HUMAN BRAIN

Chapter 3 is adapted with minimal modification from:

Guitart, X., Brunner, J. W., Ren, L., Jeong, H., Yoo, D., Porubsky, D., Hoekzema, K., Munson, K. M., Sun, K. A., Ayllon, M., Hoglin, K., McMullen, R., Pavlovic, B., Vollger, M. R., Pollen, A. A., & Eichler, E. E. (2026). *NPEPPS* segmental duplication drives position effect expression of *TBC1D3* in the human brain (p. 2026.01.14.699559). bioRxiv. https://doi.org/10.64898/2026.01.14.699559

Author Contributions: X.G. and E.E.E. conceived the project; X.G. performed the analyses relevant to all manuscript figures; J.W.B. generated and assisted in analysis of data related to human, chimpanzee, and orangutan neuron development model; L.R. conducted synonymous, missense, and nonsense analysis illustrated in Supplementary Fig. 9B,C; H.J. and D.Y. assisted in analysis of human, chimpanzee, and orangutan neuron development model; D.Y. also assisted in characterization of chromosomal inversions described in Fig. 3C and Supplementary Figs. 4-6; N.K. assisted with characterization of inversion dbSNPs in UK Biobank; D.P. assisted in characterization of human haplotypes structural diversity; M.R.V. assisted in analysis of methylation characterization of CHM13 gDNA. K.Hoekzema sequenced all ONT data. K.M.M., K.A.S., and M.A. prepared and sequenced HiFi data used in neurospheres and the comparative neuronal developmental model analyses. K.Hoglin, R.M., and B.P assisted in cell culture and experimentation of the comparative neuronal developmental model. A.A.P. assisted in the conception of the comparative neuronal developmental model. X.G. and E.E.E. wrote the manuscript with input from all authors.

3.1 ABSTRACT

In humans, the *TBC1D3* gene family is thought to play a critical role in the expansion of the frontal cortex by promoting neuronal proliferation during brain development. This gene family shows some of the greatest structural heterozygosity (~97%) with haplotype copy numbers ranging from 3-39 among different human haplotypes. This raises the question as to how a gene so crucial in the evolutionary expansion of the human frontal cortex can be so variable in the human population. Here, we characterize the regulatory architecture that explains this paradox. We show that 45-96% of *TBC1D3* expression is attributable to a single paralog located at the most telomeric position at the edge of a cluster of *TBC1D3* genes. We find that its >3-fold higher expression relative to other copies is driven by a 110 kbp segmental duplication that occurred ~8.9 million years ago, relocating a partial duplication of the puromycin-sensitive aminopeptidase gene (*NPEPPS*), including its promoter, adjacent to this *TBC1D3* locus. Using neurospheres and comparative transcriptomics of iPSC-derived cultures, we show that expression of *NPEPPSP1-TBC1D3* increases as neurons differentiate as a result of alternative splicing and differential polyadenylation usage. While the fusion exists in other ape lineages, we show subsequent deletion of the *NPEPPSP1* promoter in *Gorilla* and a separate, lineage-specific duplication in the *Pan* lineage ablated the production of this fusion product, rendering this position effect of *TBC1D3* specific to humans.

3.2 INTRODUCTION

TBC1D3 is a primate-specific gene family that promotes cellular proliferation in both neurodevelopment and cancer. Originally identified in prostate and breast cancer tumors, functional experiments determined that *TBC1D3* promotes cell transformation and proliferation by manipulating cell vesicle transport, specifically amplifying the effect and duration of insulin-like growth factor (IGF) and epidermal growth factor (EGF) pathway-associated receptors (Frittoli et al. 2008, Wainszelbaum et al. 2008 & 2012). While its cell proliferation is co-opted in cancers, endogenous *TBC1D3* plays a developmental role by potentially expanding neural progenitor populations in the primate cerebral cortex, particularly in outer radial glia (Ju et al. 2016). Subcellular localization studies suggest that the TBC1D3 protein generally localizes to the cytoplasm, but translocates to the nucleus in neuronal contexts, where it inhibits the histone methyltransferase G9a (Hou et al. 2021). This repression reduces H3K9me2-mediated gene silencing and delays differentiation of outer radial glial cells, leading to additional division rounds and potentially contributing to the exponential increase in cortical neurons associated with human brain development.

The corresponding *TBC1D3* gene family is embedded in a dynamic core duplicon and has expanded recently during primate evolution via segmental duplications (SDs) across chromosome 17 (Jiang et al. 2007). While many orphan *TBC1D3* paralogs are thought to be nonfunctional pseudogenes, in humans, transcribed *TBC1D3* copies that maintain an open reading frame (ORF) originate primarily from two major clusters, Cluster 1 and Cluster 2, mapping to chromosome 17q12 (Guitart et al. 2024). These SD clusters also promote non-allelic homologous recombination, leading to a recurrent 17q12 deletion/duplication (known as renal cyst and diabetes

or RCAD syndrome) (Mitchel & Moreno-De-Luca et al. 1993; Mefford et al. 2007, 2016). Both the duplication and deletion syndromes manifest heterogeneously in patients, though the deletion syndrome is more often associated with abnormalities of the renal and endocrine systems. In contrast, the duplication syndrome is associated with both neurodevelopmental and psychiatric abnormalities. While 75% of deletion events are *de novo*, 90% of duplications are inherited, suggesting that duplications may cause less severe phenotypes and thus persist in the general population for a few generations (Mitchel & Moreno-De-Luca et al. 1993; Mefford et al. 2016).

The origin and expansion of *TBC1D3* are complex: *TBC1D3* was proposed to originate from *USP6NL* (i.e., *RNTRE*) based on conserved exon structures across 13 of its 14 exons. However, at the amino acid level, TBC1D3 and USP6NL proteins share only 34% identity, suggesting extensive amino acid replacement over a short period of primate evolution (Frittoli et al. 2008). In our prior work, we characterized the evolution and copy number diversity of the *TBC1D3* gene family among primates (Guitart et al. 2024). We showed *TBC1D3* expanded independently across seven separate lineages of the simian infraorder but is generally absent from prosimians, also known as *Strepsirrhines*. In humans, *TBC1D3* is estimated to have expanded to include a human-specific modification of the predicted protein ~2 million years ago (MYA). This expansion or gene conversion coincides with the emergence of *Homo erectus* (Bar-Yosef et al. 2001).

All human-expressed *TBC1D3* copies carry a human-derived 58 amino acid modification of the carboxy-terminus, suggestive of a human-specific neofunctionalization (Guitart et al. 2024). Additionally, the gene family continues to evolve dynamically in the human population. Genome-wide studies have shown that *TBC1D3* is among the most copy number polymorphic and heterozygous gene families in humans, with copy numbers ranging from 3 to 39 per haplotype (Sudmant et al. 2015; Guitart et al. 2024). Using long-read transcriptome sequencing to assign

isoforms to specific paralogs, we showed that despite this variability, >90% of *TBC1D3* transcription from human fetal brain and induced pluripotent stem cells (iPSCs) maps to Cluster 2; over 80% of those transcripts stem from the last *TBC1D3* copy mapping distally at the edge of Cluster 2. We hypothesized that a position effect was responsible for regulating and restricting expression to the edge of Cluster 2.

To test this hypothesis and understand the nature of this potential position effect, we performed a detailed comparative transcriptomic and epigenetic analysis to investigate the regulation of *TBC1D3* in humans and great apes. We find that the Cluster 2 terminal *TBC1D3* paralog dominates expression as a result of fusion with an upstream duplicated gene, *NPEPPSP1*, including its brain-enriched promoter. This *NPEPPS1-TBC1D3* fusion transcript accounts for the majority of *TBC1D3* expression across various tissues. We show that the fusion is mediated by segmental duplication of *NPEPPS* in the common ancestor of African great apes but was lost in the chimpanzee, bonobo, and gorilla lineages by subsequent gene conversion and deletion events, respectively. Additionally, although both *NPEPPSP1* and *TBC1D3* are copy number polymorphic in humans, we find that the fusion gene promoter is fixed across all human genomes examined. Our findings dissect the evolution of the *TBC1D3* position effect and explain how it has become specific to the human lineage, providing a model for how new genes rapidly emerge in a species.

3.3 RESULTS

3.3.1 MATCHED *TBC1D3* COPY EXPRESSION AND REGULATION IN CHM13.

The high sequence identity (>99%) of *TBC1D3* paralogs has meant that the transcription and regulation of this gene family has essentially been excluded from large-scale studies such as ENCODE (ENCODE Project Consortium, 2012). To resolve this, we initially used matched long-read sequencing (LRS) data from the complete human reference genome (T2T-CHM13) to investigate the methylation profile of the nine *TBC1D3* copies present in this haploid source genome (Fig. 3.1A; Nurk *et al*. 2022; Methods). This analysis revealed a striking hypermethylation signature across the gene body of the terminal paralog of Cluster2, *TBC1D3-CDKL2*, whereas all other *TBC1D3* paralogs displayed uniformly low methylation signals, comparable to background levels observed for lowly expressed genes, including SD paralogs (Vollger et al. 2022). Thus, the specific hypermethylation signatures of the gene body *TBC1D3-CDKL2* aligns with one of the canonical epigenetic hallmarks of expressed genes.

To identify the putative promoter driving expression of this distal *TBC1D3-CDKL2* copy, we analyzed the flanking 100 kbp region upstream of its transcription start site (TSS) for evidence of transcription initiation. An analysis of short-read ATAC-seq data and annotated ENCODE cis-regulatory elements in the GRCh38 reference suggested a putative open chromatin accessibility region located 66 kbp from the translation initiation site of *TBC1D3* (Supplementary Fig. S1; ENCODE Project Consortium, 2012). To define the full-length gene structure based on empirical data rather than annotation alone, we mapped full-length Iso-Seq transcriptomic data from human prefrontal cortex (Fig. 3.1B). Remarkably, 43% (19/44) of transcripts did not align to the canonical RefSeq model but instead represented a fusion isoform with an annotated pseudogene,

NPEPPSP1, located directly upstream (Fig. 3.1C). Tracing this dominant isoform to its TSS revealed a distinct dip in methylation immediately upstream of the first exon—a clear hypomethylated promoter signature, consistent with active transcription initiation (Gershman et al. 2022; Vollger et al. 2022).

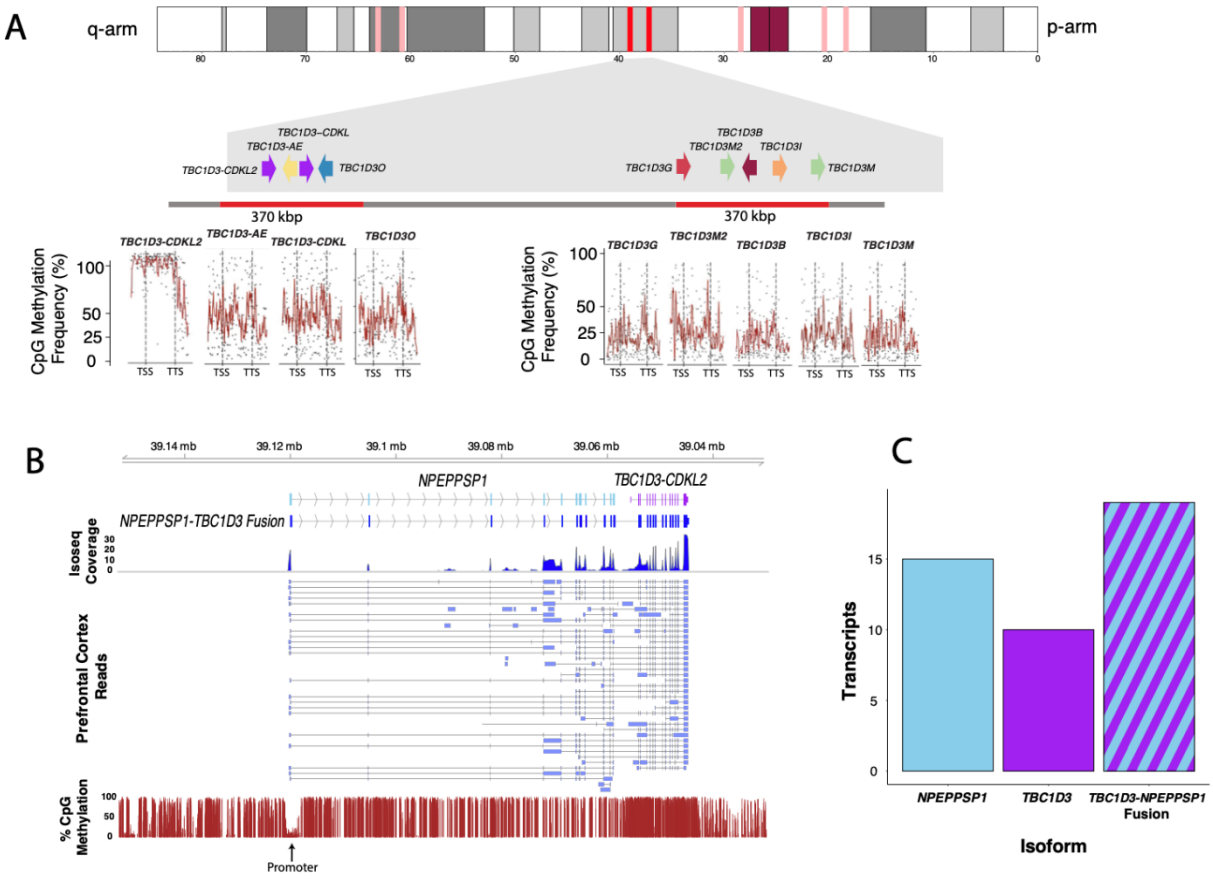


Figure 3. 1 Differential regulation and expression of *TBC1D3* gene family

(A) CpG methylation profile compared among *TBC1D3* paralogs within CHM13 based on mapping of CHM13 ONT data back to the T2T-CHM13 genome (Nurk *et al.* 2022). Percent methylation of CpG sites from transcription start site (TSS) to transcription termination site (TTS) based on rolling average of 15 consecutive CpG sites (red lines). This methylation signature is consistent with >75% of transcription originating from this distal copy (Guitart et al. 2025). **(B)** Expression and promoter definition of the terminal *TBC1D3-CDKL2* copy. A schematic of the *NPEPPSP1-TBC1D3* fusion gene structure, including overall Iso-Seq transcript coverage per exon (top panel) and various isoforms (middle panel) detected based on

Iso-Seq transcript data from human prefrontal cortex (Leung et al. 2021). Percent CpG methylation data from CHM13 overlaid with this annotation predicts the likely location of the promoter identified by the dip in methylation (bottom panel). **(C)** The absolute number of Iso-Seq transcripts identified as *NPEPPSP1-TBC1D3* fusion vs. solo *TBC1D3* or *NPEPPS1* compared for Iso-Seq data from CHM13.

3.3.2 FUSION EXPRESSION AND REGULATION IN *IN VITRO* DEVELOPING BRAIN MODEL.

While CHM13 proved useful to discover the potential site of transcription initiation, this developmental abnormality—a hydatidiform mole—is a poor proxy for the human developing brain, the principal tissue of *TBC1D3* function (Ju et al. 2016; Hou et al. 2021). To investigate *TBC1D3* regulation in a context more relevant to brain development, we first interrogated human neurospheres, an *in vitro* model for neural differentiation harboring both neural progenitor cells (NPCs) and neural stem cells (Fig. 3.2A; Real et al. 2025). We profiled both chromatin accessibility and transcriptome activity of the *TBC1D3* gene family in neurospheres derived from the HPRC sample HG02630.

Consistent with CHM13 methylation data, this analysis revealed that none of the *TBC1D3* paralogs showed evidence of a proximal promoter—defined as a peak of accessibility directly upstream of the TSS of the canonical gene model (Supplementary Table 1; Stergachis et al. 2020). Instead, the matched transcriptomic-DNA genome assembly data derived from the same neurosphere source material confirmed the initial observation from the analysis of the prefrontal cortex library; 56% (116/206) of primary *TBC1D3* transcript mappings were transcribed as a fusion product with *NPEPPSP1*, over three times the expression than any other individual *TBC1D3* paralog. Using Fiber-seq Inferred Regulatory Element (FIRE) with FIREtools (Fig. 3.2B; Vollger et al. 2025), we defined the promoter element at the *NPEPPSP1* TSS, a 302 bp unit that regulates expression of

the *TBC1D3* paralog located directly downstream of *NPEPPSP1*. Combined, these data suggest that cooption of the *NPEPPSP1* promoter plays a significant role in generating neuronal transcripts of *TBC1D3*, especially from the most distal copy of the *TBC1D3* Cluster 2 (Fig. 3.2B,C).

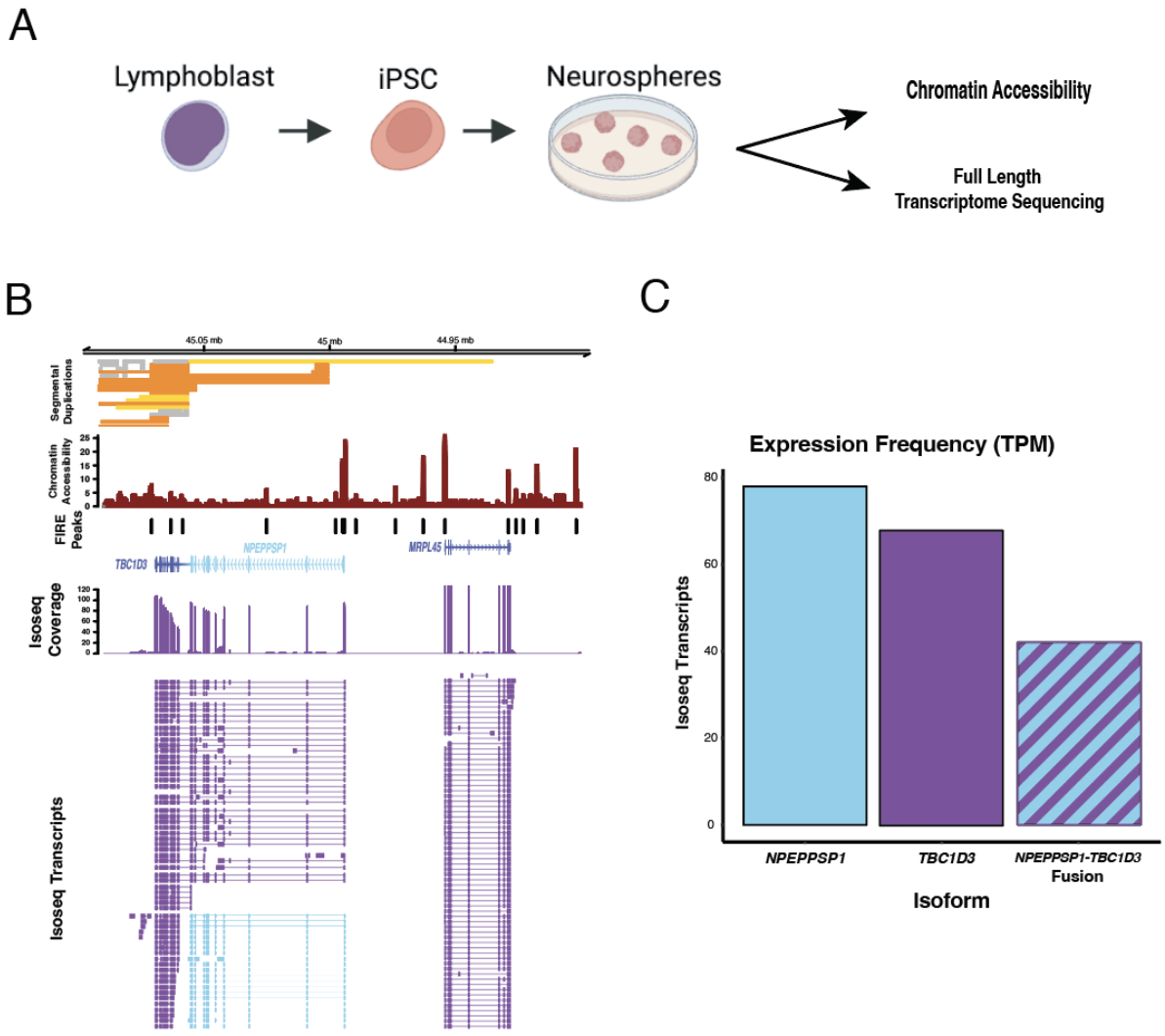


Figure 3. 2 Transcription and chromatin accessibility in human neurospheres.

(A) Schematic depicting neurosphere generation. Lymphoblasts from the 1000 Genomes Project (1KGP) sample, HG02630, were induced to create iPSCs and subsequent neurospheres. Matched Iso-Seq and Fiber-seq data were generated from neurospheres to assess the transcriptome and chromatin accessibility. **(B)** Expression and chromatin accessibility of *NPEPPSP1-TBC1D3*. Full-length transcripts mapped to the HG02630 diploid *de novo* sequence assembly are shown with alignment coverage, Fiber-seq accessibility,

and segmental duplication sequence annotated with SEDEF (Numanagic et al. 2018). **(C)** Absolute expression of *NPEPPSP1*, *TBC1D3*, and the *NPEPPSP1-TBC1D3* fusion. Transcript reads are classified into *NPEPPSP1*, *TBC1D3*, and *NPEPPSP1-TBC1D3* fusion categories.

3.3.3 JUXTAPOSITION OF NOVEL REGULATORY DNA BY SEGMENTAL DUPLICATION OF *NPEPPS*.

To understand the evolutionary origin of the regulatory sequence driving *NPEPPSP1-TBC1D3* expression, we examined its genomic context in the HG02630 diploid assembly. Using SD annotations by SEDEF (Numanagic et al. 2018), we conclude that the *NPEPPS1* duplication and its associated promoter arose as a 119,358 bp SD originating from *NPEPPS* located 9.16 Mbp distally on chromosome 17q21 (Fig. 3.3). A comparison of these two paralogous regions along with corresponding gene annotation of the ancestral locus shows that the promoter sequence of *NPEPPS* (N-peptide puromycin sensitive) matches (Fig. 3.3A) the promoter sequence associated with *NPEPPS1-TBC1D3* fusion transcripts defined by LRS (Fig. 3.2). The ancestral *NPEPP*S encodes a zinc metallopeptidase broadly expressed across tissues, and mutations of it have been associated with neurodegenerative tauopathies and Parkinsonism (Henderson Front Gent 2021; Karsten et al. 2006). Thus, the two paralogous promoter pairs regulate expression of *NPEPPS* and *NPEPPSP1*, and *MRPL45* and *MRPL5P1*, respectively (Fig. 3.3A). The duplicated region shows an average of 98.2% sequence identity by matches only, suggesting a recent origin during ape evolution.

In order to more accurately estimate the timing of the duplication, we extracted a 15 kbp region from the *NPEPPSP1*-anchored end of the SD, directly adjacent to *TBC1D3*. We used this sequence to identify homologous sequences in human, chimpanzee, gorilla, orangutan, and macaque genome assemblies (Yoo et al. 2025). Our analysis shows that the 119 kbp *NPEPPS* SD is duplicated in all

African apes assessed (human, chimpanzee, bonobo, and gorilla) (Fig. 3.3; Supplementary Figs. S2, S3), but is present as a single locus in orangutan and macaque, corresponding to *NPEPPS*. This confirms *NPEPPS* as the ancestral locus and that the duplication (*NPEPPSP1*) arose in the common ancestor of all African great apes. We also constructed a maximum likelihood phylogenetic tree by generating a multiple sequence alignment of all copies among humans and apes using macaque as an outgroup (Fig. 3.3B). Our analysis estimates, with 100% bootstrap support, that the duplication occurred in the common ancestor of African great apes approximately 8.49 MYA (95% confidence interval of 8.00–9.42 MYA) (Fig. 3.3B). A comparative analysis of primate genomes (Supplementary Fig. 3.4) reveals that the *NPEPPS1-TBC1D3* juxtaposition defines the boundary of a series of large inversions, including potential inversion toggling events that break the synteny between the African great apes and other nonhuman primate species.

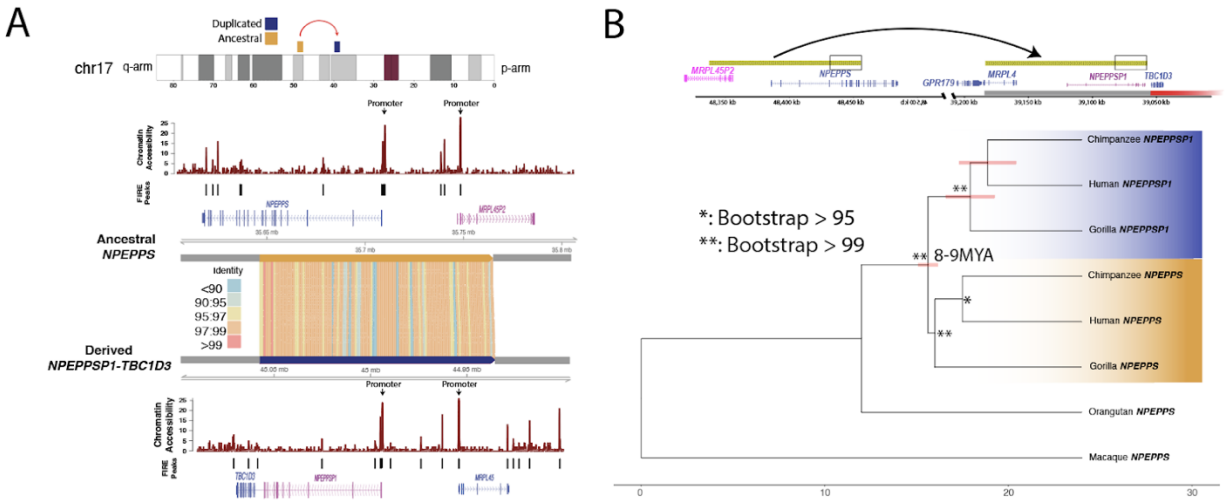


Figure 3. 3 Evolutionary origin of NPEPPSP1 regulatory sequence

(A) Comparison of the sequence and regulatory landscape of *NPEPPS* and *NPEPPSP1*. SVbyEye reveals a ~119 kbp SD with 98.2% sequence identity between the ancestral, *NPEPPS*, and duplicated, *NPEPPSP1*, loci. Fiber-seq data analyzed using FIRE compared chromatin accessibility for *NPEPPS* (above) to *NPEPPS1-TBC1D3* (below) with the most likely site of the promoter (indicated).

(B) *NPEPPS1* segmental duplication evolution. A phylogeny of 15 kbp of the *NPEPPS* duplication most proximal to *TBC1D3* and macaque (25 MYA divergence) predicts that the duplication occurred ~8-9 MYA.

3.3.4 COMPARATIVE TRANSCRIPTOMICS IN DIFFERENTIATING NEURONS.

To investigate how *NPEPPSP1-TBC1D3* fusion expression changes across neuronal development, and to compare this regulation between humans and other apes, we designed a three-stage cell culture experiment using fibroblast-derived iPSCs from human, chimpanzee, and orangutan (the latter serving as an outgroup; Fig. 3.4A). Cells were induced sequentially from iPSCs into NPCs, and finally mature neurons with the identity confirmed by marker analyses (Jeong et al. unpublished). At each stage, we collected both genomic DNA (gDNA) and full-length transcriptomes using HiFi sequencing and Kinnex Iso-Seq, respectively. We also generated a donor-specific genome assembly (DSA) for each sample using iPSC gDNA and included methylation calling during base calling to track regulatory changes throughout development. This resource allowed us to match transcript, methylation, and genomic data to specific paralogs without reference bias or cross-mapping among high-identity paralogs.

In humans, we observed stable *NPEPPSP1* expression across all three proxy developmental stages, accompanied by a stepwise increase in *TBC1D3* expression from iPSCs to mature neurons (Fig. 3.4B). This increase of *TBC1D3* expression was matched by a correlated increase in the *NPEPPSP1-TBC1D3* fusion transcript levels, increasing from a single identifiable fusion transcript in iPSCs to 72% (18/25) and 47% (98/208) of *TBC1D3* expression in NPCs and neurons, respectively. As expected, the fusion transcript was driven by the predicted juxtaposed promoter, characterized by the methylation dip at the *NPEPPSP1* TSS, consistent with the FIRE-defined regulatory elements observed in HG02630 (Fig. 3.3A; Vollger et al. 2025).

In contrast, chimpanzee *TBC1D3* expression is minimal throughout our model of neuronal differentiation. While both lineages show activation as well as a stepwise increase of the same orthologous *NPEPPSP1* promoter as cells differentiate to neurons, in chimpanzees that stepwise increase in *NPEPPSP1* expression across development is not associated with *TBC1D3* but rather a different gene, *CCL4L2* (CC chemokine ligand 4), a protein associated with inflammation and immune response (Li et al. 2023; Xu et al. 2024). Two key observations help explain the absence of *NPEPPS1-TBC1D3* expression in chimpanzee. First, the downstream *TBC1D3* paralog in the chimpanzee haplotype is inverted relative to its ORF, precluding *NPEPPSP1-TBC1D3* gene expression. Second, we identified a 44 kbp insertion of *CCL4L2* situated between the *NPEPPS1* promoter and *TBC1D3*, resulting in abundant *NPEPPSP1–CCL4L2* fusion transcription in chimpanzees (Fig. 3.4B). Thus, in chimpanzees, *CCL4L2* appears to co-opt the *NPEPPSP1* promoter analogous to human *TBC1D3* but with a very different outcome. Analysis of the bonobo (*Pan paniscus*) genome indicates that both the inversion and *CCL4L2* insertion are present in this species, suggesting that this organization is a longstanding property of the *Pan* genus likely originating in the common ancestor of the two extant lineages (Supplementary Figs. S2, S3).

Orangutan, which does not harbor the *NPEPPS1* SD or its promoter, serves as a negative control and shows scant *TBC1D3* expression of the most terminal copy (Fig. 3.4B). We note, however, that there is abundant *TBC1D3* expression in the orangutan (Supplementary Fig. S5) from other copies. However, unlike the human ortholog, orangutan *TBC1D3* is uniformly expressed by numerous internal Cluster 2 paralogs that have independently expanded in the *Pongo* genus (Guitart et al. 2024; Supplementary Fig. S6).

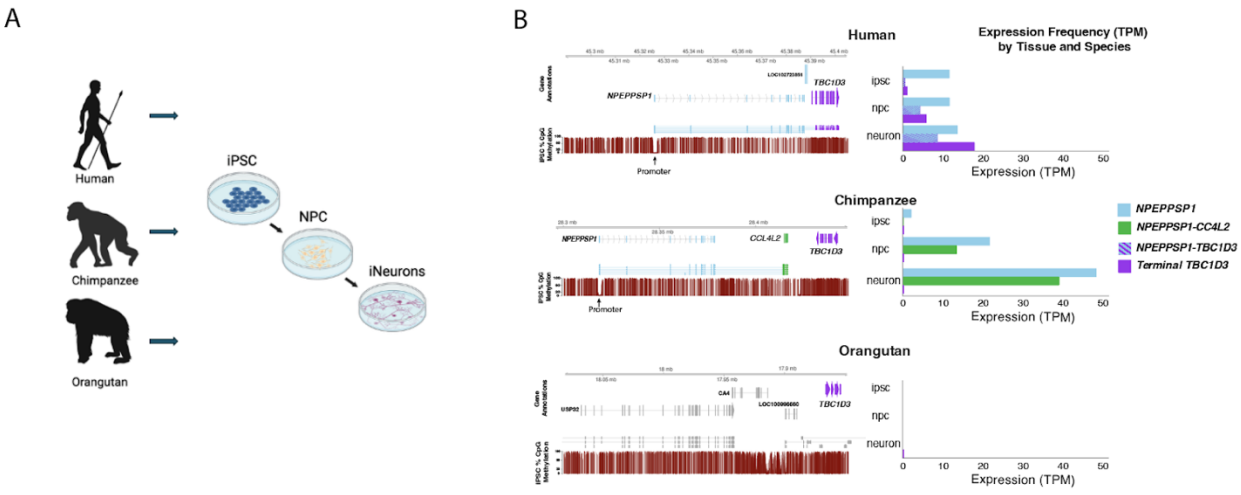


Figure 3. 4 Comparative expression of *NPEPPSP1-TBC1D3* in a neuronal developmental cell culture model of great apes.

(A) Experimental design. Fibroblasts sourced from human, chimpanzee, and gorilla samples were induced into iPSCs, NPCs, and neurons. Donor-specific genome assemblies from the iPSCs as well as gDNA and mRNA were generated using long-read sequencing protocols for each of the three stages.
(B) Comparative methylation and transcription of *NPEPPSP1*, terminal *TBC1D3*, and fusion genes in human, chimpanzee, and orangutan. A schematic of the gene organization and methylation profile (red; % of read methylated over CpG) is shown for iPSCs from each of the three species (left panels). Expression levels (transcripts per million, TPM) are compared for NPEPPS1, fusion genes, and the terminal TBC1D3 copy for each of three stages (right panels). Histogram illustrates TPM based on Kinnex sequencing (Methods).

While gorilla was not included in our neuron development cell model, we also attempted to investigate *NPEPPSP1-TBC1D3* regulation and expression by exploring the orthologous locus in the gorilla genome assembly and available Iso-Seq data from fibroblasts and testis tissue (Yoo et al. 2025). While gorilla inherited the same *NPEPPSP1* duplication, the promoter and first two exons of the pseudogene have been ablated by a subsequent 38 kbp deletion in the gorilla lineage (Fig. 3.5, Supplementary Fig. S3). Concomitantly, we find no evidence of *NPEPPSP1* promoter-

initiated *NPEPPS1-TBC1D3* fusion transcripts. We do find, however, fusion transcripts with transcript initiation starting near the *MRPL45* duplicated promoter, a gene inverted with respect to *TBC1D3* (Supplementary Fig. S7). We observe that these transcripts begin just upstream of the promoter of *MRPL45*, a gene inverted with respect to *TBC1D3*. This regulatory sequence may function as a bidirectional promoter (Trinklein et al. 2004), but *MRPL45* expression is nearly 20-fold higher, with 57 unique *MRPL45* transcripts to the three *TBC1D3* reads.

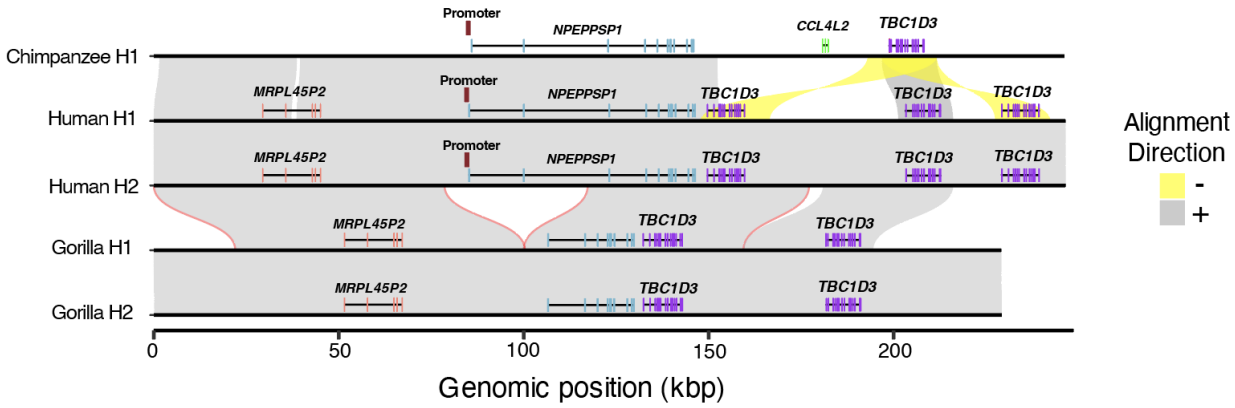


Figure 3. 5 Deletion of gorilla *NPEPPSP1* promoter.

Alignments of the *NPEPPSP1* duplication locus between one chimpanzee, two human, and two gorilla haplotypes show a 38 kbp deletion fixed in the gorilla lineage, removing both the *NPEPPSP1* promoter and first two exons.

3.3.5 THE FUSION TRANSCRIPT ENCODES TWO SEPARATE ORFS

Based on the human *NPEPPSP1-TBC1D3* fusion transcript consensus sequence, we predict two large, mutually exclusive ORFs (Supplementary Fig. S8A). The first ORF (ORF1) comprises a 449-codon reading frame corresponding to the entirety of *NPEPPSP1* and four codons located at the beginning of *TBC1D3* exon 2, traversing the fusion junction. The second ORF begins 13 bases upstream of ORF1, also at the second exon of *TBC1D3*, and represents the canonical 546-aa TBC1D3 ORF annotated by RefSeq (O'Leary et al. 2015). These ORFs are offset by two bases, placing them in different frames, precluding the formation of a fused protein product.

Unlike *TBC1D3*, there has been no characterization of *NPEPPSP1*, a supposed pseudogene, at the protein level. We searched publicly available proteomic resources using an *in silico* tryptic digest of *NPEPPSP1-TBC1D3* ORF1 and found four uniquely identifying peptides distinguishing *NPEPPSP1* from any other protein, including its ancestral paralog, *NPEPPS*. We searched for these four peptides in publicly available mass spectra datasets and identified five proteomic experiments with matches to the peptides. Although not conclusive, these data suggest *NPEPPSP1* may be translated and present in the human proteome (Supplementary Table 2; Methods).

Next, we assessed mutational tolerance of both ORFs by analyzing 295 human genomes recently sequenced as part of the Human Pangenome Reference Consortium (HPRC) (Liao et al. 2023) and the Human Genome Structural Variation Consortium (HGSVC) (Logsdon et al. 2025). Across these haplotypes, we found three instances of a premature termination of *NPEPPSP1*, driven by one splice junction in the third exon, and two nonsense mutations on the seventh exon (Supplementary Fig. S8B). In contrast, we found only a single occurrence of a splice junction

mutation (3' ss of exon 10) introducing a premature stop codon in the *TBC1D3* ORF (Supplementary Fig. S8C).

3.3.6 POST TRANSCRIPTIONAL PROCESSING OF FUSION TRANSCRIPTS.

Because *NPEPPSP1* duplicates only the first 11 exons of *NPEPPS*, we were interested in understanding how this partial duplication impacts the transcription and splicing of the gene. Using our neuronal developmental dataset, we characterized the expression of the two most common isoforms. We observe that solo *NPEPPSP1* transcripts utilize a novel polyadenylation site located 111 bp downstream of the canonical splice junction used by *NPEPPS* (Fig. 3.6A). This alternative site is strongly favored in iPSCs, being used in >98% (51/52) of full-length transcripts sequenced. The equivalent paralogous poly-adenylation site, however, was not observed in the ancestral *NPEPPS* gene and is excised within the 11th intron by a splice junction upstream of the site. Of note, we find that as cells differentiate toward neurons, *NPEPPSP1* splice site usage shifts toward the equivalent splice junction used by *NPEPPS*, losing the premature polyadenylation site and extending transcription to include the full-length *TBC1D3* ORF (Fig. 3.6B). Our results suggest that the fusion is regulated posttranscriptionally by alternative splicing of the last and first exons of *NPEPPSP1* and *TBC1D3*, respectively (Fig. 3.6).

We next sought to understand how the fusion impacts transcript processing. Given that nonsense-mediated decay (NMD) is often triggered by splice junctions located downstream of a translation stop codon, we hypothesized that *TBC1D3*-containing transcripts might be targeted for degradation and, thus, preferentially rescued upon NMD inhibition. To test this, we treated human and chimpanzee iPSCs and NPCs with cycloheximide (a known NMD inhibitor) or dimethyl sulfoxide (DMSO) as a control (Cheng et al. 2025). While we observe an overall increase in

expression of *NPEPPSP1*, *TBC1D3*, and the fusion transcript following cycloheximide treatment, we did not observe preferential rescue of *TBC1D3*. Instead, *NPEPPSP1* transcripts show the most pronounced increase (Supplementary Fig. S9), suggesting that they may be more susceptible to NMD than *NPEPPS1-TBC1D3* fusion transcripts.

Importantly, we find that *TBC1D3* expression surpasses *NPEPPSP1* in transcript abundance as human cells differentiate into neurons (Supplementary Figs. S10-S12; Supplementary Tables 3,4). Initially, we considered the possibility that this was due to expression from other *TBC1D3* paralogs. However, both epigenetic profiling and full-length transcript analysis across paralogs rule out this explanation. All 196 transcripts map to the terminal Cluster 2 *TBC1D3* copy. Only a single hypomethylation signature is identified corresponding to the promoter of the fusion transcript. We fail to identify any methylation or chromatin accessibility signatures consistent with a promoter in another paralog (Supplementary Tables 1,5).

While it is possible that this neuronal enrichment of *TBC1D3* may be due to the moderate 3' bias of the Kinnex and Iso-Seq LRS platforms, this seems unlikely since it is not observed in either NPCs or iPSCs, which were processed identically with respect to capture. We speculate, therefore, that there may be a posttranscriptional mechanism that isolates the downstream *TBC1D3* reading frame from its upstream linked-partner, *NPEPPSP1*. In support of this, we identified a Kozak sequence, AXXATGG, in the second exon of *TBC1D3*, located at the acceptor site of the splice junction that fuses the two genes and includes the start codon of ORF2. This site may serve as an internal ribosomal entry site for translation (Fig. 3.6B; Hellen, 2001).

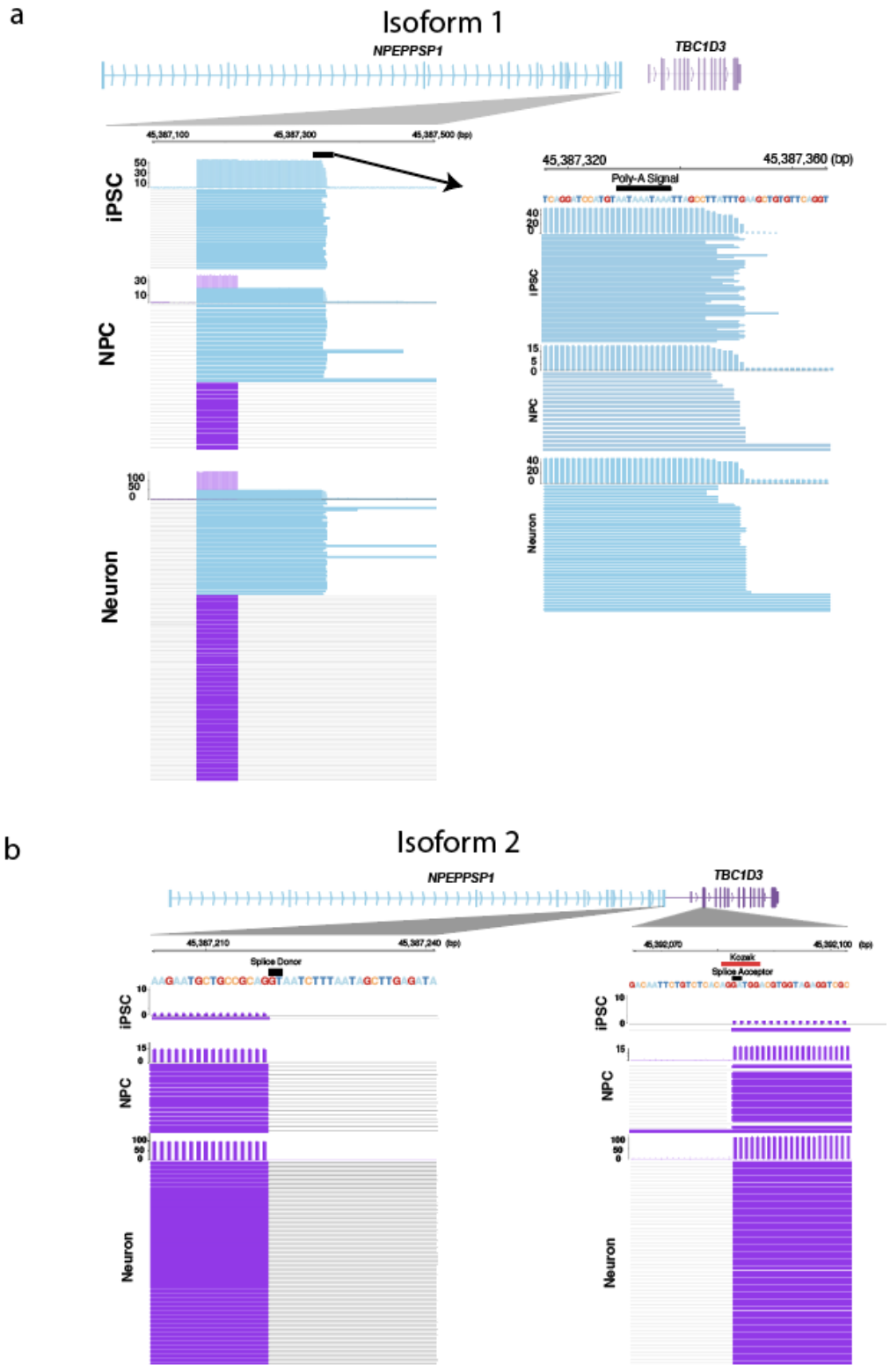


Figure 3. 6 Alternative splicing of *NPEPPSP1-TBC1D3* fusion.

(A) Solo *NPEPPSP1* isoform. Solo *NPEPPSP1* transcripts display a polyadenylation signal (chr17:46387329-46387338) terminating the transcript, a signal that is removed by splicing in fusion genes. **(B)** Fusion isoform. *NPEPPSP1-TBC1D3* transcripts display splicing between *NPEPPSP1* exon 10 and *TBC1D3* exon 2, where ORF2 begins. The splice acceptor site of *TBC1D3* overlaps with a Kozak sequence, recognized as AXXATGG.

3.4 DISCUSSION

This work advances our understanding of *TBC1D3* regulation and evolution in four major ways. First, we demonstrate that *TBC1D3* expression in humans is regulated via a position effect, primarily as a result of a fusion duplication event with the upstream duplicated gene, *NPEPPSP1*. Paralog-specific regulation via position effect is a well-documented phenomenon, as exemplified in the locus control region (LCR) regulating immediately proximal green and red opsin paralogs in chromosome X, and none of the additional copies downstream (Hayashi et al. 1999). Other examples include the beta-globin LCR (Reik et al. 1998) and the 5' LCR in *hGh* (Ho et al. 2006). Analogously, *NPEPPS1* defines an LCR of the terminal paralog of *TBC1D3* Cluster 2 with other paralogs being epigenetically repressed (Figs. 3.1 and 3.2).

Second, we identify the origin of this regulatory fusion as an SD event that duplicated ~120 kbp—encompassing the *NPEPPS* promoter and first 11 exons. The ancestral *NPEPPS* locus maps 9.16 Mbp proximally to the *TBC1D3* Cluster 2 locus on chromosome 17 (Fig. 3.3). Gene fusion arising from the juxtaposition of partial SDs has been proposed as a relatively common mechanism of neofunctionalization, particularly in primates (Marques-Bonet, Girirajan et al. 2009; Marques-Bonet, Kidd, et al. 2009; Yoo et al. 2025). For example, *HYDIN2* arose from partial duplication of *HYDIN*, resulting in a novel gene fused to *LOC101927468* (Dougherty et al. 2017) and

CHRFAM7A (i.e., *dupα7*), a human-specific fusion *CHRNA7* exons 5-10 and *FAM7A* exons A-E, regulating *CHR7A* in a dominant negative fashion (Sinkus et al. 2016).

In humans, the terminal copy of *TBC1D3* in Cluster 2 co-opted the regulatory program of *NPEPPS* to drive transcription of a fusion transcript. *NPEPPS*, also known as puromycin-sensitive aminopeptidase (PSA), is a cytoplasmic aminopeptidase conserved across metazoans that is broadly expressed across tissues and plays a key role in protein metabolism and cell cycle regulation (McLellan et al. 1988). In the brain, NPEPPS is enriched in the cerebellum and hippocampus, where it provides a protective effect against tauopathies and associated neurodegenerative diseases, including Parkinson's and dementia (Karsten et al. 2006; Muraoka et al. 2020). Disease associations remain difficult for *TBC1D3*, owing to the high sequence identity between paralogs and the significant degree of structural heterozygosity among individuals (Guitart et al. 2024). However, we find that *NPEPPSP1-TBC1D3* expression is enriched in the brain, and in particular, increases in expression in the cerebellum as the brain develops into adulthood (Supplementary Fig. S13). *In vitro* experiments of *TBC1D3* propose numerous mechanisms of *TBC1D3* function, all of which increase cell proliferation (Frittoli et al. 2008; Wainszelbaum et al. 2008, 2012; Ju et al. 2016; Hou et al. *2021*). The human cerebellum is unique among other brain regions for its delayed maturation and high degree of neuronal density, which holds ~80% of brain neurons despite constituting only 10% of volume (Van Essen et al. 2018). Recent *in vitro* studies suggest *TBC1D3* both promotes neuronal proliferation and protracts synaptic plasticity and contributes to neoteny (Hou et al. 2021; Dong et al. 2024). We suggest *TBC1D3* may contribute to human cerebellar development through this novel transcriptional context or the human-specific carboxy-terminus (Guitart et al. 2024).

Gene fusions driven by transcriptional readthrough commonly coincide with the fusion of two reading frames into a single novel protein (McCartney et al. 2019). In contrast, *NPEPPSP1-TBC1D3* retains two large, independent reading frames across the singly transcribed fusion (Supplementary Fig. S8). Polycistronic expression has been documented in eukaryotes, most notably in *C. elegans*, where approximately 15% of nuclear genes are organized into operons (Blumenthal *et al*. 2018). In nearly all other eukaryotes, polycistronic transcription events have been restricted to mitochondrial and plastid genes—both derivatives of prokaryotes (Barkan et al. 1988; Mercer et al. 2012). Nevertheless, nuclear, multi-ORF genes have been characterized. *ATF4* and *GRN*, for example, include an upstream ORF (uORF) whose retention in the transcript either promotes or inhibits translation of the downstream, main ORF (mORF) (García-Ríos et al. 1997; Ryczek et al. 2023; Vattem et al. 2004; Capell et al. 2013). In these cases, however, the uORFs tend to be small, only 16 aa on average (Wethmar et al. 2010), while the reading frame we observe for NPEPPSP1 is 479 aa.

Third, we illustrate a developmentally dependent, posttranscriptional regulation of *TBC1D3* as cells differentiate into neurons. In our neuronal differentiation system, *TBC1D3* increases in expression and fusion transcripts proportionally diminish as human stem cells differentiate into neurons (Fig. 3.6; Supplementary Figs. S10-S12), raising the possibility of independent expression of *TBC1D3*, possibly from other paralogs, or the removal of *NPEPPSP1* from the fusion gene via alternative splicing. Our epigenetic profiling reveals that nonterminal paralogs remain epigenetically silenced, lacking the promoter-associated hypomethylation and chromatin accessibility (Figs. 3.1 and 3.2). We therefore propose that the *TBC1D3* reading frame is selectively retained in the transcript via a posttranscriptional process that removes or excludes the upstream *NPEPPSP1* ORF. Supporting this, cycloheximide treatment, which inhibits NMD,

selectively rescues *NPEPPSP1* and fusion transcripts when compared to *TBC1D3* (Supplementary Fig. S9). This suggests a posttranscriptional regulatory model in which neuronal cell fates isolate *TBC1D3*, potentially influencing its translation efficiency or subcellular localization. Follow-up protein-level studies (e.g., immunofluorescence or mass spectrometry) are needed to assess the fate of the protein(s) derived from the NPEPPSP1-TBC1D3 fusion transcript.

Finally, we place this regulatory architecture in an ape evolutionary framework. The *NPEPPSP1* duplication event occurred between 8–9 MYA, in the common ancestor of African apes. This time interval corresponds to a period of rapid chromosomal restructuring, including large-scale inversions and fusions that dramatically restructured local chromosomal regions in different ape lineages potentially as a result of extensive incomplete lineage sorting (Yang et al. 2025; Mao et al. 2021, Yoo et al. 2025). The *NPEPPSP1-TBC1D3* locus exhibits structural divergence across apes, with incomplete lineage sorting and lineage-specific rearrangements contributing to distinct haplotypes in humans, chimpanzees, and gorillas, including multiple large-scale and smaller inversion events (Supplementary Figs. S2, S3). Despite the shared origin of the fusion in all African great-ape lineages, *TBC1D3* expression driven by *NPEPPSP1* appears to be human-specific. In the genus *Pan*, a subsequent 44 kbp chimpanzee duplication inserted *CCL4L2* between *NPEPPSP1* and *TBC1D3*, disrupting the fusion in both chimpanzees and bonobos. In gorilla, a 51 kbp deletion removed the *NPEPPSP1* promoter and initial exons, ablating the fusion transcript. Based on the genomic architecture, only humans are, thus, capable of producing the *NPEPPSP1-TBC1D3* fusion product at high levels during neurodevelopment.

We find that the expressed *TBC1D3* gene has been dramatically restructured during human evolution. Our previous work identified a human-specific C-terminal truncation resulting in a 58-aa modification in all expressed human copies. Here, we show that the 5' end of the major human

TBC1D3 transcript is also unique to our lineage—it includes the first 11 exons of *NPEPPSP1*. Further, we provide evidence that this modification boosts expression during neuronal development. The functional consequences of these alterations—for translation, localization, or function—remain unresolved. Experimental perturbation of *TBC1D3* remains challenging due to its high sequence identity across paralogs. However, our findings present a potential solution: targeting the fixed *NPEPPSP1* promoter, which regulates the polymorphic *TBC1D3* cluster through a position effect. This regulatory mechanism—where a fixed regulatory element controls expression of structurally variable loci—may be more common than expected, offering a new opportunity to understand, and eventually target, structurally variable and medically relevant genes that have, until recently, been impossible to characterize.

3.5 METHODS

3.5.1 METHYLATION ANALYSIS

ONT bam data used for the CHM13 methylation analysis was generated and processed described in (Mastrorosa *et al.* 2024). CHM13 genomic DNA reads used to generate the T2T reference were aligned back onto the reference using minimap2 (version 2.28; Li *et al.* 2018). Following alignment, methylation basecalls of CpG sites were processed and aggregated using the modbam2bed software (v0.10.0; github.com/epi2me-labs/modbam2bed). Following methylation aggregation, CpG sites with coverage of five reads or less were discarded. For visualization (see Fig. 3.1), percent of reads methylated over each CpG site are illustrated as black points, with a rolling average of 15 CpG sites are visualized with a red line. These results were visualized in R using ggplot2 geom_point and geom_line functions.

3.5.2 FIBER-SEQ ANALYSIS

Fiber-seq libraries of HG02630 were prepared as described in (Real et al. 2025). Reads from the dataset were re-aligned to polished HG02630 assemblies generated with Deep polisher by the Human Pangenome Reference Consortia (Mastoras *et al.* 2025). Reads were mapped back to HG02630 with the following minimap2 (version 2.24), and adenosine and cytosine methylation calls from the original base-calling were maintained with the new alignment bams. FIRE peaks were identified using Firetools (version 0.0.7; Vollger, *et al*. 2025). Fiber-seq peaks were visualized with GViz (version 1.46.1) using the following command.

3.5.3 FULL-LENGTH TRANSCRIPT ANALYSIS

CDNA libraries were aligned as previously reported in (Guitart et al. 2024). Reads were aligned to a reference assembly using minimap (version 2.24) with the settings <minimap2 -ax splice --sam-hit-only --secondary=yes -p 0.5 --eqx -K 2G>. Following alignment primary and secondary alignments were scrutinized. For paralog-specific reads, primary alignments with alignments scores greater than second-best alignments were used. Reads were visualized with R using the GViz visualization library. Reads mapping to only a single exon were excluded from the gene transcript counts.

Multi-Iso-Seq reads were gathered as described in (Dishuck *et al.* 2025). Iso-Seq libraries were normalized to transcripts per million. Read counts were then organized by hierarchical clustering with R hclust and visualized as a heat plot with ggplot2.

3.5.4 POLYADENYLATION ANALYSIS

Polyadenylation signals were identified by looking for canonical polyadenylation motif (AATAAA) motif within 20bp of the end of transcript read alignments (Higgs *et al.* 1983). Splice junctions were identified by marking AG 3’ splice acceptor and GT 5’ splice acceptor motifs at splice junctions along cDNA

alignments (Stephens & Schneider 1992). Polyadenylation signals and splice junctions were visualized with rtracklayer (version 1.62.0; Lawrence *et al.* 2009).

3.5.5 PHYLOGENETIC ANALYSIS

Multiple sequence alignments (MSA) for phylogenies were generated as follows. Homologous sequence was identified by mapping a reference target sequence to assemblies using minimap2 (-x asm20 --secondary --eqx -p -.05 -N 1000). Following sequence extraction, homologous sequences were aligned as multiple sequence alignments with MAFFT (version 7.5.25; Katoh *et al*. 2013) with the following parameters: --reorder --preservecase --adjustdirection --maxiterate 1000 --thread 16. Following MSA generation, maximum likelihood phylogenies were generated with iqtree2 (version 2.1.2; Minh *et al.* 2020) with the following parameters: -nt AUTO -m MFP -s {MSA_fasta} -o {outgroup} --prefix {output_name} -alrt 1000 -bb 1000.

For *NPEPPSP1* duplicon phylogenetic analysis, the terminal 15 kbp of the *NPEPPSP1* duplication most proximal to *TBC1D3* was extracted using SEDEF annotations (Numanagic *et al*. 2018). The sequence was then mapped to the human, chimpanzee, gorilla, orangutan, and macaque T2T assemblies using minimap2 with the default setting "-x asm20" (Yoo *et al*. 2025). These mappings were extracted with bedtools and aligned into a multiple sequence phylogeny with MAFFT (version 7.525; Katoh *et al*. 2013). Next, the multiple sequence alignment was trimmed for gaps and misalignments with both Trimal (version v1.4.rev22) and manual pruning. Next, a phylogeny and timing estimates were generated using the iqtree2 software (version 2.12; Minh *et al.* 2020), with macaque as the outgroup. Timing estimates were calculated with iqtree2, using the following divergence times: (human-chimpanzee: 6.5 MYA; human-gorilla: 8.0 MYA; human-macaque: -24 MYA**)** as estimated in the fossil record (Dunsworth 2010; Stevens et al. 2013). The following command for timing estimation was used: iqtree2 -s {MSA} --date {divergence_date list} --date-ci 100 -te {phylogeny } --keep-ident -redo -o {outgroup_name} --date-tip 0

-alrt 1000 -bb 1000). For *CCL4L2*, the same process described as above was used, but for the *CCL4L2* gene model with the upstream 5 kbp sequence was included for additional sequence.

3.5.6 Short-Read-Depth Analysis

Short reads libraries from the 1000 Genomes project (Sudmant *et al*. 2010). Illumina libraries were K-merized into 32 bp libraries with Meryl (version). Next, k-mer libraries were aligned to the CHM13 T2T (Nurk et al. 2022) reference genome using FastCN (Pendleton et al. 2018), allowing for up to two mismatches between 32-mer and assembly alignments. Copy numbers over *TBC1D3*, *NPEPPSP1*, and the *NPEPPSP1* (see Supplementary Fig. S13) inferred promoter were estimated by taking the average copy number over the given region, estimated by normalized read depth coverage to unique diploid sequence across the genome.

3.5.7 Conservation in Human Haplotypes

Gene sequences were aligned using multiple sequence alignment with MAFFT (v7.525). SNVs and indels were identified by comparing each gene sequence to the reference gene sequence from CHM13 v2.0. Variant functional consequences were annotated using VEP (v111) with gene annotation (JHU RefSeqv110 + Liftoff v5.2; Shumate *et al. 2021*). Pathogenicity predictions for missense variants were obtained using AlphaMissense (Cheng et al. Science, 2023), which were lifted over to CHM13 coordinates using UCSC LiftOver and complemented by PolyPhen and SIFT (version 20240502). Lollipop plots were generated using the R package trackViewer (v1.40.0).

3.5.8 Haplotype Sequence Annotation

Segmental duplications for *NPEPPS-TBC1D3* haplotypes were annotated in the *de novo* assemblies with SEDEF (version 1.1; Numanagic *et al*. 2018). Duplicon tracks were generated for each assembly with DupMasker (Jiang et al.) and run with the Rhodonite workflow (version 0.12; https://github.com/mrvollger/Rhodonite).

3.5.9 RNA Analysis

Short read libraries from the developmental GTEx Consortium were used for the analysis. Reads were aligned to CHM13 using bwa (version 0.7.17), and transcript abundance was normalized using pydeseq2 (version 0.5.2). Read abundance for *NPEPPSP1-TBC1D3* fusion expression was specifically quantified by counting reads spanning the merged exon boundaries across the last and first exons of *NPEPPSP1* and *TBC1D3*, respectively.

3.5.10 Reading Frame Maintenance

Gene sequences were aligned using multiple sequence alignment with MAFFT (v7.525). SNVs and indels were identified by comparing each gene sequence to the reference gene sequence from CHM13 v2.0. Variant functional consequences were annotated using VEP (v111; McLaren et al. 2016) with gene annotation (JHU RefSeqv110 + Liftoff v5.2). Pathogenicity predictions for missense variants were obtained using AlphaMissense (Cheng et al. Science, 2023), which were lifted over to CHM13 coordinates using UCSC LiftOver and complemented by PolyPhen and SIFT (20240502). Lollipop plots were generated using the R package trackViewer (v1.40.0).

3.5.11 Mass Spectrometry Analysis

Amino acid sequences for *NPEPPSP1* and *NPEPPS* were pulled from RefSeq annotations and aligned into a multiple sequence alignment with MAFFT (v7.525). Amino acid substitutions between the two paralogs were determined visually from the alignment. Next, tryptic peptides for both protein sequences were computationally inferred with expasy online tool (https://web.expasy.org/peptide_cutter/). Diverse peptide sequences between the two proteins were then selected and isolated for later analysis. These peptides were searched for in the MASSive database (https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp?redirect=auth).

CHAPTER 4. *TBC1D3* CELL CULTURE EXPERIMENTS

4.1 ABSTRACT

Here, I describe experimental efforts to investigate the functional consequences of human-specific modifications to the *TBC1D3* gene family, with a particular focus on the derived C-terminal extension that distinguishes human *TBC1D3* from nonhuman primate orthologs. I outline the experimental design, including the rationale for targeting this modification, the cellular systems employed, and the expected molecular and phenotypic readouts. Although these experiments did not yield interpretable or reproducible results, their outcomes provide insight into the technical and biological challenges associated with functional interrogation of recently duplicated, highly similar genes. I discuss potential explanations for the failure of these approaches, including a possible dosage toxicity due to a constitutive promoter and an amino acid discrepancy between this experiment and other published investigations and propose alternative strategies for future studies.

4.2 INTRODUCTION

In Chapter 2 (Guitart et al. 2024), I describe the discovery of human-specific modifications to the *TBC1D3* coding sequence. We identified a 43-nucleotide deletion unique to humans that induces a frameshift, resulting in the addition of 41 novel amino acids at the carboxy-terminus that are absent in other primates. Alterations to the C-terminus are a recurring feature of duplicated genes undergoing neo- or sub-functionalization. Generally, the carboxy-terminus of a protein is often

solvent-exposed and enriched for regulatory roles, influencing subcellular localization and posttranslational modification (Sharma *et al*. 2019). A well-characterized example is *ARHGAP11B*, in which a splice-site mutation truncates the Rho-GAP domain and introduces 47 novel amino acids to the carboxy-terminus, redirecting the protein to mitochondria and promoting glutaminolysis (Florio *et al*. 2016; Namba *et al.* 2020). The solvent-exposed C-terminus may also influence protein interactions, as observed in the *Rhodopsin* gene, where pathogenic C-terminal disruptions impair interaction with Arf4, leading to defective protein trafficking and retinal degeneration (Deretic *et al*. 2005).

Additionally, we found evidence of positive selection within six amino acids in the TBC1D3 human protein sequence. Small changes to protein amino acid sequence, even just a few substitutions, can have significant effects on function and resulting phenotype. One noteworthy example is human changes to forkhead box P2 (*FOXP2)*, where two missense mutations exclusive to humans increased dendrite length and synaptic plasticity in neurons of the striatum, which may have enhanced human capabilities of speech and language (Bornschein *et al*. 2023). Other examples include the human *HbS* allele conferring resistance against malaria (Kwiatkowski, 2005) and the pocket mouse *Mc1r* allele causing black pigmentation for individuals living in lava-rock habitats (Nachman *et al*. 2003).

Previous *in vitro* studies explored the function of TBC1D3 but largely lacked consideration of human-specific modifications. Studies by Frittoli et al. (2008), Wainszelbaum et al. (2008, 2012), He et al. (2014), and Wang et al. (2017) used cell transformation assays to demonstrate multiple mechanisms by which TBC1D3 promotes oncogenic phenotypes. In parallel, Ju et al. (2016), Hou et al. (2021), and Dong et al. (2024) described roles for *TBC1D3* in neural progenitor proliferation and synaptogenesis using both cellular and mouse transformation

models. Notably, in all of these studies, *TBC1D3* constructs were derived from human cDNA libraries, without comparison to nonhuman primate orthologs or alternative reading frames.

The goal of this study was to extend this body of work by directly testing how human-specific modifications to *TBC1D3* influence gene function during development. Using bonobo neural progenitor models, which lack endogenous *TBC1D3* expression (see Chapter 3), we introduced multiple *TBC1D3* reading frames via lentiviral transduction, with GFP serving as a reporter. We first assessed the effects of these reading frames on cell proliferation, hypothesizing that human-specific modifications would differentially affect neural progenitor proliferation relative to nonhuman primate *TBC1D3*. We subsequently planned to examine subcellular localization and protein–protein interaction phenotypes. In this section, I describe the experimental design in detail, present the results of the unsuccessful experiments, and propose potential explanations for the cytotoxicity observed across all *TBC1D3*-expressing transductions.

4.3 RESULTS

4.3.1: EXPERIMENTAL DESIGN

We conducted a cell proliferation assay using bonobo neural progenitor cells (Fig. 4.1). To test the functional consequences of human-specific *TBC1D3* modifications, we generated lentiviral vectors encoding GFP as a reporter, linked to one of eight TBC1D3 reading frames. These constructs represent all combinations of human and nonhuman primate *TBC1D3* gene bodies paired with either human or nonhuman primate C-termini, including a construct containing the human-specific C-terminal deletion (Table 1). Expression of each GFP–TBC1D3 fusion was driven by the EF1α promoter, a strong constitutive promoter, to ensure robust and uniform fluorescence for downstream quantification.

Bonobo neural progenitor cells were transduced with one of the eight lentiviral libraries. Baseline transduction efficiency and cell counts were assessed two days post-transduction, followed by a second measurement at seven days post-transduction to evaluate proliferation. Proliferation was quantified by tracking GFP-positive cells over time. In parallel, immunofluorescence assays were performed to assess maintenance of neural progenitor identity ($SOX2^+$), neuronal differentiation ($ELAV3/4^+$), and apoptosis (cleaved $CASPASE\text{-}3^+$). Fluorescent signals were quantified using both fluorescence microscopy and flow cytometry.

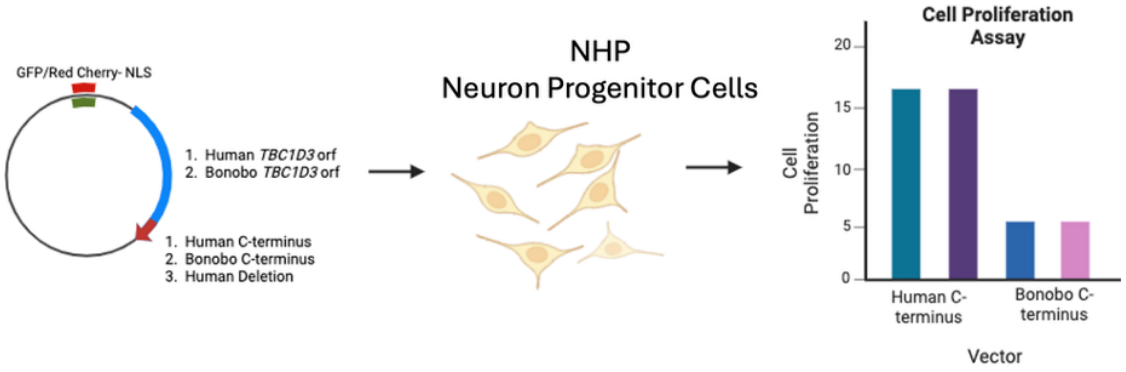


Figure 4. 1 Experimental design.

Bonobo neuron progenitor cells were transfected with a lentivirus carrying one of eight sample reading frames payloads and examined for the effect on cell proliferation. NHP: nonhuman primate

4.3.2: CELL DEATH

Following transduction, we observed that all *TBC1D3*-expressing samples proliferated at a markedly reduced rate compared to the empty-vector control (Fig. 4.2A). Quantification by flow cytometry confirmed that cells expressing any *TBC1D3* reading frame proliferated at levels comparable to the negative control, *NGN2* (Fig. 4.2B). *NGN2* is a transcription factor that drives neuronal differentiation, after which cells exit the cell cycle (Hulme *et al.* 2021).

To determine whether reduced proliferation in the *TBC1D3* samples was due to premature neuronal differentiation, we stained each condition for ELAV3/4, a marker of post-mitotic neurons (Fig. 4.2C). As expected, *NGN2*-transduced cells showed strong enrichment for ELAV3/4, consistent with neuronal differentiation. In contrast, none of the *TBC1D3*-expressing samples exhibited increased ELAV3/4 staining and instead more closely resembled the empty-vector control, indicating that *TBC1D3* expression did not induce neuronal differentiation.

We next tested whether *TBC1D3* expression was associated with cytotoxicity by staining for cleaved CASPASE-3, a marker of apoptotic cell death (Fig. 4.2D). Relative to controls, *TBC1D3*-expressing cells showed a modest but consistent increase in CASPASE-3 signal, suggesting that *TBC1D3* expression may exert a cytotoxic effect in bonobo neural progenitor cells.

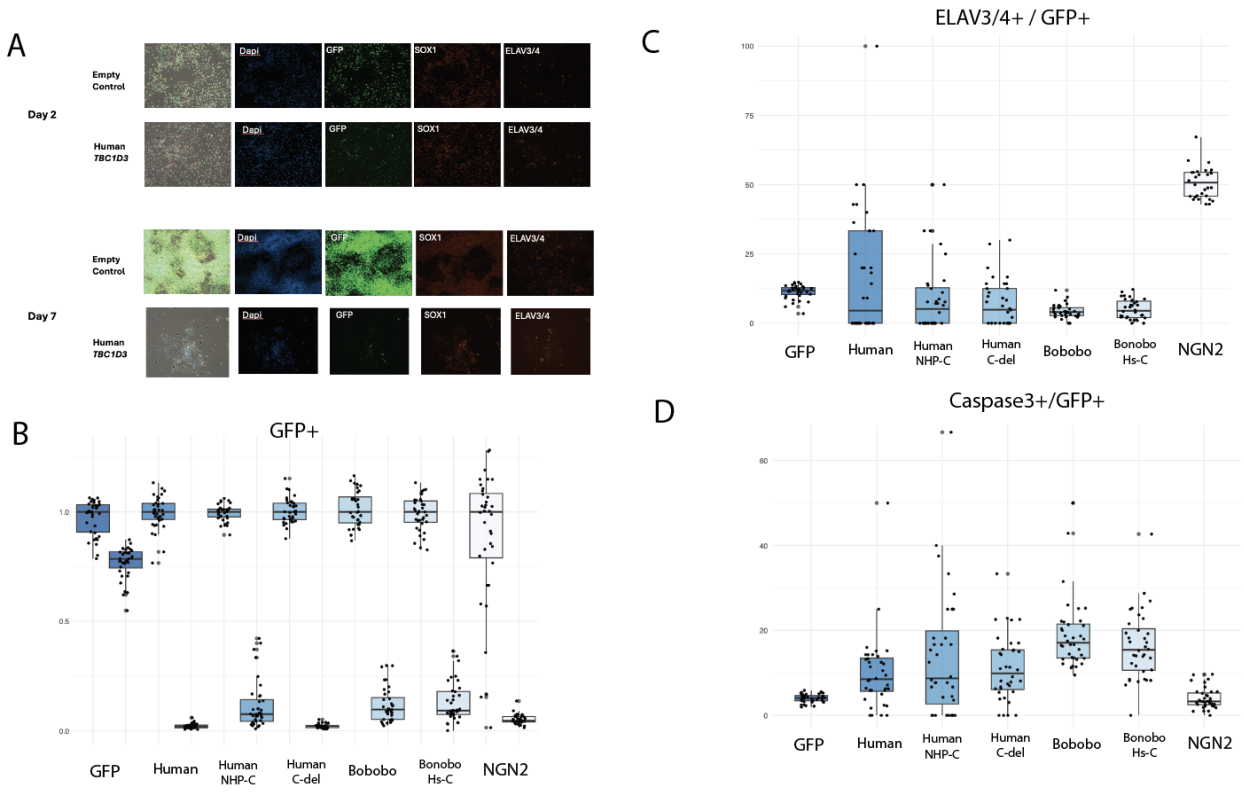


Figure 4. 2 TBC1D3-mediated cell death.

(A) Day 2 and day 7 post transfections of cells with empty control (top) and human *TBC1D3* (bottom) with stains for DAPI, GFP, SOX1 (NPC marker) and ELAV3/4 (neuronal marker). Human *TBC1D3* cells experienced significantly less growth as compared to an empty control. **(B)** GFP percentage normalized to day 2 fluorescence for GFP, human *TBC1D3*, human *TBC1D3* with NHP C-terminus, human *TBC1D3* with C-terminus deletion, bonobo *TBC1D3*, bonobo *TBC1D3* with human C-terminus, and *NGN2*. All libraries experienced significant reduction in fluorescence compared to GFP, indicative of failure to proliferate. **(C)** Percent staining of ELAV3/4, an indicator of neuronal differentiation. *NGN2*, a negative control promoting neuronal differentiation, is noticeably elevated compared to other samples, suggesting

TBC1D3 did not promote neuron differentiation. **(D)** Caspase3+ a stain for cell apoptosis. All *TBC1D3* derivatives show elevated staining of Caspase3, indicative of cell apoptosis fate.

4.4 DISCUSSION

This experiment did not yield conclusive insights into the functional consequences of human-specific TBC1D3 modifications. However, several observations help constrain the possible explanations for the uniformly reduced proliferation observed across all *TBC1D3*-expressing conditions. First, the phenotype is unlikely to be attributable to lentiviral toxicity, as cells transduced with the empty-vector control proliferated normally. Second, the NGN2 negative control exhibited robust ELAV3/4 staining, confirming neuronal differentiation, whereas *TBC1D3*-expressing cells did not, thereby ruling out premature differentiation toward a neuronal fate as the cause of reduced proliferation.

We therefore hypothesize that the observed phenotype reflects a TBC1D3-dependent toxic effect. Several non-mutually exclusive mechanisms may account for this outcome. One possibility is overexpression: as described in Chapter 3, *TBC1D3* is normally expressed at modest levels, whereas the EF1α promoter used here is a strong constitutive promoter and may have driven supraphysiological expression that is deleterious to neural progenitor cells. A second possibility is that expression of *TBC1D3* in isolation, rather than as part of the endogenous *NPEPPS–TBC1D3* fusion context, disrupts normal cellular regulation and results in toxicity. Finally, the phenotype may stem from differences in the amino acid sequence used in our constructs. The lentiviral vectors were designed using the Ensembl-annotated *TBC1D3* reading frame (transcript ID: ENST00000612727.5) (O'Leary *et al.* 2016). However, upon contacting

Dr. Zhen-Ge Luo's laboratory, which previously characterized TBC1D3 function in neural progenitors, we identified a single amino acid discrepancy between sequences: a glutamine-to-proline substitution at residue 358 (Fig. 4.3). Because glutamine is hydrophilic and proline is hydrophobic and conformationally restrictive, this substitution could substantially alter protein structure or stability, potentially contributing to the observed cytotoxicity.

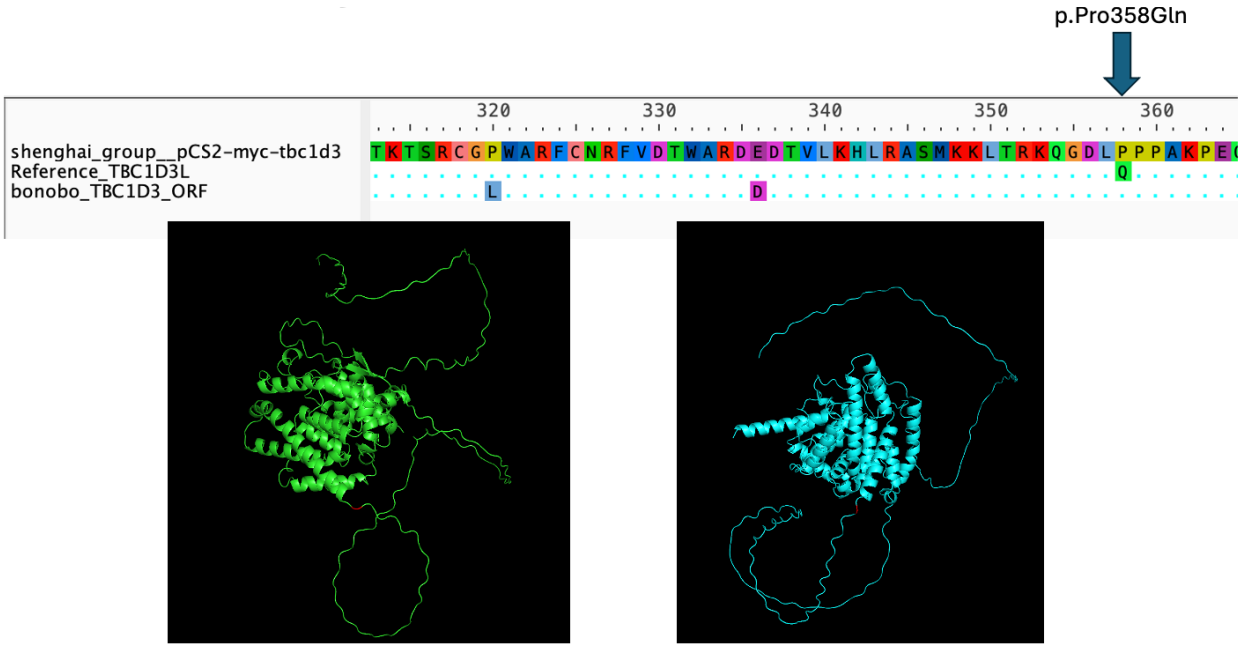


Figure 4. 3 Amino acid differences

Amino acid sequences from Dr. Luo's group experiments (top), the human reference *TBC1D3* amino acid sequence used in our experiment (middle), and bonobo amino acid sequence used by our experiment (bottom) were aligned in a multiple sequence alignment. A single amino acid substitution at site 358 differentiates Luo *et al*. TBC1D3 from the reference TBC1D3. This sequence is located in the disordered region of the carboxy-terminus (predicted with Alpha-fold)

These experiments result in more questions than answers, and as it stands these results have not been published in a peer-reviewed journal. However, the contents of this thesis will be publicly available. We still suspect that *TBC1D3*'s function must have changed in some way as a result of

these significant human-specific modifications to the protein sequence, and we hope and encourage the research community to explore this avenue further.

CHAPTER 5. SUMMARY AND FUTURE DIRECTIONS

5.1 HUMAN-SPECIFIC EVOLUTION AND REGULATORY CONTROL OF THE *TBC1D3* GENE FAMILY

In this work, I advance the evolutionary and regulatory characterization of *TBC1D3*. I show that, within humans, TBC1D3 has undergone multiple protein-level modifications, including four amino acid substitutions under positive selection and a replacement of the final 17 amino acids of the carboxy terminus with 41 novel amino acids. Further work is required to determine how these human-specific changes alter TBC1D3 function. I also demonstrate that *TBC1D3* expression is regulated through fusion with the upstream *NPEPPSP* in a developmentally programmed manner.

Several studies have investigated *TBC1D3* using *in vitro* perturbation approaches (Frittoli *et al.* 2008; Wainszelbaum *et al*. 2008, 2012; Kong *et al*. 2012; Hou *et al*. 2021; Dong *et al.* 2024). Owing to the high copy number and extreme sequence identity among *TBC1D3* paralogs, these studies predominantly relied on transcriptional ablation using RNA interference (RNAi). Collectively, they characterized roles for *TBC1D3* in promoting cell proliferation and delaying neuronal differentiation through a range of molecular mechanisms. Dong *et al.* (2024) extended this work by generating a comprehensive CRISPR knockout of all *TBC1D3* paralogs in cultured human brain slices. However, these studies did not consider the evolutionary context of the gene family and did not address human-specific modifications to *TBC1D3*.

Moreover, none of these investigations resolved functional differences among individual *TBC1D3* paralogs; instead, they globally repressed or induced the entire gene family. The high

copy number and sequence similarity of *TBC1D3* make paralog-specific perturbation particularly challenging. In Chapter 3, I show that, at the regulatory level, *TBC1D3* is predominantly expressed as a fusion with *NPEPPSP1*, which is responsible for over half *TBC1D3* expression, as measured by full-length cDNA sequencing. This finding creates two opportunities to isolate the most highly expressed copy and to test whether this human-specific regulatory configuration influences cellular development and biology. Specifically, either the fusion transcript can be targeted at the RNA level, or its promoter can be targeted at the genomic level.

First, an RNAi-based approach—using either shRNA or siRNA—could be designed to target the unique fusion junction between *NPEPPSP1* and *TBC1D3* (Agrawal *et al*. 2003). Because this sequence is exclusive to the fusion transcript, such an approach would minimize off-target effects. However, transcripts that undergo posttranscriptional processing that disrupts the fusion junction would be expected to evade RNAi-mediated repression.

Second, the promoter driving the *NPEPPSP1–TBC1D3* fusion could be targeted to achieve complete transcriptional repression. This strategy offers a simpler alternative to comprehensive *TBC1D3* CRISPR knockout approaches, as the structure and copy number of the *NPEPPSP1–TBC1D3* promoter are fixed (two copies per haploid genome). However, the ancestral *NPEPPS* locus shares the same promoter sequence, necessitating careful experimental design to ensure selective repression of the fusion gene. Our collaborators at the University of California, San Francisco—Drs. Jessie Brunner and Alex Pollen—are developing a paralog-specific repression strategy using a dCas9–KRAB system (Thakore *et al.* 2015). Targeting the *NPEPPSP1–TBC1D3* promoter with this system would enable selective repression, which could be validated by measuring transcript abundance from the *NPEPPSP1–TBC1D3* fusion and the ancestral *NPEPPS* locus using targeted RT-PCR or ddPCR (Bustin *et al*. 2009; Pinheiro *et al*. 2011).

5.2 INVESTIGATING *TBC1D3* WITH ASSOCIATION STUDIES

In vitro studies are useful for testing explicit genetic perturbations and elucidating potential molecular mechanisms. However, their results can be difficult to interpret, may be confounded by experimental artifacts, and often fail to recapitulate the complexity of tissue development *in vivo*. In contrast, association studies provide a powerful framework for characterizing the true genetic burden of a locus and for linking naturally occurring genomic variation to disease risk, clinical outcomes, and population-level phenotypes. This approach has been foundational in advancing our understanding of genome function, revealing causal—or tightly linked—variants underlying human disease. Genome-wide association studies have identified key susceptibility loci for breast cancer (e.g., *BRCA1/2* and *FGFR2*; *Kuchenbaecker et al*. 2017; Zhang *et al.* 2016), schizophrenia (e.g., C4; Sekar *et al*. 2016), cardiovascular disease (e.g., *LPA*; Clarke *et al.* 2009), and metabolic disorders such as type 2 diabetes (e.g., *TCF7L2*; *Ruiz-Narváez* 2014), directly shaping medical genetics, risk prediction, and clinical screening strategies.

Historically, genetic association studies have focused on single-nucleotide variants (SNVs), which can be readily genotyped and discretely categorized across large cohorts, enabling well-powered statistical analyses (Buniello *et al.* 2019). In contrast, structural variants (SVs) present substantially greater analytical challenges despite the fact they have been shown to have larger effects than SNVs (Sudmant *et al.* 2015). SVs are more difficult to detect and validate than SNVs, particularly using short-read sequencing, and their diverse sizes and breakpoint complexities complicate accurate genotyping across individuals (Alkan *et al*. 2011). As a result, publicly available catalogs of SVs remain far less comprehensive than single-nucleotide

polymorphism (SNP) databases, limiting both variant discovery and statistical power in association studies.

Recent large-scale initiatives, including the Human Pangenome Reference Consortium (HPRC; Liao *et al*. 2023), Human Genome Structural Variation Consortium (HGSVC; Fairley *et al*. 2020), and *All of Us* Research Program (The All of Us Research Program Investigators, 2019), are beginning to address this gap by generating high-quality, haplotype-resolved assemblies that enable accurate discovery and cataloging of structural variation across diverse human populations and linking it to human phenotypes by imputation and direct genotyping (Garimella et al., 2025). In Chapter 2, I leveraged data from the HPRC to characterize the structural organization of the *TBC1D3* locus across 69 fully assembled and validated human haplotypes. Since the publication of that work, I have extended this analysis to a total of 544 haplotypes, substantially increasing the resolution with which structural diversity at this locus can be assessed. Using phylogenetic organization of haplotypes based on sequence flanking the *NPEPPSP1–TBC1D3* fusion junction, I identified two major structural haplogroups characterized by an inversion that disrupts the *NPEPPSP1–TBC1D3* fusion (Fig. 5.1).

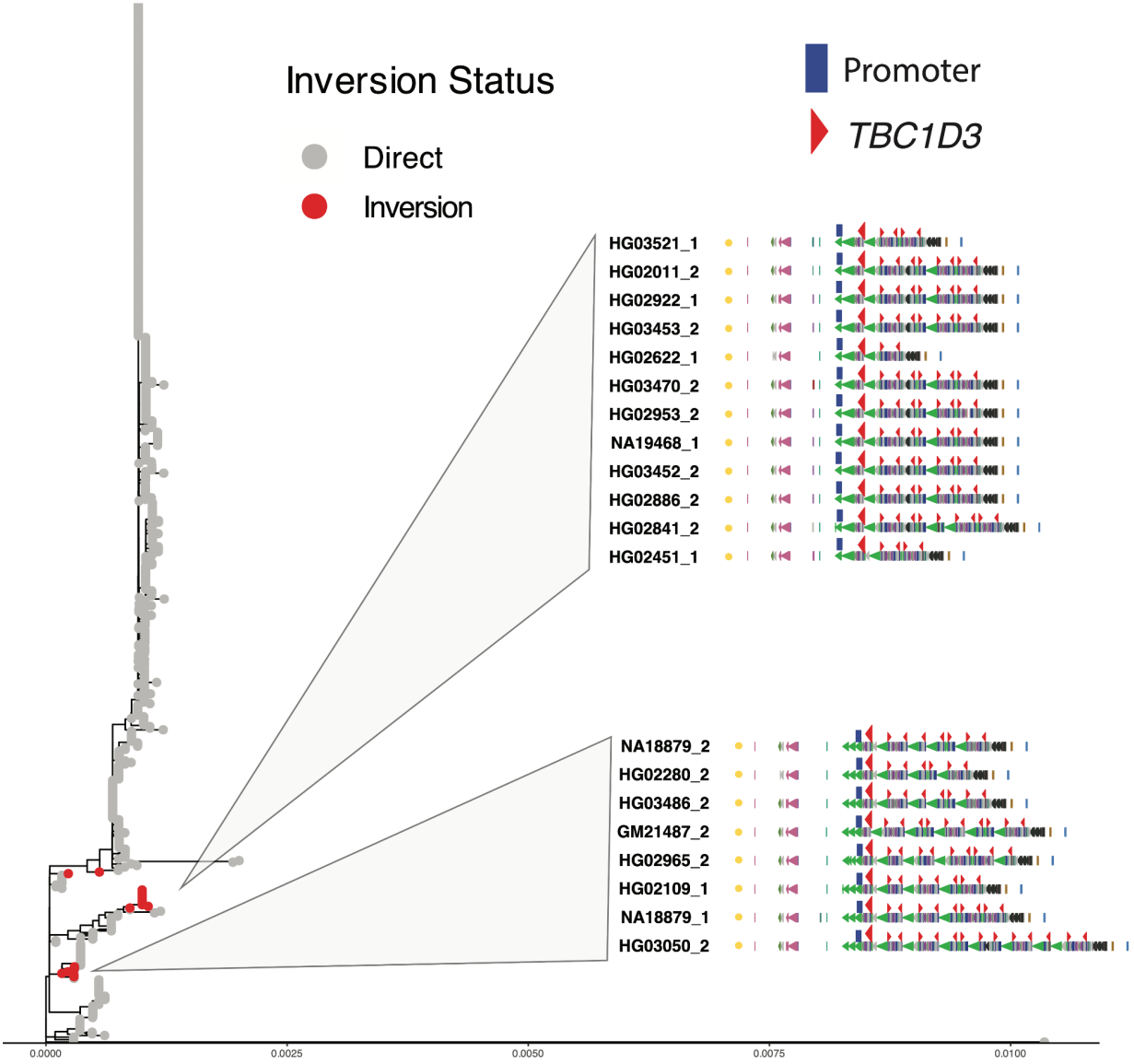


Figure 5. 1 Inversion haplogroups

Shown are 544 haplotypes assembled by the HPRC and HGSVC organized phylogenetically based on a multiple sequence alignment of flanking unique sequence. Two haplogroups (n=20) are put in focus, both of which carry an inversion of the terminal *TBC1D3* paralog (red arrow) relative to the *NPEPPSP1–TBC1D3* promoter.

The converse of genome-wide association studies (GWAS) are phenome-wide association studies (PheWAS; Pendergrass *et al.* 2012). Analogous to GWAS, PheWAS begins with a genetic variant of interest—such as a SNP or a defined structural haplotype—and tests for enrichment of that variant across a broad spectrum of phenotypes. Further work will be required to assess the phenotypic consequences associated with the two *TBC1D3* inversion haplogroups identified here (Fig. 5.1). Tagging SNPs capable of distinguishing these haplogroups can be used to query large-scale population datasets, including the UK Biobank (Sudlow *et al*. 2015), the *All of Us* Research Program, and deCODE Genetics. To date, no statistically significant phenotype associations have been detected. This limitation likely reflects both the low allele frequencies of the two haplogroups (2.2% [12/544] and 1.5% [8/544]) and their restricted population distribution, as both are currently observed exclusively in individuals of African ancestry. As a result, substantially larger and more ancestrally diverse cohorts will be required to rigorously evaluate potential phenotypic associations of these inversion haplogroups.

5.3 Closing Thoughts and Future Directions

I began my thesis work at the advent of long-read, high-fidelity sequencing and genome assembly. In the six years since I started my thesis in 2020, *de novo* assembly and haplotype-level validation have transitioned from experimental to routine (Antipov *et al*. 2025; Liao *et al*. 2023). In parallel, advances in phasing methodologies have reduced reliance on trio-based data; in many cases, haplotypes can now be resolved directly from a single individual using integrated sequencing and assembly approaches (Rautiainen *et al.* 2023). Together, these developments have enabled comprehensive population-scale characterization of structural variation and have

yielded important evolutionary and mechanistic insights into the human genome, including mutation rate (Porubsky *et al*. 2025), gene conversion (Vollger *et al*. 2023), and recurrent inversion events (Porubsky *et al.* 2022).

At the same time, *de novo* genome assembly is beginning to be leveraged in medical genetics. Long-read sequencing has already enabled the resolution of previously unsolved rare Mendelian disorders, providing closure for patients and families and, in some cases, informing improved clinical management and treatment strategies (Dawood *et al*. 2025). I anticipate that long-read sequencing will play an increasingly central role in genetic medicine. In particular, it is poised to transform diagnostics for rare disease, where conventional approaches often fail. Approximately one in ten Americans is affected by a rare disease, and an estimated 72% of these conditions have a genetic basis. Earlier and more accurate genetic diagnoses have the potential not only to guide treatment but also empower individuals with actionable knowledge about their health and risk.

TBC1D3 represents just one of many complex genic loci embedded within the human genome. I have devoted much of the past six years to studying this locus, and while I am naturally inclined to believe it will ultimately be linked to developmental phenotypes or disease, this remains an open question. What is clear, however, is that loci of comparable complexity—characterized by duplication, structural variation, and human-specific regulatory innovation—are likely to play critical roles in development and disease. As sequencing and *de novo* assembly resources continue to expand and are increasingly paired with rich, ancestrally diverse phenotypic data, it will become possible to rigorously interrogate these regions through association studies. Such efforts will both deepen our understanding of human evolution and development as well as

illuminate the genetic basis of Mendelian and complex disorders that have long remained unresolved.

REFERENCES

1. Agrawal, N., Dasaradhi, P. V. N., Mohmmed, A., Malhotra, P., Bhatnagar, R. K., & Mukherjee, S. K. (2003). RNA Interference: Biology, Mechanism, and Applications. *Microbiology and Molecular Biology Reviews*, *67*(4), 657–685. https://doi.org/10.1128/MMBR.67.4.657-685.2003
2. Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews. Genetics*, *12*(5), 363–376. https://doi.org/10.1038/nrg2958
3. Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., & Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, *23*(1), 258. https://doi.org/10.1186/s13059-022-02823-7
4. Ambreen, S., Khalil, F., & Abbasi, A. A. (2014). Integrating large-scale phylogenetic datasets to dissect the ancient evolutionary history of vertebrate genome. *Molecular Phylogenetics and Evolution*, *78*, 1–13. https://doi.org/10.1016/j.ympev.2014.05.002
5. Antipov, D., Rautiainen, M., Nurk, S., Walenz, B. P., Solar, S. J., Phillippy, A. M., & Koren, S. (2025). Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. *Genome Research*, *35*(7), 1583–1594. https://doi.org/10.1101/gr.280383.124
6. Antonacci, F., Kidd, J. M., Marques-Bonet, T., Teague, B., Ventura, M., Girirajan, S., Alkan, C., Campbell, C. D., Vives, L., Malig, M., Rosenfeld, J. A., Ballif, B. C., Shaffer, L. G., Graves, T. A., Wilson, R. K., Schwartz, D. C., & Eichler, E. E. (2010). A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nature Genetics*, *42*(9), 745–750. https://doi.org/10.1038/ng.643
7. Armengol, L., Pujana, M. A., Cheung, J., Scherer, S. W., & Estivill, X. (2003). Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Human Molecular Genetics*, *12*(17), 2201–2208. https://doi.org/10.1093/hmg/ddg223
8. *Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing | Genome Biology | Full Text*. (2024, February 23). https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02823-7
9. Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., … National Eye Institute, N. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), Article 7571. https://doi.org/10.1038/nature15393
10. Bailey, J. A., & Eichler, E. E. (2006). Primate segmental duplications: Crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, *7*(7), Article 7. https://doi.org/10.1038/nrg1895
11. Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., & Eichler, E. E. (2001). Segmental Duplications: Organization and Impact Within the Current Human Genome Project Assembly. *Genome Research*, *11*(6), 1005–1017. https://doi.org/10.1101/gr.187101
12. Barkan, A. (1988). Proteins encoded by a complex chloroplast transcription unit are each translated from both monocistronic and polycistronic mRNAs. *The EMBO Journal*, *7*(9), 2637–2644. https://doi.org/10.1002/j.1460-2075.1988.tb03116.x
13. Bar-Yosef, O., & Belfer-Cohen, A. (2001). From Africa to Eurasia—Early dispersals. *Quaternary International*, *75*(1), 19–28. https://doi.org/10.1016/S1040-6182(00)00074-4
14. Bekpen, C., Tastekin, I., Siswara, P., Akdis, C. A., & Eichler, E. E. (2012). Primate segmental duplication creates novel promoters for the LRRC37 gene family within the 17q21.31 inversion polymorphism region. *Genome Research*, *22*(6), 1050–1058. https://doi.org/10.1101/gr.134098.111

15. Bekpen, C., & Tautz, D. (2019). Human core duplicon gene families: Game changers or game players? *Briefings in Functional Genomics*, *18*(6), 402–411. https://doi.org/10.1093/bfgp/elz016
16. Bitar, M., Kuiper, S., O'Brien, E. A., & Barry, G. (2019). Genes with human-specific features are primarily involved withbrain, immune and metabolic evolution. *BMC Bioinformatics*, *20*(9), 406. https://doi.org/10.1186/s12859-019-2886-2
17. Blumenthal, T., Davis, P., & Garrido-Lecca, A. (2018). Operon and non-operon gene clusters in the C. elegans genome. In *WormBook: The Online Review of C. elegans Biology [Internet]*. WormBook. https://www.ncbi.nlm.nih.gov/books/NBK293639/
18. Bolognini, D., Halgren, A., Lou, R. N., Raveane, A., Rocha, J. L., Guarracino, A., Soranzo, N., Chin, C.-S., Garrison, E., & Sudmant, P. H. (2024). Recurrent evolution and selection shape structural diversity at the amylase locus. *Nature*, *634*(8034), 617–625. https://doi.org/10.1038/s41586-024-07911-1
19. Bornschein, U., Zeberg, H., Enard, W., Hevers, W., & Pääbo, S. (2023). Functional dissection of two amino acid substitutions unique to the human FOXP2 protein. *Scientific Reports*, *13*(1), 3747. https://doi.org/10.1038/s41598-023-30663-3
20. Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., Plessis, L. du, Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, *15*(4), e1006650. https://doi.org/10.1371/journal.pcbi.1006650
21. Bukhman, Y. V., Morin, P. A., Meyer, S., Chu, L.-F., Jacobsen, J. K., Antosiewicz-Bourget, J., Mamott, D., Gonzales, M., Argus, C., Bolin, J., Berres, M. E., Fedrigo, O., Steill, J., Swanson, S. A., Jiang, P., Rhie, A., Formenti, G., Phillippy, A. M., Harris, R. S., … Stewart, R. (2024). A High-Quality Blue Whale Genome, Segmental Duplications, and Historical Demography. *Molecular Biology and Evolution*, *41*(3), msae036. https://doi.org/10.1093/molbev/msae036
22. Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120
23. Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., & Wittwer, C. T. (2009). The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, *55*(4), 611–622. https://doi.org/10.1373/clinchem.2008.112797
24. Capell, A., Fellerer, K., & Haass, C. (2014). Progranulin transcripts with short and long 5' untranslated regions (UTRs) are differentially expressed via posttranscriptional and translational repression. *The Journal of Biological Chemistry*, *289*(37), 25879–25889. https://doi.org/10.1074/jbc.M114.560128
25. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), Article 2. https://doi.org/10.1038/s41592-020-01056-5
26. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science (New York, N.Y.)*, *381*(6664), eadg7492. https://doi.org/10.1126/science.adg7492
27. Cheng, Y.-H. H., Sedeño-Cortés, A. E., Ranchalis, J. E., Munson, K. M., Vollger, M. R., Balton, E., Genetti, C. A., Wojcik, M. H., Beggs, A. H., Bamshad, M. J., Wei, C.-L., Dipple, K. M., Kumar, R. D., Blue, E. E., Jarvik, G., Chong, J. X., Witten, D. M., O'Donnell-Luria, A., & Stergachis, A. B. (2025). Long-read transcriptome analysis using IsoRanker for identifying pathogenic variants in Mendelian conditions. *medRxiv*, 2025.11.07.25339764. https://doi.org/10.1101/2025.11.07.25339764
28. Cheung, W. A., Johnson, A. F., Rowell, W. J., Farrow, E., Hall, R., Cohen, A. S. A., Means, J. C., Zion, T. N., Portik, D. M., Saunders, C. T., Koseva, B., Bi, C., Truong, T. K., Schwendinger-Schreck, C.,

Yoo, B., Johnston, J. J., Gibson, M., Evrony, G., Rizzo, W. B., … Pastinen, T. (2023). Direct haplotype-resolved 5-base HiFi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-38782-1

29. Clarke, R., Peden, J. F., Hopewell, J. C., Kyriakou, T., Goel, A., Heath, S. C., Parish, S., Barlera, S., Franzosi, M. G., Rust, S., Bennett, D., Silveira, A., Malarstig, A., Green, F. R., Lathrop, M., Gigante, B., Leander, K., Faire, U. de, Seedorf, U., … Farrall, M. (2009). Genetic Variants Associated with Lp(a) Lipoprotein Level and Coronary Disease. *New England Journal of Medicine*, *361*(26), 2518–2528. https://doi.org/10.1056/NEJMoa0902604
30. Coorens, T. H. H., Guillaumet-Adkins, A., Kovner, R., Linn, R. L., Roberts, V. H. J., Sule, A., & Van Hoose, P. M. (2025). The human and non-human primate developmental GTEx projects. *Nature*, *637*(8046), 557–564. https://doi.org/10.1038/s41586-024-08244-9
31. Dawood, M., Heavner, B., Wheeler, M. M., Ungar, R. A., LoTempio, J., Wiel, L., Berger, S., Bernstein, J. A., Chong, J. X., Délot, E. C., Eichler, E. E., Lupski, J. R., Shojaie, A., Talkowski, M. E., Wagner, A. H., Wei, C.-L., Wellington, C., Wheeler, M. T., GREGoR Partner Members, … Genomics Research to Elucidate the Genetics of Rare Diseases (GREGoR) Consortium. (2025). GREGoR: Accelerating genomics for rare diseases. *Nature*, *647*(8089), 331–342. https://doi.org/10.1038/s41586-025-09613-8
32. *deCODE in Iceland agrees to sequence half of UK Biobank participants*. (2019, September 11). deCODE Genetics. https://www.decode.com/decode-in-iceland-agrees-to-sequence-half-of-uk-biobank-participants/
33. Dennis, M. Y., Harshman, L., Nelson, B. J., Penn, O., Cantsilieris, S., Huddleston, J., Antonacci, F., Penewit, K., Denman, L., Raja, A., Baker, C., Mark, K., Malig, M., Janke, N., Espinoza, C., Stessman, H. A. F., Nuttle, X., Hoekzema, K., Lindsay-Graves, T. A., … Eichler, E. E. (2017). The evolution and population diversity of human-specific segmental duplications. *Nature Ecology & Evolution*, *1*(3), 69. https://doi.org/10.1038/s41559-016-0069
34. Dennis, M. Y., Nuttle, X., Sudmant, P. H., Antonacci, F., Graves, T. A., Nefedov, M., Rosenfeld, J. A., Sajjadian, S., Malig, M., Kotkiewicz, H., Curry, C. J., Shafer, S., Shaffer, L. G., de Jong, P. J., Wilson, R. K., & Eichler, E. E. (2012). Evolution of Human-Specific Neural *SRGAP2* Genes by Incomplete Segmental Duplication. *Cell*, *149*(4), 912–922. https://doi.org/10.1016/j.cell.2012.03.033
35. Deretic, D., Williams, A. H., Ransom, N., Morel, V., Hargrave, P. A., & Arendt, A. (2005). Rhodopsin C terminus, the site of mutations causing retinal disease, regulates trafficking by binding to ADP-ribosylation factor 4 (ARF4). *Proceedings of the National Academy of Sciences*, *102*(9), 3301–3306. https://doi.org/10.1073/pnas.0500095102
36. Dibbens, L. M., Mullen, S., Helbig, I., Mefford, H. C., Bayly, M. A., Bellows, S., Leu, C., Trucks, H., Obermeier, T., Wittig, M., Franke, A., Caglayan, H., Yapici, Z., EPICURE Consortium, Sander, T., Eichler, E. E., Scheffer, I. E., Mulley, J. C., & Berkovic, S. F. (2009). Familial and sporadic 15q13.3 microdeletions in idiopathic generalized epilepsy: Precedent for disorders with complex inheritance. *Human Molecular Genetics*, *18*(19), 3626–3631. https://doi.org/10.1093/hmg/ddp311
37. Dishuck, P. C., Munson, K. M., Lewis, A. P., Dougherty, M. L., Underwood, J. G., Harvey, W. T., Hsieh, P., Pastinen, T., & Eichler, E. E. (2025). Structural variation, selection, and diversification of the NPIP gene family from the human pangenome. *Cell Genomics*, *5*(10), 100977. https://doi.org/10.1016/j.xgen.2025.100977
38. Dishuck, P. C., Rozanski, A. N., Logsdon, G. A., Porubsky, D., & Eichler, E. E. (2022). GAVISUNK: Genome assembly validation via inter-SUNK distances in Oxford Nanopore reads. *Bioinformatics*, *39*(1), btac714. https://doi.org/10.1093/bioinformatics/btac714
39. Dong, J., Zhu, X.-N., Zeng, P.-M., Cao, D.-D., Yang, Y., Hu, J., & Luo, Z.-G. (2024). A hominoid-specific signaling axis regulating the tempo of synaptic maturation. *Cell Reports*, *43*(8), 114548. https://doi.org/10.1016/j.celrep.2024.114548
40. Dougherty, M. L., Nuttle, X., Penn, O., Nelson, B. J., Huddleston, J., Baker, C., Harshman, L., Duyzend, M. H., Ventura, M., Antonacci, F., Sandstrom, R., Dennis, M. Y., & Eichler, E. E. (2017). The birth of a human-specific neural gene by incomplete duplication and gene fusion. *Genome Biology*, *18*(1), 49. https://doi.org/10.1186/s13059-017-1163-9

41. Dougherty, M. L., Underwood, J. G., Nelson, B. J., Tseng, E., Munson, K. M., Penn, O., Nowakowski, T. J., Pollen, A. A., & Eichler, E. E. (2018). Transcriptional fates of human-specific segmental duplications in brain. *Genome Research*, *28*(10), 1566–1576. https://doi.org/10.1101/gr.237610.118
42. Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, *29*(8), 1969–1973. https://doi.org/10.1093/molbev/mss075
43. Dunsworth, H. M. (2010). Origin of the Genus Homo. *Evolution: Education and Outreach*, *3*(3), Article 3. https://doi.org/10.1007/s12052-010-0247-8
44. Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Mari, R. S., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science (New York, N.Y.)*, *372*(6537), eabf7117. https://doi.org/10.1126/science.abf7117
45. Enard, W., Przeworski, M., Fisher, S. E., Lai, C. S. L., Wiebe, V., Kitano, T., Monaco, A. P., & Pääbo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, *418*(6900), 869–872. https://doi.org/10.1038/nature01025
46. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247
47. EPI2ME. (2025). *Epi2me-labs/modbam2bed* [C]. https://github.com/epi2me-labs/modbam2bed (Original work published 2021)
48. Fairley, S., Lowy-Gallego, E., Perry, E., & Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, *48*(D1), D941–D947. https://doi.org/10.1093/nar/gkz836
49. Fiddes, I. T., Lodewijk, G. A., Mooring, M., Bosworth, C. M., Ewing, A. D., Mantalas, G. L., Novak, A. M., van den Bout, A., Bishara, A., Rosenkrantz, J. L., Lorig-Roach, R., Field, A. R., Haeussler, M., Russo, L., Bhaduri, A., Nowakowski, T. J., Pollen, A. A., Dougherty, M. L., Nuttle, X., … Haussler, D. (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell*, *173*(6), 1356-1369.e22. https://doi.org/10.1016/j.cell.2018.03.051
50. Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F. K., Peters, J., Guhr, E., Klemroth, S., Prüfer, K., Kelso, J., Naumann, R., Nüsslein, I., Dahl, A., Lachmann, R., Pääbo, S., & Huttner, W. B. (2015). Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. *Science (New York, N.Y.)*, *347*(6229), 1465–1470. https://doi.org/10.1126/science.aaa1975
51. Florio, M., Heide, M., Pinson, A., Brandl, H., Albert, M., Winkler, S., Wimberger, P., Huttner, W. B., & Hiller, M. (2018). Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *eLife*, *7*, e32332. https://doi.org/10.7554/eLife.32332
52. Frittoli, E., Palamidessi, A., Pizzigoni, A., Lanzetti, L., Garrè, M., Troglio, F., Troilo, A., Fukuda, M., Di Fiore, P. P., Scita, G., & Confalonieri, S. (2008). The Primate-specific Protein TBC1D3 Is Required for Optimal Macropinocytosis in a Novel ARF6-dependent Pathway. *Molecular Biology of the Cell*, *19*(4), 1304–1316. https://doi.org/10.1091/mbc.e07-06-0594
53. Gao, S., Oshima, K. K., Chuang, S.-C., Loftus, M., Montanari, A., Gordon, D. S., Consortium, H. G. S. V., Consortium, H. P. R., Hsieh, P., Konkel, M. K., Ventura, M., & Logsdon, G. A. (2025). *A global view of human centromere variation and evolution* (p. 2025.12.09.693231). bioRxiv. https://doi.org/10.64898/2025.12.09.693231
54. García-Ríos, M., Fujita, T., LaRosa, P. C., Locy, R. D., Clithero, J. M., Bressan, R. A., & Csonka, L. N. (1997). Cloning of a polycistronic cDNA from tomato encoding gamma-glutamyl kinase and gamma-glutamyl phosphate reductase. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(15), 8249–8254. https://doi.org/10.1073/pnas.94.15.8249
55. *GenomeArck*. (n.d.). [Dataset]. Retrieved https://registry.opendata.aws/genomeark
56. Gershman, A., Sauria, M. E. G., Guitart, X., Vollger, M. R., Hook, P. W., Hoyt, S. J., Jain, M., Shumate, A., Razaghi, R., Koren, S., Altemose, N., Caldas, G. V., Logsdon, G. A., Rhie, A., Eichler, E. E.,

Schatz, M. C., O'Neill, R. J., Phillippy, A. M., Miga, K. H., & Timp, W. (2022). Epigenetic patterns in a complete human genome. *Science (New York, N.Y.)*, *376*(6588), eabj5089. https://doi.org/10.1126/science.abj5089

57. Gray, T. A., Saitoh, S., & Nicholls, R. D. (1999). An imprinted, mammalian bicistronic transcript encodes two independent proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(10), 5616–5621. https://doi.org/10.1073/pnas.96.10.5616
58. Guitart, X., Brunner, J. W., Ren, L., Jeong, H., Yoo, D., Porubsky, D., Hoekzema, K., Munson, K. M., Sun, K. A., Ayllon, M., Hoglin, K., McMullen, R., Pavlovic, B., Vollger, M. R., Pollen, A. A., & Eichler, E. E. (2026). *NPEPPS segmental duplication drives position effect expression of TBC1D3 in the human brain* (p. 2026.01.14.699559). bioRxiv. https://doi.org/10.64898/2026.01.14.699559
59. Guitart, X., Porubsky, D., Yoo, D., Dougherty, M. L., Dishuck, P. C., Munson, K. M., Lewis, A. P., Hoekzema, K., Knuth, J., Chang, S., Pastinen, T., & Eichler, E. E. (2024). Independent expansion, selection, and hypervariability of the TBC1D3 gene family in humans. *Genome Research*, *34*(11), 1798–1810. https://doi.org/10.1101/gr.279299.124
60. *GWAS Catalog*. (2025, December 31). https://www.ebi.ac.uk/gwas/
61. Hayashi, T., Motulsky, A. G., & Deeb, S. S. (1999). Position of a "green-red" hybrid gene in the visual pigment array determines colour-vision phenotype. *Nature Genetics*, *22*(1), Article 1. https://doi.org/10.1038/8798
62. Heffer, A., & Pick, L. (2013). Conservation and Variation in *Hox* Genes: How Insect Models Pioneered the Evo-Devo Field. *Annual Review of Entomology*, *58*(1), 161–179. https://doi.org/10.1146/annurev-ento-120811-153601
63. Hellen, C. U. T., & Sarnow, P. (2001). Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes & Development*, *15*(13), 1593–1612. https://doi.org/10.1101/gad.891101
64. Higgs, D. R., Goodbourn, S. E., Lamb, J., Clegg, J. B., Weatherall, D. J., & Proudfoot, N. J. (1983). Alpha-thalassaemia caused by a polyadenylation signal mutation. *Nature*, *306*(5941), 398–400. https://doi.org/10.1038/306398a0
65. Hirschberg, Y., Valle-Tamayo, N., Dols-Icardo, O., Engelborghs, S., Buelens, B., Vandenbroucke, R. E., Vermeiren, Y., Boonen, K., & Mertens, I. (2023). Proteomic comparison between non-purified cerebrospinal fluid and cerebrospinal fluid-derived extracellular vesicles from patients with Alzheimer's, Parkinson's and Lewy body dementia. *Journal of Extracellular Vesicles*, *12*(12), 12383. https://doi.org/10.1002/jev2.12383
66. Ho, Y., Elefant, F., Liebhaber, S. A., & Cooke, N. E. (2006). Locus Control Region Transcription Plays an Active Role in Long-Range Gene Activation. *Molecular Cell*, *23*(3), 365–375. https://doi.org/10.1016/j.molcel.2006.05.041
67. Hodzic, D., Kong, C., Wainszelbaum, M. J., Charron, A. J., Su, X., & Stahl, P. D. (2006). TBC1D3, a hominoid oncoprotein, is encoded by a cluster of paralogues located on chromosome 17q12. *Genomics*, *88*(6), 731–736. https://doi.org/10.1016/j.ygeno.2006.05.009
68. Holland, P. W. H. (2013). Evolution of homeobox genes. *Wiley Interdisciplinary Reviews. Developmental Biology*, *2*(1), 31–45. https://doi.org/10.1002/wdev.78
69. Hou, Q.-Q., Xiao, Q., Sun, X.-Y., Ju, X.-C., & Luo, Z.-G. (2021). TBC1D3 promotes neural progenitor proliferation by suppressing the histone methyltransferase G9a. *Science Advances*, *7*(3), eaba8053. https://doi.org/10.1126/sciadv.aba8053
70. Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics*, *116*(1), 153–159.
71. Hughes, J. F., Skaletsky, H., Brown, L. G., Pyntikova, T., Graves, T., Fulton, R. S., Dugan, S., Ding, Y., Buhay, C. J., Kremitzki, C., Wang, Q., Shen, H., Holder, M., Villasana, D., Nazareth, L. V., Cree, A., Courtney, L., Veizer, J., Kotkiewicz, H., … Page, D. C. (2012). Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature*, *483*(7387), 82–86. https://doi.org/10.1038/nature10843

72. Hulme, A. J., Maksour, S., St-Clair Glover, M., Miellet, S., & Dottori, M. (2021). Making neurons, made easy: The use of Neurogenin-2 in neuronal differentiation. *Stem Cell Reports*, *17*(1), 14–34. https://doi.org/10.1016/j.stemcr.2021.11.015
73. Huttner, W. B., Heide, M., Mora-Bermúdez, F., & Namba, T. (2024). Neocortical neurogenesis in development and evolution—Human-specific features. *Journal of Comparative Neurology*, *532*(2), e25576. https://doi.org/10.1002/cne.25576
74. Itsara, A., Cooper, G. M., Baker, C., Girirajan, S., Li, J., Absher, D., Krauss, R. M., Myers, R. M., Ridker, P. M., Chasman, D. I., Mefford, H., Ying, P., Nickerson, D. A., & Eichler, E. E. (2009). Population Analysis of Large Copy Number Variants and Hotspots of Human Genetic Disease. *The American Journal of Human Genetics*, *84*(2), 148–161. https://doi.org/10.1016/j.ajhg.2008.12.014
75. Jeong, H., Dishuck, P. C., Yoo, D., Harvey, W. T., Munson, K. M., Lewis, A. P., Kordosky, J., Garcia, G. H., Consortium (HGSVC), H. G. S. V., Yilmaz, F., Hallast, P., Lee, C., Pastinen, T., & Eichler, E. E. (2024). *Structural polymorphism and diversity of human segmental duplications*. https://doi.org/10.1101/2024.06.04.597452
76. Jiang, Z., Hubley, R., Smit, A., & Eichler, E. E. (2008). DupMasker: A tool for annotating primate segmental duplications. *Genome Research*, *18*(8), 1362–1368. https://doi.org/10.1101/gr.078477.108
77. Jiang, Z., Tang, H., Ventura, M., Cardone, M. F., Marques-Bonet, T., She, X., Pevzner, P. A., & Eichler, E. E. (2007). Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nature Genetics*, *39*(11), 1361–1368. https://doi.org/10.1038/ng.2007.9
78. Johnson, M. E., National Institute of Health Intramural Sequencing Center Comparative Sequencing Program, Cheng, Z., Morrison, V. A., Scherer, S., Ventura, M., Gibbs, R. A., Green, E. D., & Eichler, E. E. (2006). Recurrent duplication-driven transposition of DNA during hominoid evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(47), 17626–17631. https://doi.org/10.1073/pnas.0605426103
79. Ju, X.-C., Hou, Q.-Q., Sheng, A.-L., Wu, K.-Y., Zhou, Y., Jin, Y., Wen, T., Yang, Z., Wang, X., & Luo, Z.-G. (2016). The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. *eLife*, *5*, e18197. https://doi.org/10.7554/eLife.18197
80. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
81. Kalebic, N., Gilardi, C., Albert, M., Namba, T., Long, K. R., Kostic, M., Langen, B., & Huttner, W. B. (2018). Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex. *eLife*, *7*, e41241. https://doi.org/10.7554/eLife.41241
82. Karsten, S. L., Sang, T.-K., Gehman, L. T., Chatterjee, S., Liu, J., Lawless, G. M., Sengupta, S., Berry, R. W., Pomakian, J., Oh, H. S., Schulz, C., Hui, K.-S., Wiedau-Pazos, M., Vinters, H. V., Binder, L. I., Geschwind, D. H., & Jackson, G. R. (2006). A genomic screen for modifiers of tauopathy identifies puromycin-sensitive aminopeptidase as an inhibitor of tau-induced neurodegeneration. *Neuron*, *51*(5), 549–560. https://doi.org/10.1016/j.neuron.2006.07.019
83. Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436
84. Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010
85. Kim, K.-W., De-Kayne, R., Gordon, I. J., Omufwoko, K. S., Martins, D. J., ffrench-Constant, R., & Martin, S. H. (2022). Stepwise evolution of a butterfly supergene via duplication and inversion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1856), 20210207. https://doi.org/10.1098/rstb.2021.0207

86. Kim, P. M., Lam, H. Y. K., Urban, A. E., Korbel, J. O., Affourtit, J., Grubert, F., Chen, X., Weissman, S., Snyder, M., & Gerstein, M. B. (2008). Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Research*, *18*(12), 1865–1874. https://doi.org/10.1101/gr.081422.108
87. Kim, T.-K., & Shiekhattar, R. (2015). Architectural and functional commonalities between enhancers and promoters. *Cell*, *162*(5), 948–959. https://doi.org/10.1016/j.cell.2015.08.008
88. King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, *188*(4184), 107–116. https://doi.org/10.1126/science.1090005
89. Kiryutin, B., Souvorov, A., & Tatusova, T. (2017). *ProSplign* [Computer software].
90. Kong, C., Samovski, D., Srikanth, P., Wainszelbaum, M. J., Charron, A. J., Liu, J., Lange, J. J., Chen, P.-I., Pan, Z.-Q., Su, X., & Stahl, P. D. (2012). Ubiquitination and Degradation of the Hominoid-Specific Oncoprotein TBC1D3 Is Mediated by CUL7 E3 Ligase. *PLoS ONE*, *7*(9), e46485. https://doi.org/10.1371/journal.pone.0046485
91. Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116
92. Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., Porubsky, D., Kuhn, K., Mueller, K. A., Low, W. Y., Hiendleder, S., Fedrigo, O., Liachko, I., Hall, R. J., Phillippy, A. M., Eichler, E. E., Williams, J. L., Smith, T. P. L., Jarvis, E. D., … Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-020-20536-y
93. Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K.-A., Mooij, T. M., Roos-Blom, M.-J., Jervis, S., van Leeuwen, F. E., Milne, R. L., Andrieu, N., Goldgar, D. E., Terry, M. B., Rookus, M. A., Easton, D. F., Antoniou, A. C., & and the BRCA1 and BRCA2 Cohort Consortium. (2017). Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA*, *317*(23), 2402–2416. https://doi.org/10.1001/jama.2017.7112
94. Kumara Mastrorosa, F., Oshima, K. K., Rozanski, A. N., Harvey, W. T., Eichler, E. E., & Logsdon, G. A. (2024). Identification and annotation of centromeric hypomethylated regions with Centromere Dip Region (CDR)-Finder. *bioRxiv*, 2024.11.01.621587. https://doi.org/10.1101/2024.11.01.621587
95. Kwiatkowski, D. P. (2005). How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *The American Journal of Human Genetics*, *77*(2), 171–192. https://doi.org/10.1086/432519
96. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … The Wellcome Trust: (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062
97. Lawrence, M., Gentleman, R., & Carey, V. (2009). rtracklayer: An R package for interfacing with genome browsers. *Bioinformatics*, *25*(14), 1841–1842. https://doi.org/10.1093/bioinformatics/btp328
98. Lee, J. A., Carvalho, C. M. B., & Lupski, J. R. (2007). A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell*, *131*(7), 1235–1247. https://doi.org/10.1016/j.cell.2007.11.037
99. Leung, S. K., Jeffries, A. R., Castanho, I., Jordan, B. T., Moore, K., Davies, J. P., Dempster, E. L., Bray, N. J., O'Neill, P., Tseng, E., Ahmed, Z., Collier, D. A., Jeffery, E. D., Prabhakar, S., Schalkwyk, L., Jops, C., Gandal, M. J., Sheynkman, G. M., Hannon, E., & Mill, J. (2021). Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. *Cell Reports*, *37*(7). https://doi.org/10.1016/j.celrep.2021.110022
100. Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. *Nature*, *276*(5688), 565–570. https://doi.org/10.1038/276565a0
101. Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

102. Li, H., Feng, X., & Chu, C. (2020). The design and construction of reference pangenome graphs with minigraph. *Genome Biology*, *21*(1), 265. https://doi.org/10.1186/s13059-020-02168-z
103. Li, X., Sun, H., Li, H., Li, D., Cai, Z., Xu, J., & Ma, R. (2023). A Single-Cell RNA-Sequencing Analysis of Distinct Subsets of Synovial Macrophages in Rheumatoid Arthritis. *DNA and Cell Biology*, *42*(4), 212–222. https://doi.org/10.1089/dna.2022.0509
104. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), Article 7960. https://doi.org/10.1038/s41586-023-05896-x
105. Liedigk, R., Roos, C., Brameier, M., & Zinner, D. (2014). Mitogenomics of the Old World monkey tribe Papionini. *BMC Evolutionary Biology*, *14*, 176. https://doi.org/10.1186/s12862-014-0176-1
106. Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews. Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x
107. Maggiolini, F. A. M., Cantsilieris, S., D'Addabbo, P., Manganelli, M., Coe, B. P., Dumont, B. L., Sanders, A. D., Pang, A. W. C., Vollger, M. R., Palumbo, O., Palumbo, P., Accadia, M., Carella, M., Eichler, E. E., & Antonacci, F. (2019). Genomic inversions and GOLGA core duplicons underlie disease instability at the 15q25 locus. *PLoS Genetics*, *15*(3), e1008075. https://doi.org/10.1371/journal.pgen.1008075
108. Makova, K. D., Pickett, B. D., Harris, R. S., Hartley, G. A., Cechova, M., Pal, K., Nurk, S., Yoo, D., Li, Q., Hebbar, P., McGrath, B. C., Antonacci, F., Aubel, M., Biddanda, A., Borchers, M., Bornberg-Bauer, E., Bouffard, G. G., Brooks, S. Y., Carbone, L., … Phillippy, A. M. (2024). *The Complete Sequence and Comparative Analysis of Ape Sex Chromosomes*. https://doi.org/10.1101/2023.11.30.569198
109. Mao, Y., Catacchio, C. R., Hillier, L. W., Porubsky, D., Li, R., Sulovari, A., Fernandes, J. D., Montinaro, F., Gordon, D. S., Storer, J. M., Haukness, M., Fiddes, I. T., Murali, S. C., Dishuck, P. C., Hsieh, P., Harvey, W. T., Audano, P. A., Mercuri, L., Piccolo, I., … Eichler, E. E. (2021). A high-quality bonobo genome refines the analysis of hominid evolution. *Nature*, *594*(7861), 77–81. https://doi.org/10.1038/s41586-021-03519-x
110. Mao, Y., Harvey, W. T., Porubsky, D., Munson, K. M., Hoekzema, K., Lewis, A. P., Audano, P. A., Rozanski, A., Yang, X., Zhang, S., Yoo, D., Gordon, D. S., Fair, T., Wei, X., Logsdon, G. A., Haukness, M., Dishuck, P. C., Jeong, H., Rosario, R. del, … Eichler, E. E. (2024). Structurally divergent and recurrently mutated regions of primate genomes. *Cell*, *0*(0). https://doi.org/10.1016/j.cell.2024.01.052
111. Mark, M., Rijli, F. M., & Chambon, P. (1997). Homeobox Genes in Embryogenesis and Pathogenesis. *Pediatric Research*, *42*(4), 421–429. https://doi.org/10.1203/00006450-199710000-00001
112. Marques-Bonet, T., & Eichler, E. E. (2009). The Evolution of Human Segmental Duplications and the Core Duplicon Hypothesis. *Cold Spring Harbor Symposia on Quantitative Biology*, *74*, 355–362. https://doi.org/10.1101/sqb.2009.74.011
113. Marques-Bonet, T., Girirajan, S., & Eichler, E. E. (2009). The origins and impact of primate segmental duplications. *Trends in Genetics : TIG*, *25*(10), 443–454. https://doi.org/10.1016/j.tig.2009.08.002
114. Marques-Bonet, T., Kidd, J. M., Ventura, M., Graves, T. A., Cheng, Z., Hillier, L. W., Jiang, Z., Baker, C., Malfavon-Borja, R., Fulton, L. A., Alkan, C., Aksay, G., Girirajan, S., Siswara, P., Chen, L., Cardone, M. F., Navarro, A., Mardis, E. R., Wilson, R. K., & Eichler, E. E. (2009). A Burst of Segmental Duplications in the African Great Ape Ancestor. *Nature*, *457*(7231), 877–881. https://doi.org/10.1038/nature07744
115. Mastoras, M., Asri, M., Brambrink, L., Hebbar, P., Kolesnikov, A., Cook, D. E., Nattestad, M., Lucas, J., Won, T. S., Chang, P.-C., Carroll, A., Paten, B., Shafin, K., & and the Human Pangenome

Reference Consortium. (2025). Highly accurate assembly polishing with DeepPolisher. *Genome Research*, *35*(7), 1595–1608. https://doi.org/10.1101/gr.280149.124

116. McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), Article 6328. https://doi.org/10.1038/351652a0

117. McLellan, R. A., Oscarson, M., Alexandrie, A. K., Seidegård, J., Evans, D. A., Rannug, A., & Ingelman-Sundberg, M. (1997). Characterization of a human glutathione S-transferase mu cluster containing a duplicated GSTM1 gene that causes ultrarapid enzyme activity. *Molecular Pharmacology*, *52*(6), 958–965. https://doi.org/10.1124/mol.52.6.958

118. McLellan, S., Dyer, S. H., Rodriguez, G., & Hersh, L. B. (1988). Studies on the Tissue Distribution of the Puromycin-Sensitive Enkephalin-Degrading Aminopeptidases. *Journal of Neurochemistry*, *51*(5), 1552–1559. https://doi.org/10.1111/j.1471-4159.1988.tb01124.x

119. Mefford, H. (2016). 17q12 Recurrent Duplication. In M. P. Adam, S. Bick, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, & A. Amemiya (Eds.), *GeneReviews®*. University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK344340/

120. Mefford, H. C., Clauin, S., Sharp, A. J., Moller, R. S., Ullmann, R., Kapur, R., Pinkel, D., Cooper, G. M., Ventura, M., Ropers, H. H., Tommerup, N., Eichler, E. E., & Bellanne-Chantelot, C. (2007). Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *American Journal of Human Genetics*, *81*(5), 1057–1069. https://doi.org/10.1086/522591

121. Mercer, T. R., Neph, S., Dinger, M. E., Crawford, J., Smith, M. A., Shearwood, A.-M. J., Haugen, E., Bracken, C. P., Rackham, O., Stamatoyannopoulos, J. A., Filipovska, A., & Mattick, J. S. (2011). The human mitochondrial transcriptome. *Cell*, *146*(4), 645–658. https://doi.org/10.1016/j.cell.2011.06.051

122. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

123. Mitchel, M. W., Moreno-De-Luca, D., Myers, S. M., Levy, R. V., Turner, S., Ledbetter, D. H., & Martin, C. L. (2016). 17q12 Recurrent Deletion Syndrome. In M. P. Adam, S. Bick, G. M. Mirzaa, R. A. Pagon, S. E. Wallace, & A. Amemiya (Eds.), *GeneReviews®*. University of Washington, Seattle. http://www.ncbi.nlm.nih.gov/books/NBK401562/

124. Mohajeri, K., Cantsilieris, S., Huddleston, J., Nelson, B. J., Coe, B. P., Campbell, C. D., Baker, C., Harshman, L., Munson, K. M., Kronenberg, Z. N., Kremitzki, M., Raja, A., Catacchio, C. R., Graves, T. A., Wilson, R. K., Ventura, M., & Eichler, E. E. (2016). Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. *Genome Research*, *26*(11), 1453–1467. https://doi.org/10.1101/gr.211284.116

125. Muraoka, S., Jedrychowski, M. P., Yanamandra, K., Ikezu, S., Gygi, S. P., & Ikezu, T. (2020). Proteomic Profiling of Extracellular Vesicles Derived from Cerebrospinal Fluid of Alzheimer's Disease Patients: A Pilot Study. *Cells*, *9*(9), 1959. https://doi.org/10.3390/cells9091959

126. Nachman, M. W., Hoekstra, H. E., & D'Agostino, S. L. (2003). The genetic basis of adaptive melanism in pocket mice. *Proceedings of the National Academy of Sciences*, *100*(9), 5268–5273. https://doi.org/10.1073/pnas.0431157100

127. Namba, T., Dóczi, J., Pinson, A., Xing, L., Kalebic, N., Wilsch-Bräuninger, M., Long, K. R., Vaid, S., Lauer, J., Bogdanova, A., Borgonovo, B., Shevchenko, A., Keller, P., Drechsel, D., Kurzchalia, T., Wimberger, P., Chinopoulos, C., & Huttner, W. B. (2020). Human-Specific ARHGAP11B Acts in Mitochondria to Expand Neocortical Progenitors by Glutaminolysis. *Neuron*, *105*(5), 867-881.e9. https://doi.org/10.1016/j.neuron.2019.11.027

128. Numanagic, I., Gökkaya, A. S., Zhang, L., Berger, B., Alkan, C., & Hach, F. (2018). Fast characterization of segmental duplications in genome assemblies. *Bioinformatics (Oxford, England)*, *34*(17), i706–i714. https://doi.org/10.1093/bioinformatics/bty586

129. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A.,

Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science (New York, N.Y.)*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

130. Nuttle, X., Giannuzzi, G., Duyzend, M. H., Schraiber, J. G., Narvaiza, I., Sudmant, P. H., Penn, O., Chiatante, G., Malig, M., Huddleston, J., Benner, C., Camponeschi, F., Ciofi-Baffoni, S., Stessman, H. A. F., Marchetto, M. C. N., Denman, L., Harshman, L., Baker, C., Raja, A., … Eichler, E. E. (2016). Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature*, *536*(7615), 205–209. https://doi.org/10.1038/nature19075
131. Ohno, Susumu. (1970). *Evolution by Gene Duplication*. Springer Berlin, Heidelberg.
132. O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189
133. Osoegawa, K., Woon, P. Y., Zhao, B., Frengen, E., Tateno, M., Catanese, J. J., & de Jong, P. J. (1998). An Improved Approach for Construction of Bacterial Artificial Chromosome Libraries. *Genomics*, *52*(1), 1–8. https://doi.org/10.1006/geno.1998.5423
134. Ousley, O., Rockers, K., Dell, M. L., Coleman, K., & Cubells, J. F. (2007). A review of neurocognitive and behavioral profiles associated with 22q11 deletion syndrome: Implications for clinical evaluation and treatment. *Current Psychiatry Reports*, *9*(2), 148–158. https://doi.org/10.1007/s11920-007-0085-8
135. Pacific Biosciences. (n.d.). *Isoseq3* (Version 4.0.0) [Computer software].
136. Paulding, C. A., Ruvolo, M., & Haber, D. A. (2003). The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proceedings of the National Academy of Sciences*, *100*(5), 2507–2511. https://doi.org/10.1073/pnas.0437015100
137. Pei, L., Peng, Y., Yang, Y., Ling, X. B., van Eyndhoven, W. G., Nguyen, K. C. Q., Rubin, M., Hoey, T., Powers, S., & Li, J. (2002). PRC17, a Novel Oncogene Encoding a Rab GTPase-activating Protein, Is Amplified in Prostate Cancer. *Cancer Research*, *62*(19), 5420–5424.
138. Pendergrass, S. A., Dudek, S. M., Crawford, D. C., & Ritchie, M. D. (2012). Visually integrating and exploring high throughput Phenome-Wide Association Study (PheWAS) results using PheWAS-View. *BioData Mining*, *5*, 5. https://doi.org/10.1186/1756-0381-5-5
139. Pendleton, A. L., Shen, F., Taravella, A. M., Emery, S., Veeramah, K. R., Boyko, A. R., & Kidd, J. M. (2018). Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biology*, *16*(1), 64. https://doi.org/10.1186/s12915-018-0535-2
140. Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., & Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, *39*(10), Article 10. https://doi.org/10.1038/ng2123
141. Pinheiro, L. B., Coleman, V. A., Hindson, C. M., Herrmann, J., Hindson, B. J., Bhat, S., & Emslie, K. R. (2012). Evaluation of a Droplet Digital Polymerase Chain Reaction Format for DNA Copy Number Quantification. *Analytical Chemistry*, *84*(2), 1003–1011. https://doi.org/10.1021/ac202578x
142. Plender, E. G., Prodanov, T., Hsieh, P., Nizamis, E., Harvey, W. T., Sulovari, A., Munson, K. M., Kaufman, E. J., O'Neal, W. K., Valdmanis, P. N., Marschall, T., Bloom, J. D., & Eichler, E. E. (2024). Structural and genetic diversity in the secreted mucins MUC5AC and MUC5B. *The American Journal of Human Genetics*, *111*(8), 1700–1716. https://doi.org/10.1016/j.ajhg.2024.06.007
143. Popesco, M. C., Maclaren, E. J., Hopkins, J., Dumas, L., Cox, M., Meltesen, L., McGavran, L., Wyckoff, G. J., & Sikela, J. M. (2006). Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science (New York, N.Y.)*, *313*(5791), 1304–1307. https://doi.org/10.1126/science.1127980
144. Porubsky, D., Dashnow, H., Sasani, T. A., Logsdon, G. A., Hallast, P., Noyes, M. D., Kronenberg, Z. N., Mokveld, T., Koundinya, N., Nolan, C., Steely, C. J., Guarracino, A., Dolzhenko, E., Harvey,

W. T., Rowell, W. J., Grigorev, K., Nicholas, T. J., Goldberg, M. E., Oshima, K. K., … Eichler, E. E. (2025). Human de novo mutation rates from a four-generation pedigree reference. *Nature*, *643*(8071), 427–436. https://doi.org/10.1038/s41586-025-08922-2

145. Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F. A., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Human Genome Structural Variation Consortium (HGSVC), Lee, C., … Korbel, J. O. (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, *185*(11), 1986-2005.e26. https://doi.org/10.1016/j.cell.2022.04.017
146. Pramanik, S., Cui, X., Wang, H.-Y., Chimge, N.-O., Hu, G., Shen, L., Gao, R., & Li, H. (2011). Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region. *BMC Genomics*, *12*, 78. https://doi.org/10.1186/1471-2164-12-78
147. Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., Schierup, M. H., & Jensen, T. H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. *Science (New York, N.Y.)*, *322*(5909), 1851–1854. https://doi.org/10.1126/science.1164096
148. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033
149. Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., Eichler, E. E., Phillippy, A. M., & Koren, S. (2023). Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nature Biotechnology*, *41*(10), Article 10. https://doi.org/10.1038/s41587-023-01662-6
150. Real, T. D., Hebbar, P., Yoo, D., Antonacci, F., Pačar, I., Diekhans, M., Mikol, G. J., Popoola, O. G., Mallory, B. J., Vollger, M. R., Dishuck, P. C., Guitart, X., Rozanski, A. N., Munson, K. M., Hoekzema, K., Ranchalis, J. E., Neph, S. J., Sedeño-Cortes, A. E., Paten, B., … Eichler, E. E. (2025). *Genetic diversity and regulatory features of human-specific NOTCH2NL duplications*. https://doi.org/10.1101/2025.03.14.643395
151. Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., … Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature*, *444*(7118), Article 7118. https://doi.org/10.1038/nature05329
152. Reik, A., Telling, A., Zitnik, G., Cimbora, D., Epner, E., & Groudine, M. (1998). The Locus Control Region Is Necessary for Gene Expression in the Human β-Globin Locus but Not the Maintenance of an Open Chromatin Structure in Erythroid Cells. *Molecular and Cellular Biology*, *18*(10), 5992–6000. https://doi.org/10.1128/mcb.18.10.5992
153. Ruiz-Narváez, E. A. (2014). Redundant enhancers and causal variants in the TCF7L2 gene. *European Journal of Human Genetics*, *22*(11), 1243–1246. https://doi.org/10.1038/ejhg.2014.17
154. Ryczek, N., Łyś, A., & Makałowska, I. (2023). The Functional Meaning of 5'UTR in Protein-Coding Genes. *International Journal of Molecular Sciences*, *24*(3), 2976. https://doi.org/10.3390/ijms24032976
155. Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., & Lander, E. S. (2006). Positive Natural Selection in the Human Lineage. *Science*, *312*(5780), 1614–1620. https://doi.org/10.1126/science.1124309
156. Salter, J. P., Choe, Y., Albrecht, H., Franklin, C., Lim, K.-C., Craik, C. S., & McKerrow, J. H. (2002). Cercarial Elastase Is Encoded by a Functionally Conserved Gene Family across Multiple Species of Schistosomes*. *Journal of Biological Chemistry*, *277*(27), 24618–24624. https://doi.org/10.1074/jbc.M202364200
157. Schrödinger, LLC. (n.d.). *PyMol* (Version 2.0) [Computer software].

158. Sekar, A., Bialas, A. R., de Rivera, H., Davis, A., Hammond, T. R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., Genovese, G., Rose, S. A., Handsaker, R. E., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Daly, M. J., Carroll, M. C., Stevens, B., & McCarroll, S. A. (2016). Schizophrenia risk from complex variation of complement component 4. *Nature*, *530*(7589), 177–183. https://doi.org/10.1038/nature16549

159. Sharma, S., & Schiller, M. R. (2019). The carboxy-terminus, a key regulator of protein function. *Critical Reviews in Biochemistry and Molecular Biology*, *54*(2), 85–102. https://doi.org/10.1080/10409238.2019.1586828

160. She, X., Liu, G., Ventura, M., Zhao, S., Misceo, D., Roberto, R., Cardone, M. F., Rocchi, M., Program, N. C. S., Green, E. D., Archidiacano, N., & Eichler, E. E. (2006). A preliminary comparative analysis of primate segmental duplications shows elevated substitution rates and a great-ape expansion of intrachromosomal duplications. *Genome Research*, *16*(5), 576–583. https://doi.org/10.1101/gr.4949406

161. Shi, J., Levinson, D. F., Duan, J., Sanders, A. R., Zheng, Y., Pe'er, I., Dudbridge, F., Holmans, P. A., Whittemore, A. S., Mowry, B. J., Olincy, A., Amin, F., Cloninger, C. R., Silverman, J. M., Buccola, N. G., Byerley, W. F., Black, D. W., Crowe, R. R., Oksenberg, J. R., … Gejman, P. V. (2009). Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*, *460*(7256), 753–757. https://doi.org/10.1038/nature08192

162. Shumate, A., & Salzberg, S. L. (2021). Liftoff: Accurate mapping of gene annotations. *Bioinformatics (Oxford, England)*, *37*(12), 1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016

163. Singh, U., & Wurtele, E. S. (2021). orfipy: A fast and flexible tool for extracting ORFs. *Bioinformatics*, *37*(18), 3019–3020. https://doi.org/10.1093/bioinformatics/btab090

164. Sinkus, M. L., Graw, S., Freedman, R., Ross, R. G., Lester, H. A., & Leonard, S. (2015). The Human CHRNA7 and CHRFAM7A Genes: A Review of the Genetics, Regulation, and Function. *Neuropharmacology*, *96*(0 0), 274–288. https://doi.org/10.1016/j.neuropharm.2015.02.006

165. Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Molecular Biology and Evolution*, *32*(5), 1342–1353. https://doi.org/10.1093/molbev/msv022

166. Spoor, F., Gunz, P., Neubauer, S., Stelzer, S., Scott, N., Kwekason, A., & Dean, M. C. (2015). Reconstructed Homo habilis type OH 7 suggests deep-rooted species diversity in early Homo. *Nature*, *519*(7541), Article 7541. https://doi.org/10.1038/nature14224

167. Stankiewicz, P., Park, S.-S., Inoue, K., & Lupski, J. R. (2001). The Evolutionary Chromosome Translocation 4;19 in Gorilla gorilla is Associated with Microduplication of the Chromosome Fragment Syntenic to Sequences Surrounding the Human Proximal CMT1A-REP. *Genome Research*, *11*(7), 1205–1210. https://doi.org/10.1101/gr.181101

168. Stedman, H. H., Kozyak, B. W., Nelson, A., Thesier, D. M., Su, L. T., Low, D. W., Bridges, C. R., Shrager, J. B., Minugh-Purvis, N., & Mitchell, M. A. (2004). Myosin gene mutation correlates with anatomical changes in the human lineage. *Nature*, *428*(6981), 415–418. https://doi.org/10.1038/nature02358

169. Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P. H., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J. E., Hansen, T., Jakobsen, K. D., Muglia, P., Francks, C., Matthews, P. M., Gylfason, A., Halldorsson, B. V., Gudbjartsson, D., Thorgeirsson, T. E., … Stefansson, K. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature*, *455*(7210), 232–236. https://doi.org/10.1038/nature07229

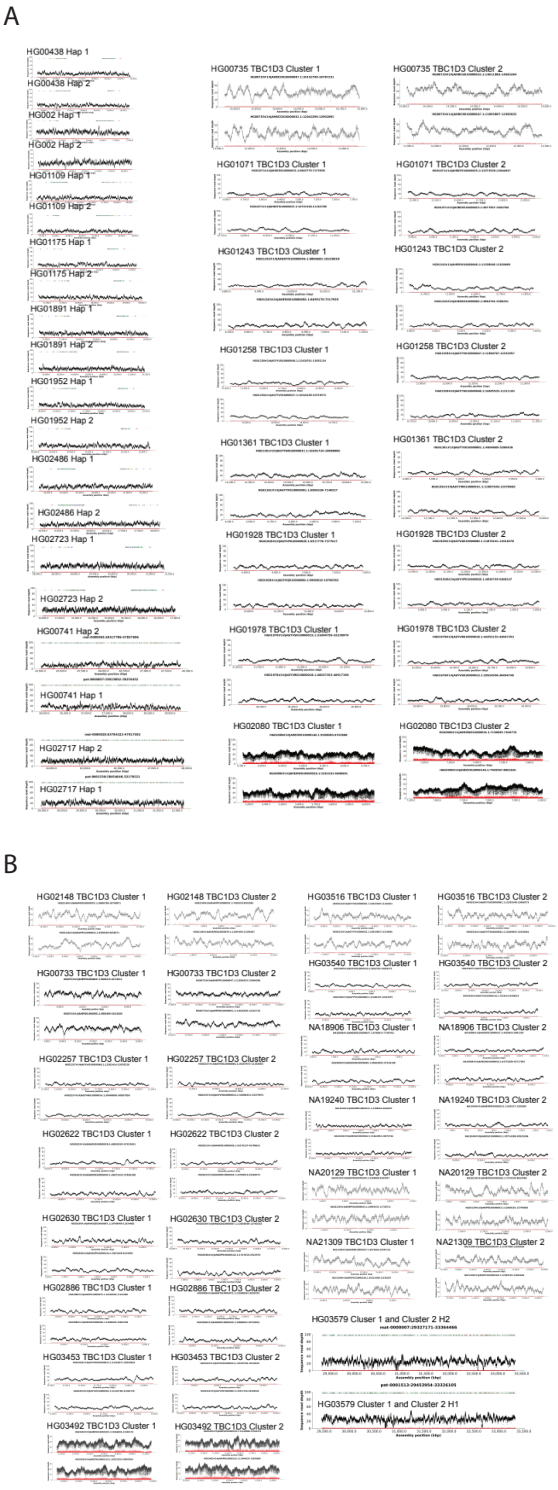
170. Stephens, R. M., & Schneider, T. D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *Journal of Molecular Biology*, *228*(4), 1124–1136. https://doi.org/10.1016/0022-2836(92)90320-j

171. Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S., & Stamatoyannopoulos, J. A. (2020). Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science (New York, N.Y.)*, *368*(6498), 1449–1454. https://doi.org/10.1126/science.aaz1646
172. Stevens, N. J., Seiffert, E. R., O'Connor, P. M., Roberts, E. M., Schmitz, M. D., Krause, C., Gorscak, E., Ngasala, S., Hieronymus, T. L., & Temu, J. (2013). Palaeontological evidence for an Oligocene divergence between Old World monkeys and apes. *Nature*, *497*(7451), Article 7451. https://doi.org/10.1038/nature12161
173. Stone, A. C., Battistuzzi, F. U., Kubatko, L. S., Perry, G. H., Trudeau, E., Lin, H., & Kumar, S. (2010). More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1556), 3277–3288. https://doi.org/10.1098/rstb.2010.0096
174. Stringer, C. (2016). The origin and evolution of Homo sapiens. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*, 20150237. https://doi.org/10.1098/rstb.2015.0237
175. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779
176. Sudmant, P. H., Huddleston, J., Catacchio, C. R., Malig, M., Hillier, L. W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R. E., Persengiev, S., Antonacci, F., Ventura, M., Prado-Martinez, J., Marques-Bonet, T., & Eichler, E. E. (2013). Evolution and diversity of copy number variation in the great ape lineage. *Genome Research*, *23*(9), 1373–1382. https://doi.org/10.1101/gr.158543.113
177. Sudmant, P. H., Kitzman, J. O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., 1000 Genomes Project, & Eichler, E. E. (2010). Diversity of Human Copy Number Variation and Multicopy Genes. *Science*, *330*(6004), 641–646. https://doi.org/10.1126/science.1197005
178. Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M. H.-Y., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Casale, F. P., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394
179. Sulak, M., Fong, L., Mika, K., Chigurupati, S., Yon, L., Mongan, N. P., Emes, R. D., & Lynch, V. J. (2026). TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *eLife*, *5*, e11994. https://doi.org/10.7554/eLife.11994
180. Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(Web Server issue), W609-612. https://doi.org/10.1093/nar/gkl315
181. Tajima, F. (1989). Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics*, *123*(3), 585–595.
182. Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E., & Gersbach, C. A. (2015). Highly Specific Epigenome Editing by CRISPR/Cas9 Repressors for Silencing of Distal Regulatory Elements. *Nature Methods*, *12*(12), 1143–1149. https://doi.org/10.1038/nmeth.3630
183. *The "All of Us" Research Program | New England Journal of Medicine*. (2025, December 31). https://www.nejm.org/doi/full/10.1056/NEJMsr1809937
184. The GTEx Consortium atlas of genetic regulatory effects across human tissues. (2020). *Science (New York, N.Y.)*, *369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776
185. Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., & Myers, R. M. (2004). An Abundance of Bidirectional Promoters in the Human Genome. *Genome Research*, *14*(1), 62–66. https://doi.org/10.1101/gr.1982804

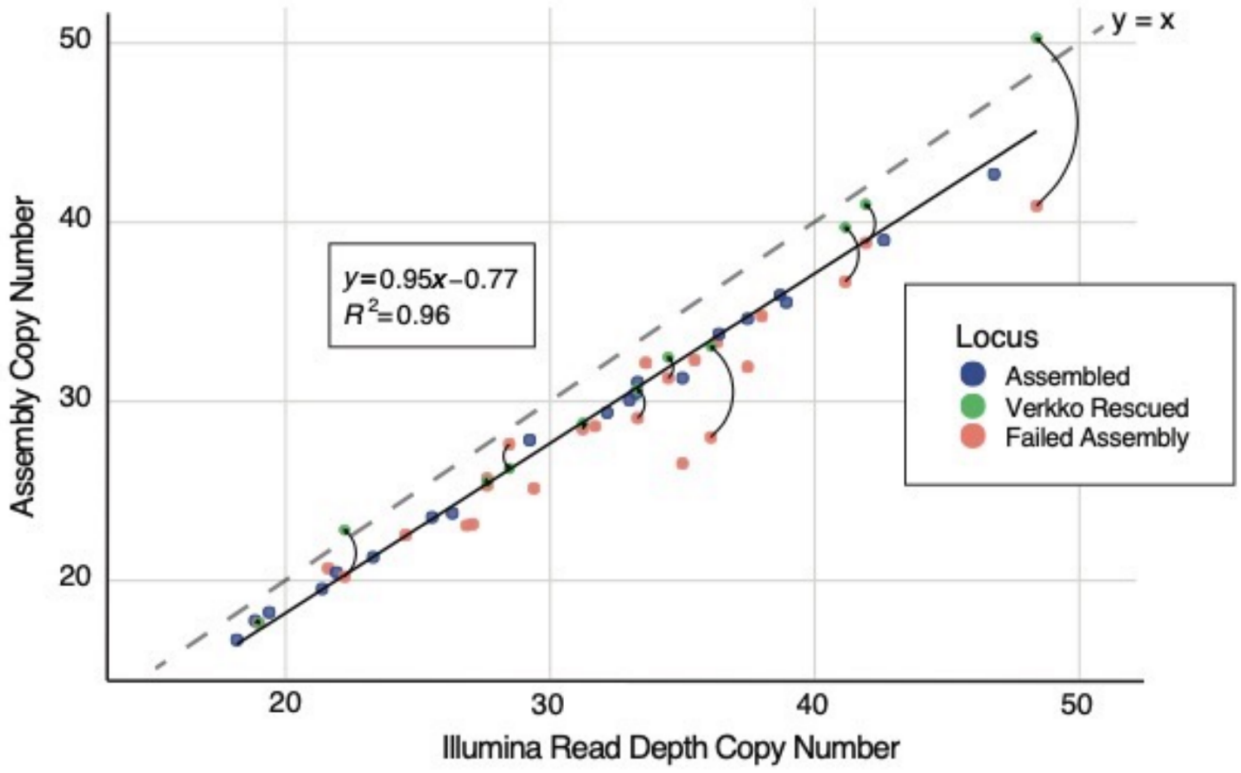
186. Van Bibber, N. W., Haerle, C., Khalife, R., Dayhoff, G. W. I., & Uversky, V. N. (2020). Intrinsic Disorder in Human Proteins Encoded by Core Duplicon Gene Families. *The Journal of Physical Chemistry B*, *124*(37), 8050–8070. https://doi.org/10.1021/acs.jpcb.0c07676
187. Van Essen, D. C., Donahue, C. J., & Glasser, M. F. (2018). Development and Evolution of Cerebral and Cerebellar Cortex. *Brain, Behavior and Evolution*, *91*(3), 158–169. https://doi.org/10.1159/000489943
188. Vattem, K. M., & Wek, R. C. (2004). Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(31), 11269–11274. https://doi.org/10.1073/pnas.0400541101
189. Vollger, M. R., Dishuck, P. C., Harvey, W. T., DeWitt, W. S., Guitart, X., Goldberg, M. E., Rozanski, A. N., Lucas, J., Asri, M., Munson, K. M., Lewis, A. P., Hoekzema, K., Logsdon, G. A., Porubsky, D., Paten, B., Harris, K., Hsieh, P., & Eichler, E. E. (2023). Increased mutation and gene conversion within human segmental duplications. *Nature*, *617*(7960), Article 7960. https://doi.org/10.1038/s41586-023-05895-y
190. Vollger, M. R., Dishuck, P. C., Sorensen, M., Welch, A. E., Dang, V., Dougherty, M. L., Graves-Lindsay, T. A., Wilson, R. K., Chaisson, M. J. P., & Eichler, E. E. (2019). Long-read sequence and assembly of segmental duplications. *Nature Methods*, *16*(1), Article 1. https://doi.org/10.1038/s41592-018-0236-3
191. Vollger, M. R., Guitart, X., Dishuck, P. C., Mercuri, L., Harvey, W. T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K. M., Lewis, A. P., Hoekzema, K., Porubsky, D., Li, R., Nurk, S., Koren, S., Miga, K. H., Phillippy, A. M., Timp, W., Ventura, M., & Eichler, E. E. (2022). Segmental duplications and their variation in a complete human genome. *Science (New York, N.Y.)*, *376*(6588), eabj6965. https://doi.org/10.1126/science.abj6965
192. Vollger, M. R., Korlach, J., Eldred, K. C., Swanson, E., Underwood, J. G., Bohaczuk, S. C., Mao, Y., Cheng, Y.-H. H., Ranchalis, J., Blue, E. E., Schwarze, U., Munson, K. M., Saunders, C. T., Wenger, A. M., Allworth, A., Chanprasert, S., Duerden, B. L., Glass, I., Horike-Pyne, M., … Stergachis, A. B. (2025). Synchronized long-read genome, methylome, epigenome and transcriptome profiling resolve a Mendelian condition. *Nature Genetics*, *57*(2), 469–479. https://doi.org/10.1038/s41588-024-02067-0
193. Vollger, M. R., Neph, S., & Bohaczuk, S. (2025). *fiberseq/FIRE: V0.1.3* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.15677355
194. Vollger, M. R., Swanson, E. G., Neph, S. J., Ranchalis, J., Munson, K. M., Ho, C.-H., Cheng, Y. H. H., Sedeño-Cortés, A. E., Fondrie, W. E., Bohaczuk, S. C., Dippel, M. A., Mao, Y., Parmalee, N. L., Mallory, B. J., Harvey, W. T., Kwon, Y., Garcia, G. H., Hoekzema, K., Meyer, J. G., … Stergachis, A. B. (2025). A haplotype-resolved view of human gene regulation. *bioRxiv*, 2024.06.14.599122. https://doi.org/10.1101/2024.06.14.599122
195. Wainszelbaum, M. J., Charron, A. J., Kong, C., Kirkpatrick, D. S., Srikanth, P., Barbieri, M. A., Gygi, S. P., & Stahl, P. D. (2008). The Hominoid-specific Oncogene *TBC1D3* Activates Ras and Modulates Epidermal Growth Factor Receptor Signaling and Trafficking*. *Journal of Biological Chemistry*, *283*(19), 13233–13242. https://doi.org/10.1074/jbc.M800234200
196. Wainszelbaum, M. J., Liu, J., Kong, C., Srikanth, P., Samovski, D., Su, X., & Stahl, P. D. (2012). TBC1D3, a Hominoid-Specific Gene, Delays IRS-1 Degradation and Promotes Insulin Signaling by Modulating p70 S6 Kinase Activity. *PLOS ONE*, *7*(2), e31225. https://doi.org/10.1371/journal.pone.0031225
197. Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M. A. R., Green, T., Platt, O. S., Ruderfer, D. M., Walsh, C. A., Altshuler, D., Chakravarti, A., Tanzi, R. E., Stefansson, K., Santangelo, S. L., Gusella, J. F., … Daly, M. J. (2008). Association between Microdeletion and Microduplication at 16p11.2 and Autism. *New England Journal of Medicine*, *358*(7), 667–675. https://doi.org/10.1056/NEJMoa075974
198. *Welcome to MassIVE*. (2025, December 31). https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp?redirect=auth

199. Wethmar, K., Smink, J. J., & Leutz, A. (2010). Upstream open reading frames: Molecular switches in (patho)physiology. *Bioessays*, *32*(10), 885–893. https://doi.org/10.1002/bies.201000037
200. Xu, J., Zheng, M., Feng, Z., & Lin, Q. (2024). CCL4L2 participates in tendinopathy progression by promoting macrophage inflammatory responses: A single-cell analysis. *Journal of Orthopaedic Surgery and Research*, *19*(1), 836. https://doi.org/10.1186/s13018-024-05268-9
201. Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. https://doi.org/10.1093/molbev/msm088
202. Yang, Z., Wong, W. S. W., & Nielsen, R. (2005). Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Molecular Biology and Evolution*, *22*(4), 1107–1118. https://doi.org/10.1093/molbev/msi097
203. Yang, Z., Zhang, L., Jiang, X., Yang, X., Ma, K., Yoo, D., Lu, Y., Zhang, S., Chen, J., Nie, Y., Bian, X., Han, J., Fu, L., Zhang, J., Ventura, M., Zhang, G., Sun, Q., Eichler, E. E., & Mao, Y. (2025). Incomplete lineage sorting of segmental duplications defines the human chromosome 2 fusion site early during African great ape speciation. *Cell Genomics*, *0*(0). https://doi.org/10.1016/j.xgen.2025.101079
204. Yoo, D., Rhie, A., Hebbar, P., Antonacci, F., Logsdon, G. A., Solar, S. J., Antipov, D., Pickett, B. D., Safonova, Y., Montinaro, F., Luo, Y., Malukiewicz, J., Storer, J. M., Lin, J., Sequeira, A. N., Mangan, R. J., Hickey, G., Monfort Anez, G., Balachandran, P., … Eichler, E. E. (2025). Complete sequencing of ape genomes. *Nature*, *641*(8062), 401–418. https://doi.org/10.1038/s41586-025-08816-3
205. Yu, G. (2022). *Data Integration, Manipulation and Visualization of Phylogenetic Trees* (1st ed.). Chapman and Hall/CRC.
206. Zhang, F., Zhang, T., Dong, H., Jiang, J., Yang, G., Seim, I., & Tian, R. (2025). Comparative Genomics Uncovers Molecular Adaptations for Cetacean Deep-Sea Diving. *Molecular Ecology*, *34*(22), e17678. https://doi.org/10.1111/mec.17678
207. Zhang, Y., Zeng, X., Liu, P., Hong, R., Lu, H., Ji, H., Lu, L., & Li, Y. (2016). Association between FGFR2 (rs2981582, rs2420946 and rs2981578) polymorphism and breast cancer susceptibility: A meta-analysis. *Oncotarget*, *8*(2), 3454–3470. https://doi.org/10.18632/oncotarget.13839
208. Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A Greedy Algorithm for Aligning DNA Sequences. *Journal of Computational Biology*, *7*(1–2), 203–214. https://doi.org/10.1089/10665270050081478
209. Zody, M. C., Garber, M., Adams, D. J., Sharpe, T., Harrow, J., Lupski, J. R., Nicholson, C., Searle, S. M., Wilming, L., Young, S. K., Abouelleil, A., Allen, N. R., Bi, W., Bloom, T., Borowsky, M. L., Bugalter, B. E., Butler, J., Chang, J. L., Chen, C.-K., … Nusbaum, C. (2006). DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. *Nature*, *440*(7087), 1045–1049. https://doi.org/10.1038/nature04689
210. Zody, M. C., Garber, M., Sharpe, T., Young, S. K., Rowen, L., O'Neill, K., Whittaker, C. A., Kamal, M., Chang, J. L., Cuomo, C. A., Dewar, K., FitzGerald, M. G., Kodira, C. D., Madan, A., Qin, S., Yang, X., Abbasi, N., Abouelleil, A., Arachchi, H. M., … Nusbaum, C. (2006). Analysis of the DNA sequence and duplication history of human chromosome 15. *Nature*, *440*(7084), Article 7084. https://doi.org/10.1038/nature04601

APPENDIX A. SUPPLEMENT FOR CHAPTER 2

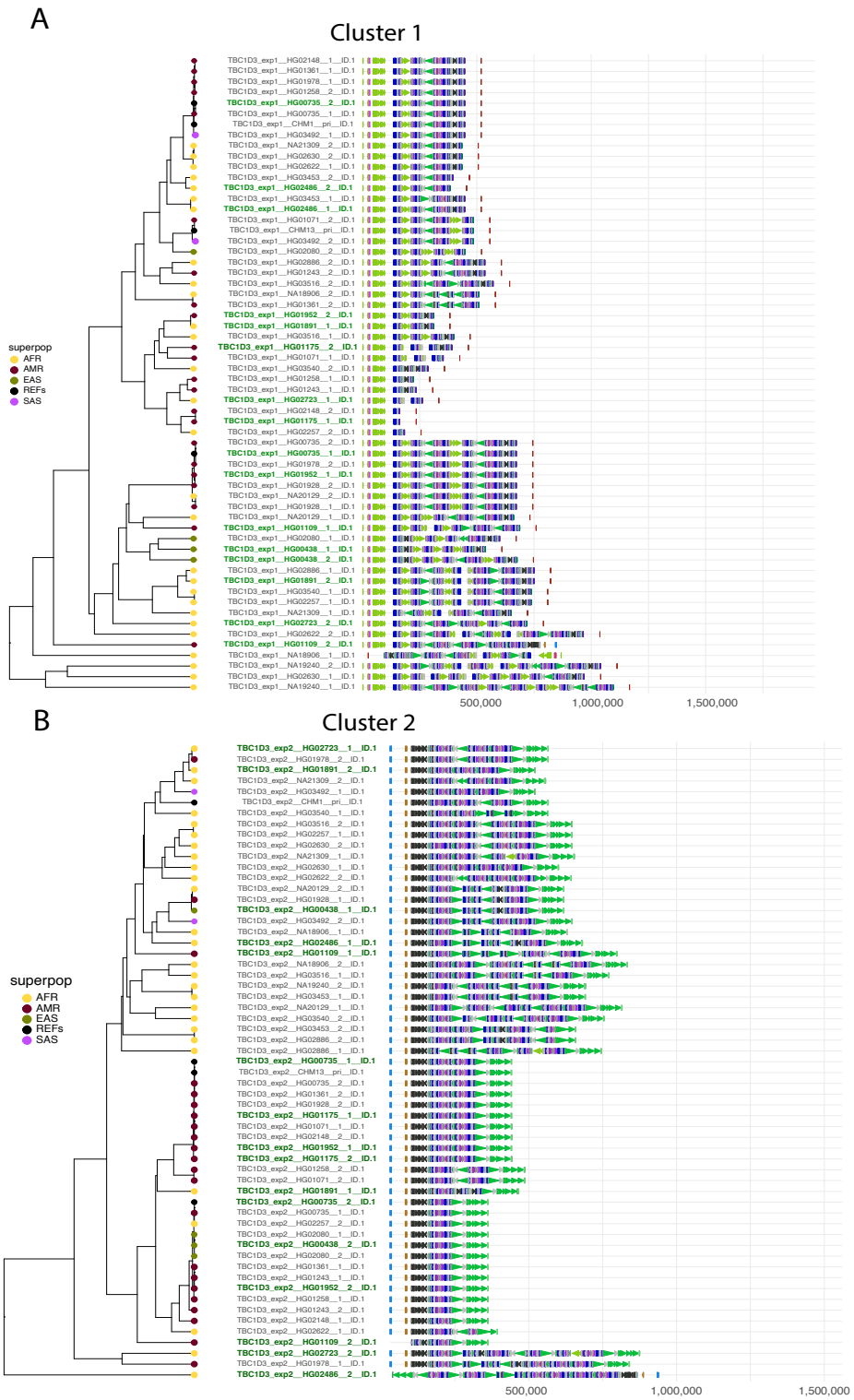


Supplemental Figure S1: NucFreq Validation



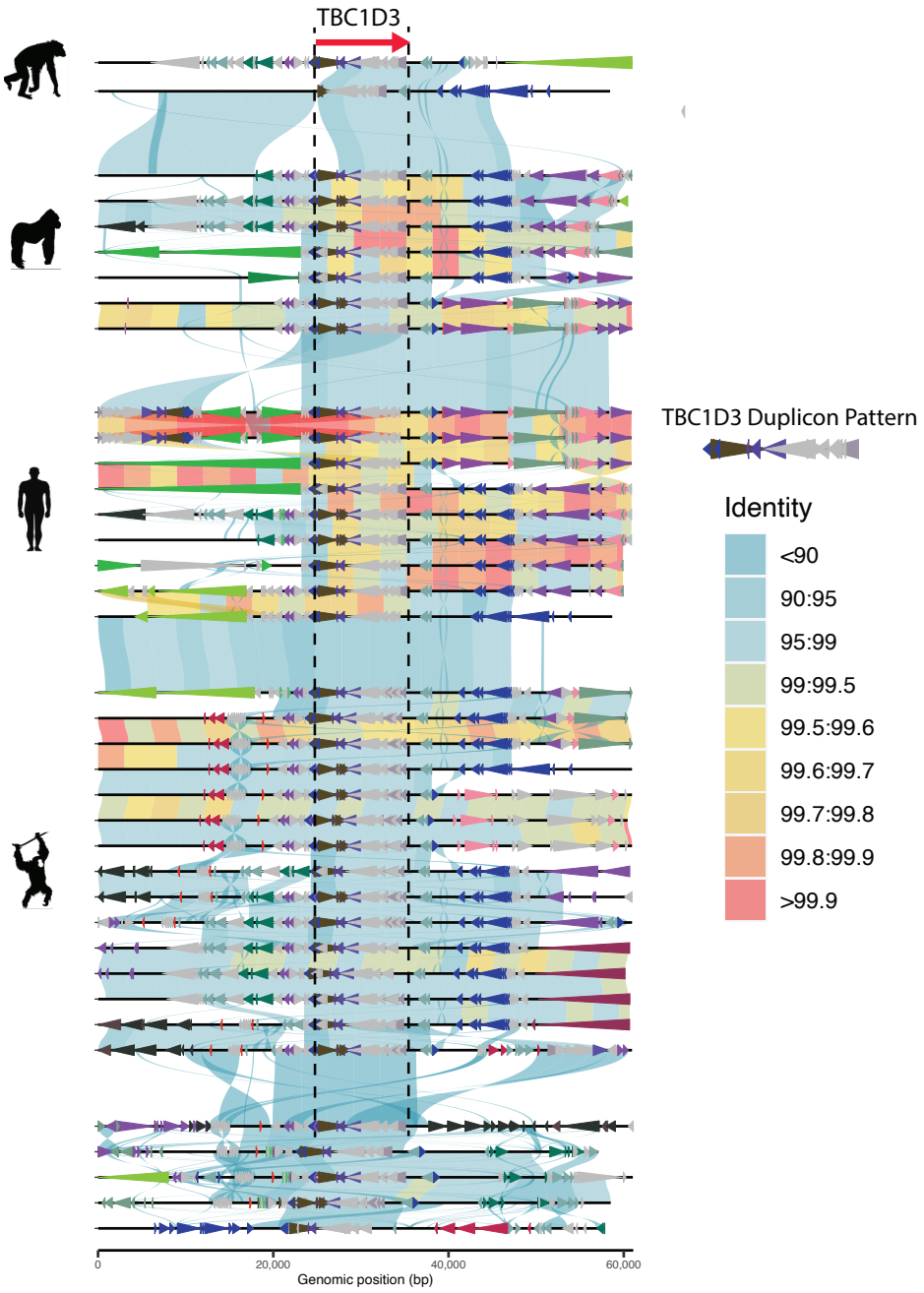
Supplemental Figure S2. Assembly and rescue of TBC1D3 phased haplotypes

Samples were first phased and assembled using exclusively HiFi sequence. The y-axis represents the diploid copy number of these assemblies, and the x-axis represents diploid copy number based on Illumina sequence, with y=x axis marked with a dashed line. Assemblies were validated by read depth estimates of HiFi and ONT, colored in blue if validated, and salmon colored if failing validation. We attempted to rescue sample haplotypes using a novel assembly approach leveraging both HiFi and ultra-long ONT. These samples, assembled by Verkko, are indicated in green.



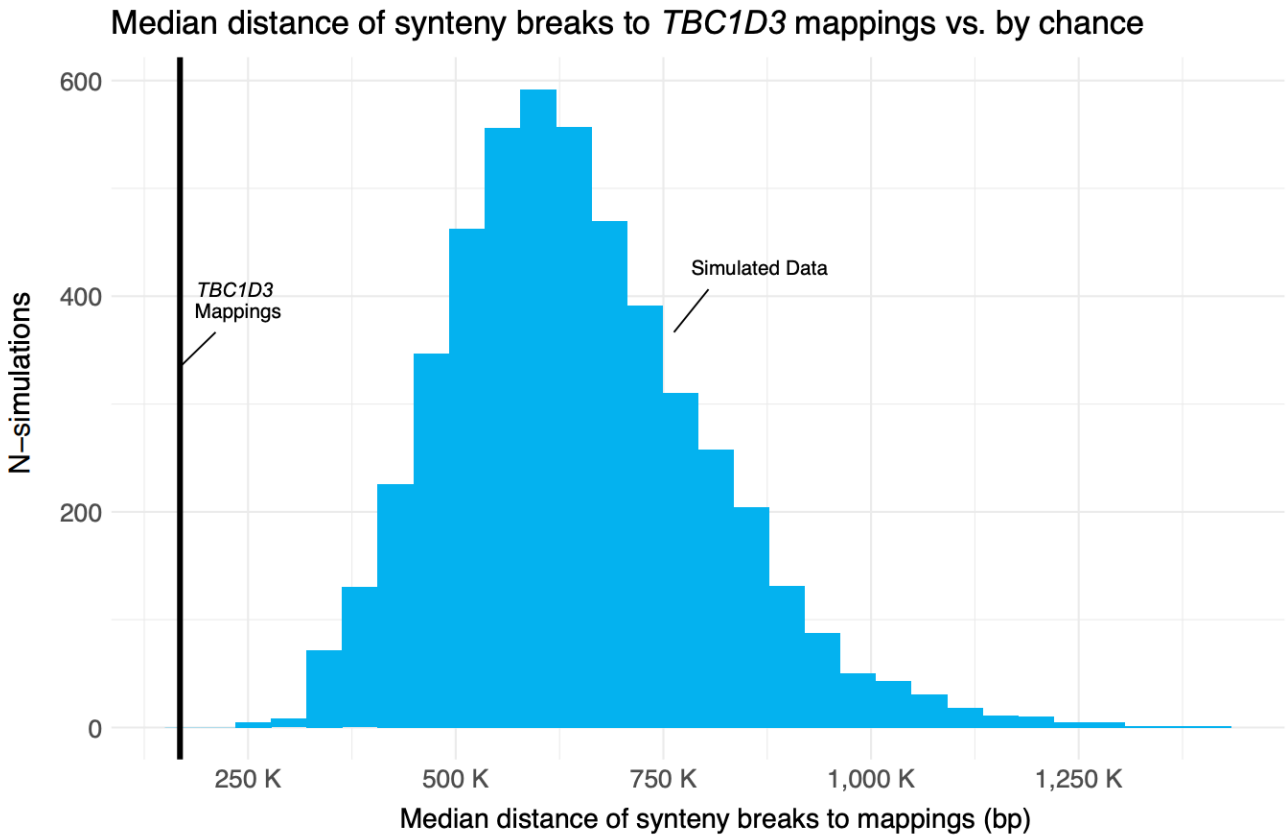
Supplemental Figure S3. UPGMA clustering of TBC1D3

We hierarchically clustered validated assemblies based on duplicon content using UPGMA for cluster 1 **(A)** and 2 **(B)** (Methods). Superpopulations for each haplotype are included in the dendrogram, and assemblies rescued with Verkko are colored in green.



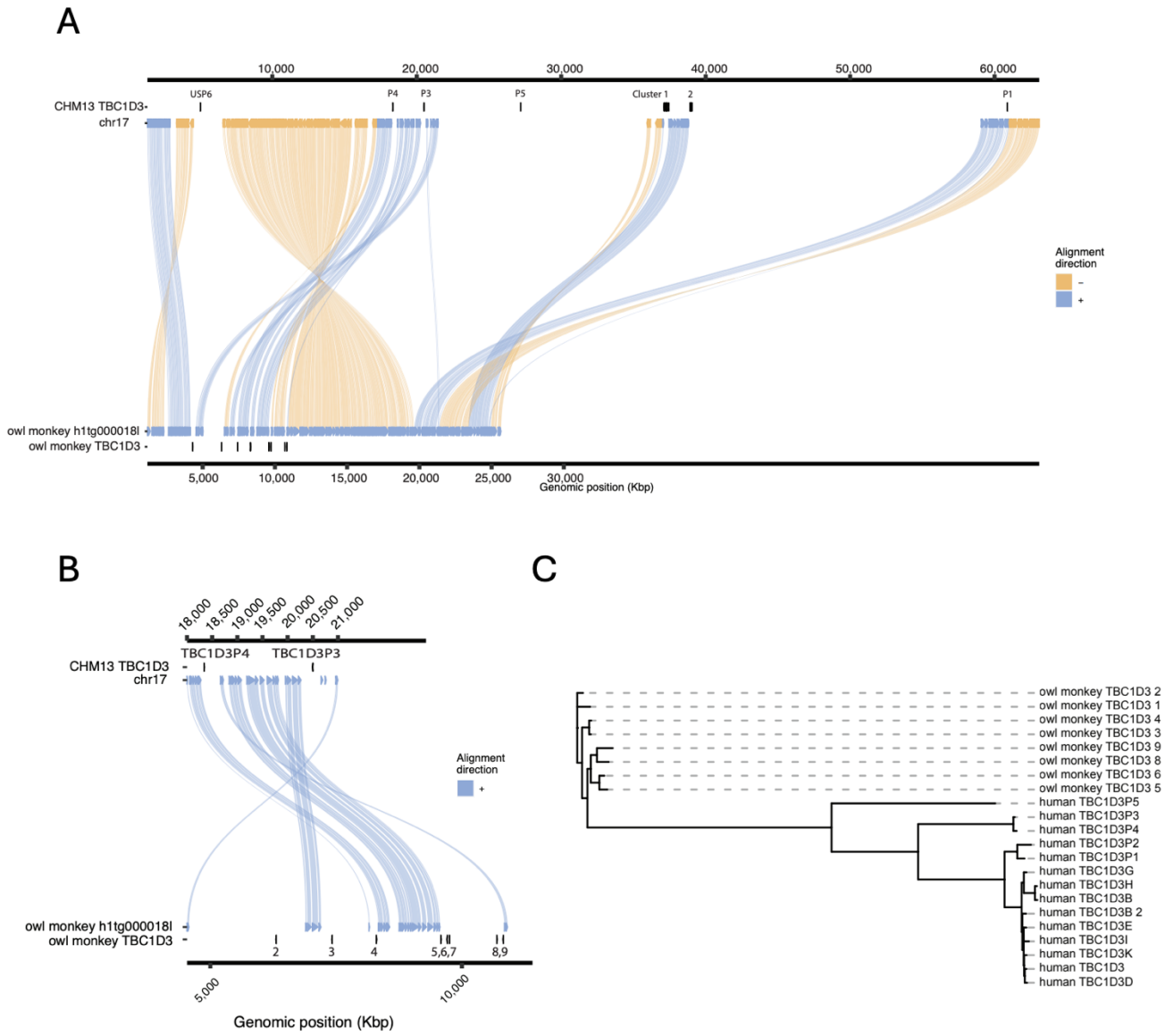
Supplemental Figure S4. Local TBC1D3 duplicon structure supports independent expansion

TBC1D3 copies from clusters 1 and 2, along with 25 kbp flanking sequence, were extracted and mapped to one another. We observe that these paralogs consistently map best, with highest sequence identity and contiguity, to paralogs from their same species of origin, consistent with the hypothesis of independent expansion. This contiguity is similarly observed in underlying duplicon content, shown with colored arrows and annotated with DupMasker.



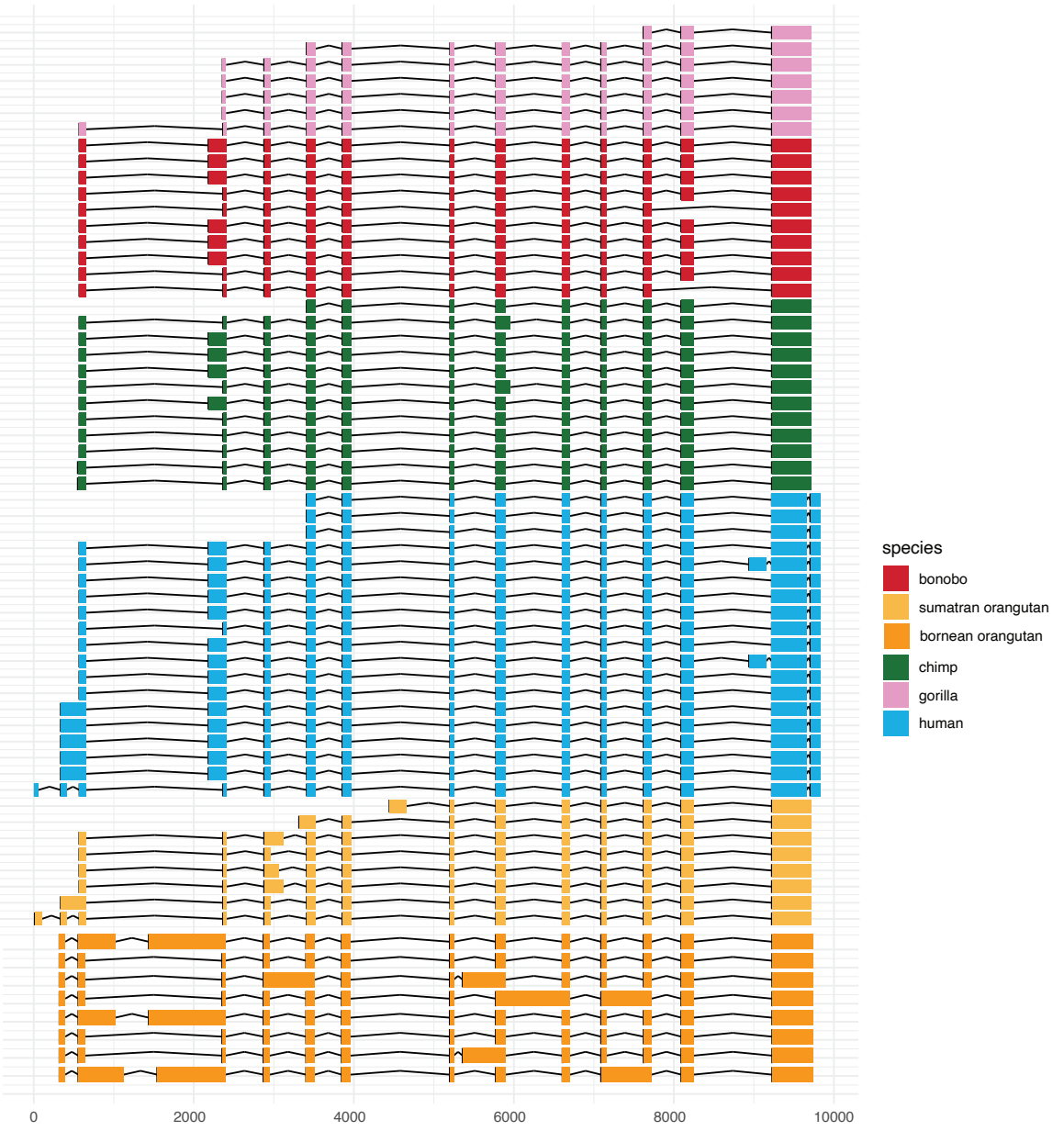
Supplemental Figure S5. TBC1D3 vs. random genomic sequence permutation

Sequences of 11 kbp were randomly selected from orthologous primate Chromosome 17 contigs at the same quantity as the observed TBC1D3 copies contained within the chromosome. We calculated the median distance of this sampling and repeated this experiment in 5000 permutations, comparing median distance relative to true TBC1D3 mappings, marked in black.



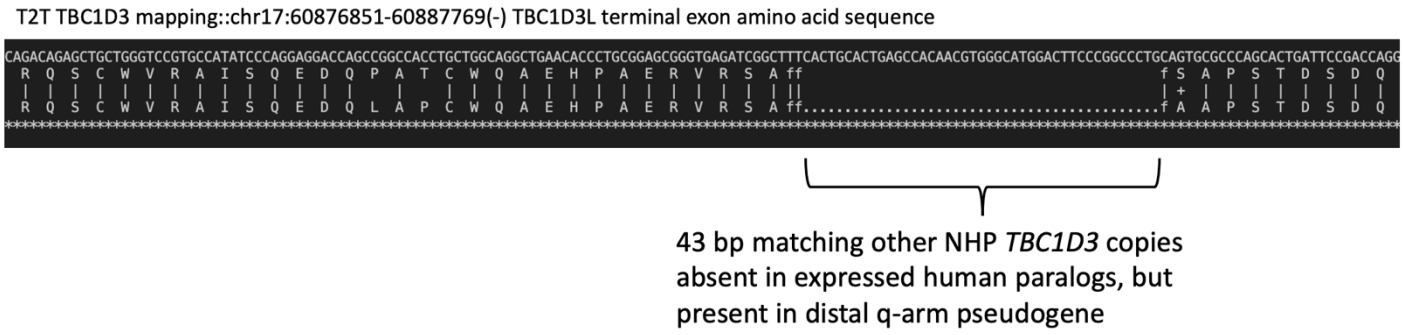
Supplemental Figure S6. Search for ancestral simian TBC1D3 in human and owl monkey assemblies

(A) Human Chr. 17 was aligned to the single owl monkey contig marked with TBC1D3 copies to identify synteny at TBC1D3 locations. **(B)** Zoomed-in synteny between human TBC1D3P4 and TBC1D3P3 with several owl monkey paralogs suggests human TBC1D3P3 or P4 may represent the ancestral copy. **(C)** Maximum likelihood phylogeny of all human vs. owl monkey TBC1D3 paralogs contradicts (B) suggesting TBC1D3P5 may in fact represent the ancestral copy in humans.



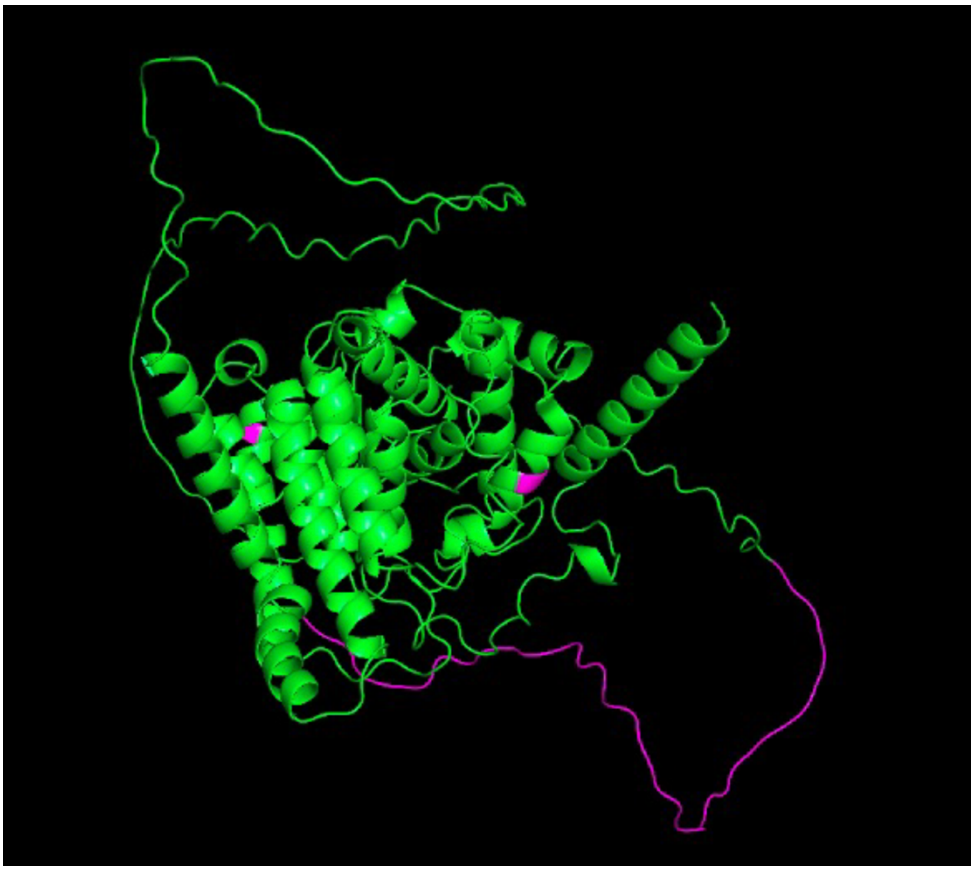
Supplemental Figure S7. Expressed TBC1D3 paralog isoforms

Paralog-specific isoforms were selected for each primate based on their length, mapping quality, and expression support. We observe expression of TBC1D3 from all ape lineages examined; however, for ORF analysis Bornean orangutan and gorilla isoforms were removed due to lack of expression support or intron retention.



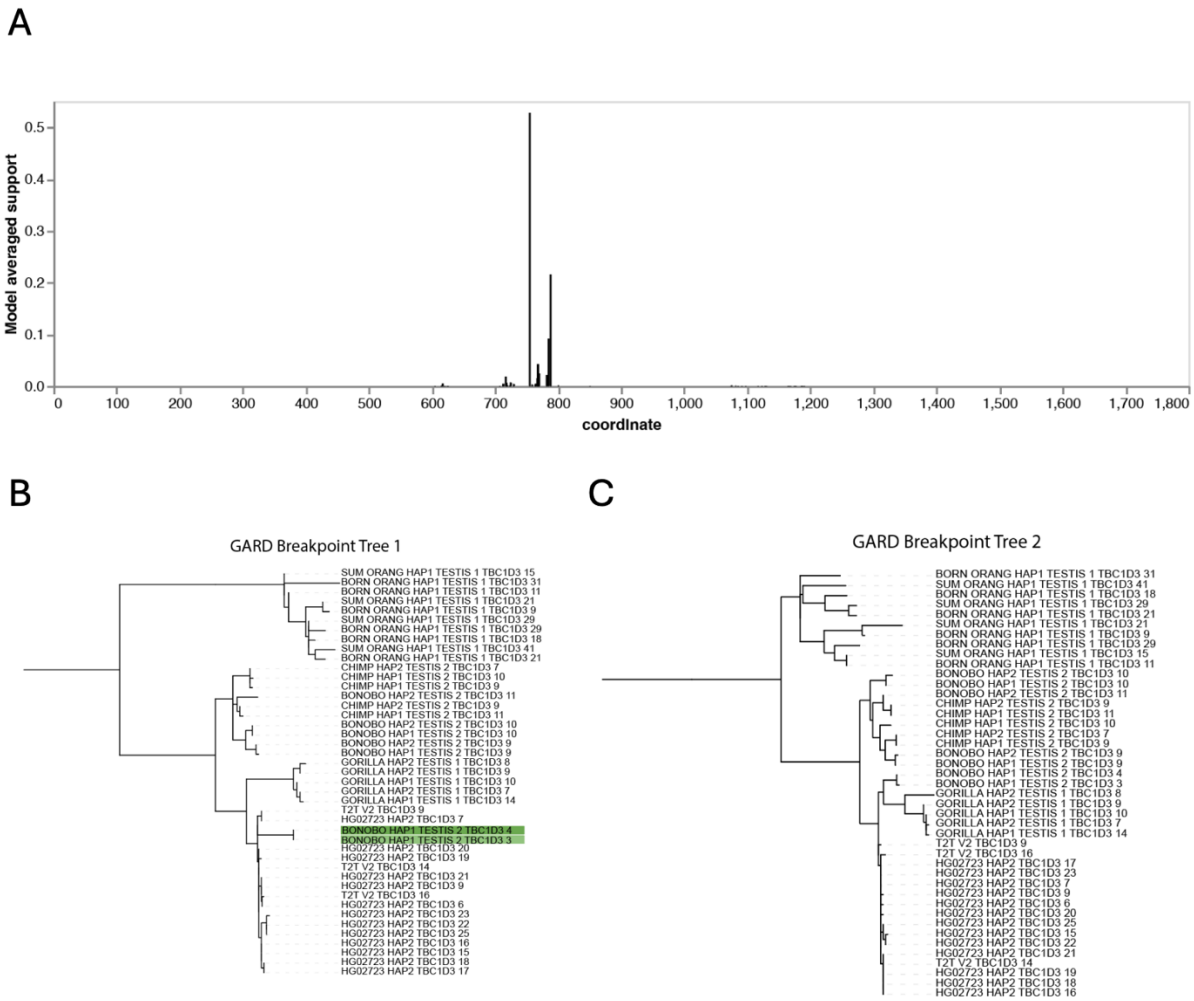
Supplemental Figure S8. Missing 43 bp deletion in human distal TBC1D3 pseudogenes

Human-expressed TBC1D3, with ORF modified by a 43 bp deletion, was mapped to all primate TBC1D3 copies, including all human TBC1D3 copies, with ProSplign (Kiriyutin et al. 2017). ProSplign predicts the genomic nucleotides that represent codons of a given amino acid. We observe that distal TBC1D3 pseudogenes lack the 43 bp deletion universally present in cluster 1 and 2 human TBC1D3 copies.



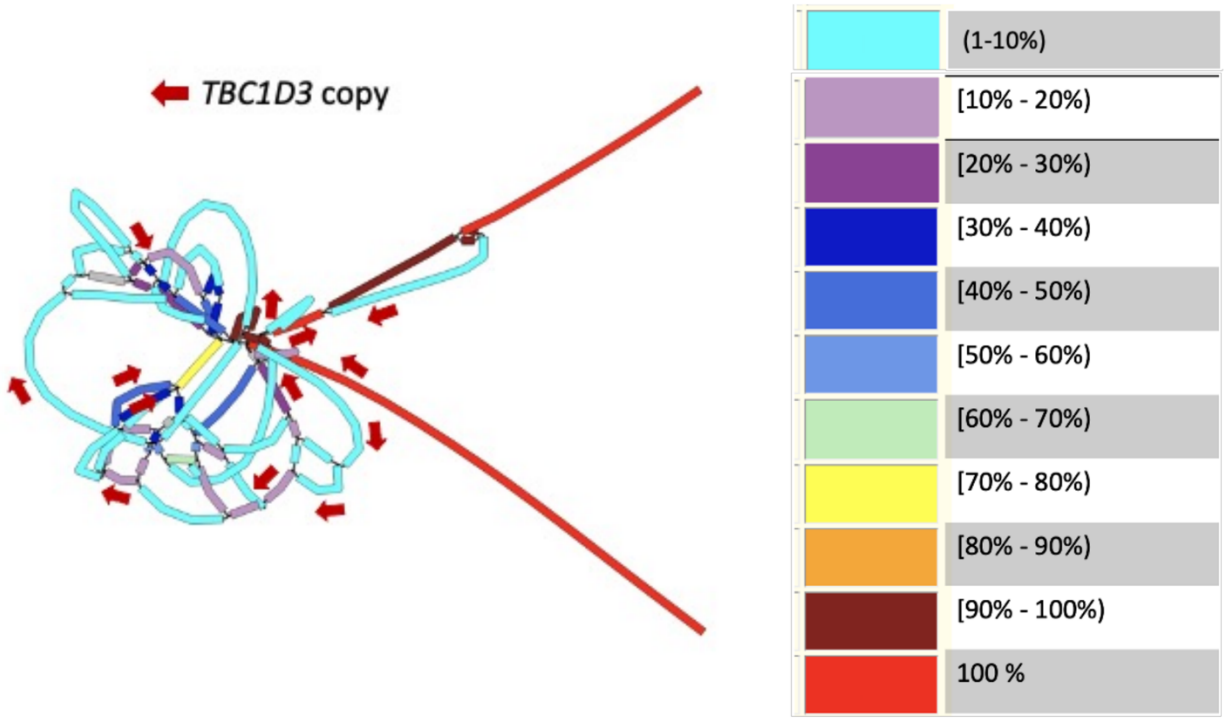
Supplemental Figure S9. Human TBC1D3 predicted tertiary structure

Human TBC1D3 was predicted with Alphafold2 (https://alphafold.ebi.ac.uk/). Human lineage amino acid changes, including the modified carboxy terminus, are indicated with violet. We observe that the 41 aa novel C-terminus tertiary structure could not be predicted and is disordered.



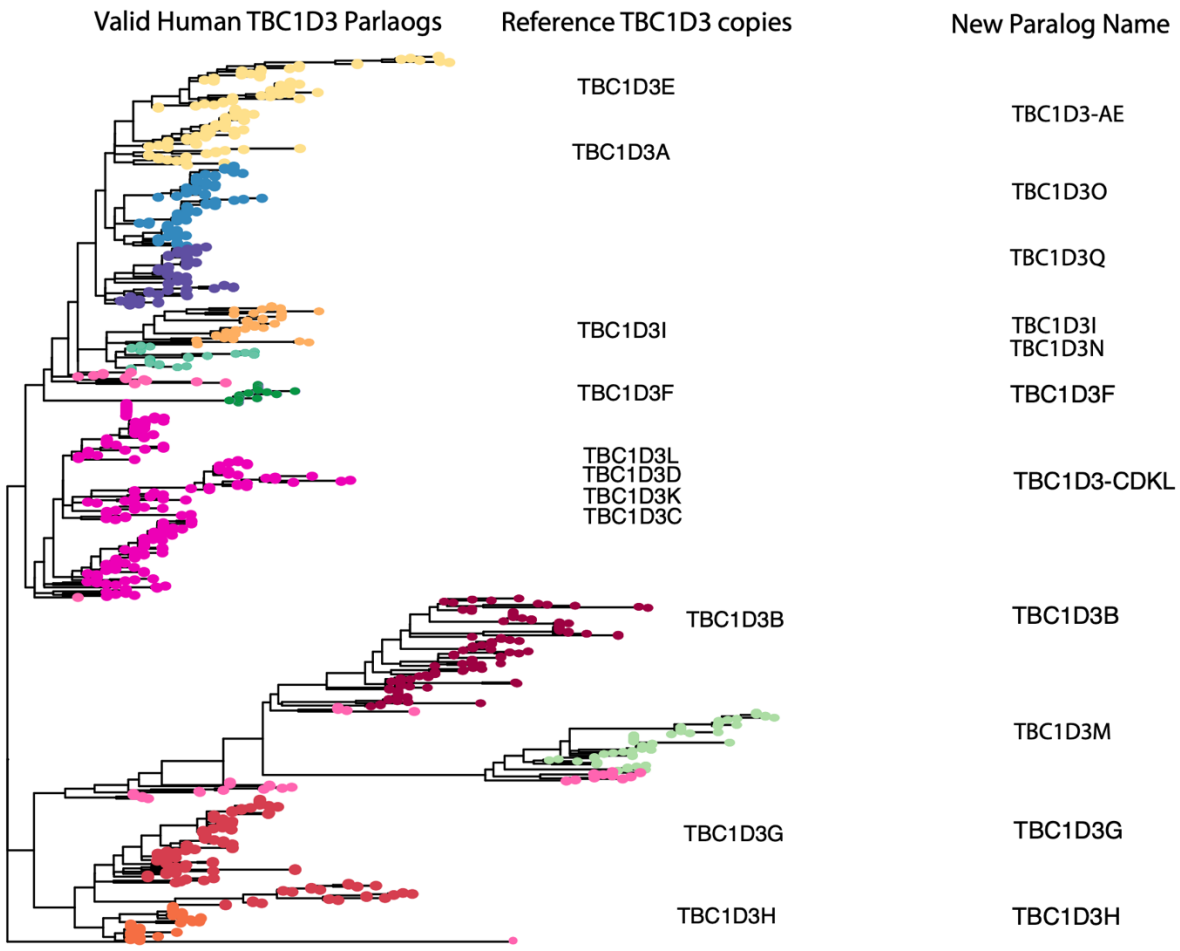
Supplemental Figure S10. Recombination within TBC1D3 gene model

GARD was run on the codon-aligned MSA of expressed TBC1D3 paralogs to investigate if sites under selection or lineage-specific expansions may be explained by gene conversion. **(A)** GARD identified evidence for one breakpoint TBC1D3 bp 755, corresponding to AA 252, and segmenting the multiple sequence alignment in two subsequences. **(B)** Inferred phylogeny of subsequence 1, the first 755 bp of TBC1D3. Green colored lineages illustrate bonobo paralogs nested among human TBC1D3 copies. **(C)** Inferred phylogeny for the last 1045 bp of TBC1D3. For both (B) and (C), species separation is largely maintained, and the inferred breakpoint is not proximal to any sites identified to be undergoing positive selection.



Supplemental Figure S11. Minigraph pangenome graph

We generated a pangenome graph for TBC1D3 using validated human haplotypes with minigraph (settings -S -xggs -L 250 -r 100000). Graph segments are colored to represent the proportion of haplotypes that span the given segment, with light red indicating 100% representation and light blue indicating a single-haplotype traversal. TBC1D3 paralogs are marked with arrows along the graph. We observe that TBC1D3 structural variation is poorly reduced by minigraph, where most copies reduce to nodes with single-haplotype support.



Supplemental Figure S12. Pangenomic clustering and naming

Maximum likelihood phylogeny with ~9600 bp of all human TBC1D3 cluster 1 and 2 copies, outgrouped to chimpanzee TBC1D3. The paralog groups are defined by a heuristic intra-group allelic cutoff based on expected allelic variation in SD sequence (Methods). The first column of labels shows reference GRCh28 paralog locations within the phylogeny. The final column shows the new name given to the associated clusters. Most inherited the name assigned in GRCh28, or a concatenation when multiple paralogs mapped to a common cluster (TBC1D3-AE; TBC1D3-CDKL). Four novel population paralogs not included in GRCh38 were identified (TBC1D3- M,TBC1D3-N, TBC1D3-O, TBC1D3-Q).

Supplemental Table S1

Sample	hap	Super Pop	Assembly method	Total Bp	N Seqs	mean	50%	min	max	N50	auN
GRCh38	1	NA	NA	3.09Gbp	24	128.68Mbp	133.54 Mbp	46.71 Mbp	248.96Mbp	156.04 Mbp	153.57 Mbp
CHM1	1	EUR	NA	3.03Gbp	639	4.75 Mbp	43.46 Kbp	8.01Kbp	118.38Mbp	59.89 Mbp	65.52 Mbp
CHM13	1	EUR	NA	3.11Gbp	25	124.48Mbp	133.32 Mbp	16.57 Kbp	248.39Mbp	150.62 Mbp	154.63Mbp
HG00438	1	EAS	Verkko 1.1	3.01Gbp	352	8.57 Mbp	49.09 Kbp	11.90 Kbp	193.25Mbp	103.64 Mbp	108.21Mbp
HG00438	2	EAS	Verkko 1.1	3.02Gbp	236	12.79 Mbp	53.00 Kbp	6.67Kbp	191.28Mbp	103.03 Mbp	109.90Mbp
HG00733	1	AMR	Verkko 1.1	3.00Gbp	400	7.50 Mbp	36.65 Kbp	6.50Kbp	203.20Mbp	134.85 Mbp	123.49Mbp
HG00733	2	AMR	Verkko 1.1	3.00Gbp	232	12.92 Mbp	46.75 Kbp	7.80Kbp	201.24Mbp	136.67 Mbp	125.16Mbp
HG00735	1	AMR	Trio-Hifiasm (v.0.14)	3.03Gbp	299	10.15 Mbp	312.08 Kbp	19.02 Kbp	159.64Mbp	53.42 Mbp	67.21 Mbp
HG00735	2	AMR	Trio-Hifiasm (v.0.14)	3.04Gbp	244	12.45 Mbp	911.95 Kbp	16.56 Kbp	123.74Mbp	56.47 Mbp	59.62 Mbp
HG01071	1	AMR	Trio-Hifiasm (v.0.14)	3.06Gbp	330	9.26 Mbp	550.86 Kbp	22.63 Kbp	148.18Mbp	55.59 Mbp	57.76 Mbp
HG01071	2	AMR	Trio-Hifiasm (v.0.14)	3.01Gbp	236	12.77 Mbp	1.08 Mbp	27.10 Kbp	184.36Mbp	50.13 Mbp	64.23 Mbp

HG01109	1	AMR	Verkko 1.1	3.01Gbp	562	5.36 Mbp	34.98 Kbp	5.21Kbp	236.40Mbp	81.66 Mbp	96.38 Mbp
HG01109	2	AMR	Verkko 1.1	2.89Gbp	356	8.13 Mbp	36.64 Kbp	4.00Kbp	164.30Mbp	82.20 Mbp	89.01 Mbp
HG01175	1	AMR	Verkko 1.2	3.02Gbp	278	10.88 Mbp	58.49 Kbp	9.32Kbp	200.53Mbp	87.48 Mbp	89.52 Mbp
HG01175	2	AMR	Verkko 1.2	3.02Gbp	247	12.21 Mbp	64.67 Kbp	7.74Kbp	182.17Mbp	109.50 Mbp	99.33 Mbp
HG01243	1	AMR	Trio-Hifiasm (v.0.14)	2.91Gbp	436	6.67 Mbp	295.20 Kbp	16.75 Kbp	78.53 Mbp	29.12 Mbp	33.06 Mbp
HG01243	2	AMR	Trio-Hifiasm (v.0.14)	3.03Gbp	345	8.78 Mbp	1.98 Mbp	16.57 Kbp	118.53Mbp	31.38 Mbp	37.59 Mbp
HG01258	1	AMR	Trio-Hifiasm (v.0.14)	2.92Gbp	324	9.00 Mbp	445.02 Kbp	16.54 Kbp	116.99Mbp	49.86 Mbp	55.47 Mbp
HG01258	2	AMR	Trio-Hifiasm (v.0.14)	3.03Gbp	346	8.76 Mbp	349.97 Kbp	14.86 Kbp	157.90Mbp	56.64 Mbp	62.84 Mbp
HG01361	1	AMR	Trio-Hifiasm (v.0.14)	3.01Gbp	288	10.45 Mbp	431.32 Kbp	10.29 Kbp	174.83Mbp	47.18 Mbp	57.26 Mbp
HG01361	2	AMR	Trio-Hifiasm (v.0.14)	3.03Gbp	295	10.26 Mbp	791.00 Kbp	16.56 Kbp	132.68Mbp	45.12 Mbp	58.51 Mbp
HG01891	1	AFR	Verkko 1.2	3.03Gbp	395	7.68 Mbp	34.32 Kbp	7.75Kbp	239.95Mbp	123.76 Mbp	117.17Mbp
HG01891	2	AFR	Verkko 1.2	3.01Gbp	327	9.22 Mbp	33.30 Kbp	10.37 Kbp	243.96Mbp	133.11 Mbp	126.55Mbp
HG01928	1	AMR	Trio-Hifiasm (v.0.14)	2.92Gbp	317	9.22 Mbp	424.27 Kbp	20.43 Kbp	154.53Mbp	45.70 Mbp	59.38 Mbp

HG01928	2	AMR	Trio-Hifiasm (v.0.14)	3.03Gbp	244	12.40 Mbp	1.19 Mbp	16.56 Kbp	152.51Mbp	53.72 Mbp	57.55 Mbp
HG01952	1	AMR	Verkko 1.0	2.96Gbp	1011	2.93 Mbp	153.37 Kbp	12.08 Kbp	65.93 Mbp	19.04 Mbp	22.79 Mbp
HG01952	2	AMR	Verkko 1.0	2.82Gbp	1153	2.44 Mbp	127.15 Kbp	9.79Kbp	68.79 Mbp	19.66 Mbp	22.92 Mbp
HG01978	1	AMR	Trio-Hifiasm (v.0.14)	3.06Gbp	286	10.68 Mbp	566.14 Kbp	21.64 Kbp	137.44Mbp	52.81 Mbp	55.62 Mbp
HG01978	2	AMR	Trio-Hifiasm (v.0.14)	3.05Gbp	247	12.36 Mbp	848.41 Kbp	16.57 Kbp	128.13Mbp	60.49 Mbp	59.11 Mbp
HG02080	1	EAS	Trio-Hifiasm (v.0.14)	3.02Gbp	485	6.24 Mbp	687.41 Kbp	17.38 Kbp	75.66 Mbp	24.29 Mbp	28.55 Mbp
HG02080	2	EAS	Trio-Hifiasm (v.0.14)	3.03Gbp	422	7.19 Mbp	1.84 Mbp	16.56 Kbp	83.35 Mbp	20.23 Mbp	27.12 Mbp
HG02148	1	AMR	Trio-Hifiasm (v.0.14)	3.03Gbp	487	6.21 Mbp	156.84 Kbp	14.34 Kbp	111.34Mbp	41.87 Mbp	42.08 Mbp
HG02148	2	AMR	Trio-Hifiasm (v.0.14)	3.04Gbp	423	7.18 Mbp	241.54 Kbp	14.93 Kbp	104.12Mbp	39.94 Mbp	45.84 Mbp
HG02257	1	AFR	Trio-Hifiasm (v.0.14)	3.04Gbp	306	9.94 Mbp	461.33 Kbp	17.03 Kbp	134.40Mbp	57.98 Mbp	62.54 Mbp
HG02257	2	AFR	Trio-Hifiasm (v.0.14)	3.03Gbp	292	10.38 Mbp	452.14Kbp	16.57 Kbp	121.81Mbp	59.04 Mbp	63.70 Mbp
HG02486	1	AFR	Verkko 1.2	3.06Gbp	340	9.01 Mbp	32.89 Kbp	7.84Kbp	235.77Mbp	140.68 Mbp	131.21Mbp
HG02486	2	AFR	Verkko 1.2	2.94Gbp	330	8.91 Mbp	33.82 Kbp	5.82Kbp	227.69Mbp	132.65 Mbp	127.40Mbp

HG02622	1	AFR	Trio-Hifiasm (v.0.14)	3.04Gbp	270	11.27 Mbp	205.57Kbp	20.60 Kbp	103.76Mbp	51.21 Mbp	55.29 Mbp
HG02622	2	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	292	10.43 Mbp	145.55Kbp	16.57 Kbp	139.51Mbp	60.04 Mbp	65.20 Mbp
HG02630	1	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	532	5.74 Mbp	126.84Kbp	10.24 Kbp	76.91 Mbp	29.25 Mbp	31.68 Mbp
HG02630	2	AFR	Trio-Hifiasm (v.0.14)	3.04Gbp	497	6.12 Mbp	215.10Kbp	16.57 Kbp	83.92 Mbp	25.38 Mbp	29.89 Mbp
HG02723	1	AFR	Verkko 1.1	3.05Gbp	348	8.77 Mbp	47.06 Kbp	5.05Kbp	178.60Mbp	94.50 Mbp	97.30 Mbp
HG02723	2	AFR	Verkko 1.1	3.09Gbp	254	12.16 Mbp	55.63 Kbp	6.17Kbp	242.28Mbp	96.32 Mbp	110.64Mbp
HG02886	1	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	503	6.06 Mbp	292.53Kbp	15.28 Kbp	87.30 Mbp	29.11 Mbp	31.41 Mbp
HG02886	2	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	461	6.61 Mbp	230.68Kbp	16.57 Kbp	114.51Mbp	28.88 Mbp	33.91 Mbp
HG03453	1	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	480	6.36 Mbp	392.59Kbp	16.57 Kbp	63.91 Mbp	27.05 Mbp	29.27 Mbp
HG03453	2	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	439	6.94 Mbp	1.05Mbp	16.57 Kbp	90.18 Mbp	26.38 Mbp	27.55 Mbp
HG03492	1	SAS	Trio-Hifiasm (v.0.14)	2.92Gbp	618	4.73 Mbp	413.77Kbp	15.77 Kbp	51.97 Mbp	20.16 Mbp	21.88 Mbp
HG03492	2	SAS	Trio-Hifiasm (v.0.14)	3.02Gbp	495	6.11 Mbp	1.46Mbp	16.57 Kbp	68.99 Mbp	18.86 Mbp	24.60 Mbp
HG03516	1	AFR	Trio-Hifiasm (v.0.14)	3.07Gbp	369	8.31 Mbp	178.76Kbp	15.71 Kbp	158.66Mbp	55.48 Mbp	62.20 Mbp
HG03516	2	AFR	Trio-Hifiasm (v.0.14)	3.03Gbp	320	9.48 Mbp	621.01Kbp	16.10 Kbp	134.21Mbp	44.77 Mbp	50.42 Mbp

HG03540	1	AFR	Trio-Hifiasm (v.0.14)	3.07Gbp	512	5.99 Mbp	130.70Kbp	15.27 Kbp	88.87 Mbp	34.16 Mbp	34.16 Mbp
HG03540	2	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	435	7.01 Mbp	385.66Kbp	16.57 Kbp	96.32 Mbp	30.47 Mbp	33.87 Mbp
NA18906	1	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	304	10.02 Mbp	525.06Kbp	18.30 Kbp	109.21Mbp	43.52 Mbp	45.95 Mbp
NA18906	2	AFR	Trio-Hifiasm (v.0.14)	3.06Gbp	271	11.28 Mbp	1.82Mbp	15.32 Kbp	96.36 Mbp	40.08 Mbp	43.31 Mbp
NA19240	1	AFR	Trio-Hifiasm (v.0.14)	3.04Gbp	453	6.71 Mbp	768.11Kbp	15.87 Kbp	70.97 Mbp	25.20 Mbp	28.93 Mbp
NA19240	2	AFR	Trio-Hifiasm (v.0.14)	3.03Gbp	387	7.83 Mbp	989.05Kbp	16.57 Kbp	71.66 Mbp	28.90 Mbp	31.47 Mbp
NA20129	1	AFR	Trio-Hifiasm (v.0.14)	3.03Gbp	481	6.30 Mbp	1.02Mbp	13.66 Kbp	58.33 Mbp	22.42 Mbp	23.09 Mbp
NA20129	2	AFR	Trio-Hifiasm (v.0.14)	3.05Gbp	448	6.80 Mbp	1.44Mbp	16.37 Kbp	92.88 Mbp	21.15 Mbp	28.32 Mbp
NA21309	1	AFR	Trio-Hifiasm (v.0.14)	3.03Gbp	608	4.98 Mbp	901.50Kbp	13.19 Kbp	53.37 Mbp	17.44 Mbp	20.91 Mbp
NA21309	2	AFR	Trio-Hifiasm (v.0.14)	3.04Gbp	501	6.06 Mbp	1.52Mbp	19.66 Kbp	70.48 Mbp	20.56 Mbp	23.77 Mbp
HG03579	1	AFR	Verkko 1.4	3.04Gbp	689	4.42 Mbp	29.21 Kbp	9.77Kbp	246.12Mbp	132.32 Mbp	124.92Mbp
HG03579	2	AFR	Verkko 1.4	2.94Gbp	454	6.48 Mbp	30.43 Kbp	7.84Kbp	234.80Mbp	102.87 Mbp	109.40Mbp
HG00741	1	AMR	Verkko 1.4	3.05Gbp	239	12.76 Mbp	49.32 Kbp	8.45Kbp	202.81Mbp	133.17 Mbp	123.66Mbp
HG00741	2	AMR	Verkko 1.4	3.03Gbp	210	14.44 Mbp	47.07 Kbp	9.82Kbp	252.54Mbp	142.03 Mbp	138.30Mbp
HG02717	1	AFR	Verkko 1.4	3.06Gbp	381	8.02 Mbp	23.78 Kbp	6.13Kbp	243.23Mbp	116.09 Mbp	120.92Mbp

HG02717	2	AFR	Verkko 1.4	2.95Gbp	336	8.77 Mbp	23.61 Kbp	6.14Kbp	241.66Mbp	109.73 Mbp	123.16Mbp
HG002	1	EUR	NA	2.95Gbp	25	117.92Mbp	109.23Mbp	16.57 Kbp	252.06Mbp	146.79 Mbp	154.64Mbp
HG002	2	EUR	NA	3.05Gbp	25	122.07Mbp	133.58Mbp	16.57 Kbp	244.02Mbp	154.34 Mbp	155.80Mbp

Supplemental Table S2: Human Assembly HiFi and ONT Sequencing Statistics

Sample	ONT Read N50	ONT Gb	ONT coverage	ONT 100kb+	HiFi Read N50	HiFi Gbp	HiFi coverage
CHM1	NA	NA	NA	NA	NA	NA	NA
CHM13	NA	NA	NA	NA	NA	NA	NA
GRCh38chrOnly	NA	NA	NA	NA	NA	NA	NA
HG002	NA	NA	710	66	14446	550.89	180.34
HG00438	73532	136.61	41.4	14.67	22619	127.47	41.73
HG00735	76977	102.66	31.11	11.68	22328	133.5	43.70
HG00741	79828	145.37	44.05	17.29	24843	119.23	39.03
HG01071	76951	132.38	40.12	15.29	26926	108.93	35.66
HG01109	-	-	59	2.5	21163	99.78	32.66
HG01175	75225	148.25	44.93	16.59	21602	111.94	36.64
HG01243	-	-	60	6	19528	109.24	35.76
HG01258	91791	69.81	21.16	9.73	20342	111.15	36.38
HG01361	67140	107.71	32.63	10.69	20026	130	42.55
HG01891	79570	120.77	36.6	14.72	17039	115.13	37.69
HG01928	82260	85.01	25.76	10.49	23609	108.9	35.65
HG01952	72206	103.55	31.38	11.06	21109	127.85	41.85
HG01978	77617	128.37	38.91	14.67	26203	117.57	38.49
HG02080	-	-	56	6	21313	105.9	34.66
HG02148	61223	60.91	18.45	5.63	16942	123.24	40.34
HG02257	76136	85.04	25.77	9.56	19246	110.64	36.22
HG02486	64367	95	28.79	8.62	18626	123.9	40.56
HG02622	87046	47.62	14.43	6.29	22882	143.99	47.14
HG02630	88601	64.53	19.56	8.72	19196	151.01	49.43
HG02717	72917	88.93	26.94	9.75	16751	140.3	45.93
HG02723	63.22	-	131.74	38.64	15693	143.29	46.91
HG02886	81798	106.81	32.37	13.26	17624	137.41	44.98
HG03453	93224	36.08	10.94	5.07	19883	160.25	52.46
HG03492	NA	NA	51	3	21008	104.36	34.16
HG03516	75811	141.52	42.89	16.3	18663	110.25	36.09
HG03540	81809	95.38	28.91	11.78	20527	153.41	50.22
HG03579	87729	51.94	15.73	6.91	19408	149.99	49.10
NA18906	NA	NA	NA	NA	21232	130.85	42.83
NA19240	NA	NA	NA	NA	17682	125.18	40.98
NA20129	NA	NA	NA	NA	17708	116.26	38.06
NA21309	NA	NA	NA	NA	16551	122.41	40.07

Supplemental Table S3: Human Assembly TBC1D3 Statistics

Sample	hap	assembly_copy_num	TBC_exp1_copy_num	TBC exp2 copy_num	TBC Exp 1 size	TBC_exp2_size	Kmer diplotype copy_numm	illumina_WGS_wssd_diplotype_copy_number
GRCh38 chrOnly	1	14	6	6	534298	595537	NA	NA
CHM1	1	14	6	6	534307	595654	NA	NA
CHM13	1	11	5	4	572148	471455	NA	NA
HG00438	1	15	6	7	624623	648304	26.3	28.5
HG00438	2	12	8	2	762494	391802	26.3	28.5
HG00733	1	19	9	8	769671	684890	28.46	31.23
HG00733	2	12	6	4	544270	525504	28.46	31.23
HG00735	1	10	6	2	534428	391473	23.5	25.5
HG00735	2	15	9	4	759809	471485	23.5	25.5
HG01071	1	10	4	4	440634	471443	20.5	21.9
HG01071	2	11	5	4	572077	515820	20.5	21.9
HG01109	1	19	8	9	774207	827909	32.5	34.5
HG01109	2	16	10	4	2073405	1589665	32.5	34.5
HG01175	1	7	1	4	250249	471080	17.6	19
HG01175	2	11	5	4	478249	471535	17.6	19
HG01243	1	7	3	2	322430	391516	18.2	19.4
HG01243	2	12	7	2	622679	391523	18.2	19.4
HG01258	1	6	2	2	309731	391920	17.8	18.8
HG01258	2	12	6	4	534400	515481	17.8	18.8
HG01361	1	10	6	2	534218	391529	21.3	23.3
HG01361	2	12	6	4	596303	471101	21.3	23.3
HG01891	1	9	3	4	397758	494196	25.5	27.6
HG01891	2	18	10	6	837564	550994	25.5	27.6

HG01928	1	18	9	7	759781	648239	30.1	33
HG01928	2	15	9	4	759856	471487	30.1	33
HG01952	1	15	9	4	759697	471497	22.8	22.2
HG01952	2	7	3	2	397674	391881	22.8	22.2
HG01978	1	20	6	12	534232	870344	33.8	36.4
HG01978	2	17	9	6	759822	595326	33.8	36.4
HG02080	1	12	8	2	687148	391843	19.5	21.4
HG02080	2	8	4	2	535886	391807	19.5	21.4
HG02148	1	10	6	2	534205	391822	16.7	18.2
HG02148	2	7	1	4	250236	471511	16.7	18.2
HG02257	1	19	9	8	824977	675619	23.8	26.3
HG02257	2	6	2	2	271436	391456	23.8	26.3
HG02486	1	16	6	8	534487	711372	33.1	36.1
HG02486	2	20	4	14	470484	959636	33.1	36.1
HG02622	1	10	5	3	521596	423803	30.5	33.2
HG02622	2	23	13	8	1052836	673635	30.5	33.2
HG02630	1	28	12	8	1056262	630477	39	42.6
HG02630	2	15	5	8	521981	675007	39	42.6
HG02723	1	11	3	6	349226	595446	30.7	33.3
HG02723	2	22	9	11	806089	904953	30.7	33.3
HG02886	1	20	10	8	837679	775073	34.6	37.5
HG02886	2	18	7	9	622574	687898	34.6	37.5
HG03453	1	16	6	8	534514	721244	29.4	32.2
HG03453	2	16	5	9	482942	687867	29.4	32.2
HG03492	1	14	6	6	534368	550729	27.8	29.2
HG03492	2	16	5	8	572121	676340	27.8	29.2

HG03516	1	16	4	10	485950	800534	31.3	35
HG03516	2	18	8	8	658644	675642	31.3	35
HG03540	1	17	9	6	824836	595540	31.1	33.3
HG03540	2	16	4	10	373713	784206	31.1	33.3
NA18906	1	19	10	7	860724	659968	35.5	38.9
NA18906	2	19	6	11	596139	863389	35.5	38.9
NA19240	1	23	14	7	1182507	646481	42.7	46.8
NA19240	2	24	14	8	1127904	721084	42.7	46.8
NA20129	1	21	8	11	747191	845142	35.9	38.7
NA20129	2	18	9	7	759784	648271	35.9	38.7
NA21309	1	22	8	7	736759	683282	33.2	34.5
NA21309	2	13	5	6	521464	586315	33.2	34.5
HG03579	1	24	10	12	837616	880074	41	41.9
HG03579	2	19	11	6	959115	595995	41	41.9
HG00741	1	12	6	4	534380	471469	22.5	24.5
HG00741	2	12	6	4	534322	471481	22.5	24.5
HG02717	1	24	10	12	928410	967440	39.7	41.2
HG02717	2	17	9	6	811551	595443	39.7	41.2
HG002	1	17	6	9	534346	726341	29.9	33.6
HG002	2	16	3	11	322509	805748	29.9	33.6

Supplemental Table S4: Primate Assembly Statistics

Species	Scientific Name	Sample Name	Hap	Assembly method	TotalBp (Gbp)	N50	AuN
human	Homo sapiens	T2T	1	NA	3.117	150.62	154.571
human	Homo sapiens	HG02723	2	Verkko 1.4	3.088	96.32	110.644
gorilla	Gorilla gorilla	Jim	1	Verkko 1.4	3.505	151.56	161.831
gorilla	Gorilla gorilla	Jim	2	Verkko 1.4	3.35	150.8	157.04
chimp	Pan troglodytes	AG18354_PTR	1	Verkko 1.4	3.146	147.43	147.435
chimp	Pan troglodytes	AG18354_PTR	2	Verkko 1.4	3.02	140.84	145.971
bonobo	Pan paniscus	PR00251_PPA	1	Verkko 1.4	3.206	147.03	149.437
bonobo	Pan paniscus	PR00251_PPA	2	Verkko 1.4	3.071	147.48	146.475
Sumatran orangutan	Pongo abelii	AG06213_PAB	1	Verkko 1.4	3.151	144.74	149.937
Sumatran orangutan	Pongo abelii	AG06213_PAB	2	Verkko 1.4	3.088	140.6	147.769
Bornean orangutan	Pongo pygmaeus	AG05252_PPY	1	Verkko 1.4	3.153	140.99	148.428
Bornean orangutan	Pongo pygmaeus	AG05252_PPY	2	Verkko 1.4	3.059	140.97	146.581
macaque	Macaca mulata	AG07107	1	hifiasm 0.15.2	3.111	18.81	23.74
macaque	Macaca mulata	AG07107	2	hifiasm 0.15.2	3.121	19.01	36.949
marmoset	Callithrix jacchus	CJ1700_CJA	1	hifiasm 0.15.2	2.931	103.97	95
marmoset	Callithrix jacchus	CJ1700_CJA	2	hifiasm 0.15.2	2.909	87.06	84.954
gelada	Theropithecus gelada	DRT_2020_14	1	hifiasm 0.15.2	3.112	109.07	108.157
gelada	Theropithecus gelada	DRT_2020_14	2	hifiasm 0.15.2	3.11	93.89	89.908
gibbon	Symphalangus syndactylus	Jambi_SSY	1	Verkko 1.4	3.241	144.67	136.725
gibbon	Symphalangus syndactylus	Jambi_SSY	2	Verkko 1.4	3.106	145.5	134.762
owl monkey	Aotus lemurinus	86718_ANA	1	hifiasm 0.15.2	3.098	55.92	55.764
owl monkey	Aotus lemurinus	86718_ANA	2	hifiasm 0.15.2	3.041	44.99	54.014
mouse lemur	Microcebus murinus	Inina_MMUR	1	hifiasm 0.15.2	2.355	29.39	37.127
mouse lemur	Microcebus murinus	Inina_MMUR	2	hifiasm 0.15.2	2.337	26.77	33.053

Supplemental Table S5: Primate TBC1D3 copy number

Species	Scientific Name	Number of TBC1D3 loci per haplotype | (Hap1 | Hap2)	Number of TBC1D3 genes (Hap1 | Hap2)
Human ***	Homo sapiens	4 | 4	11 | 22
Gorilla*	Gorilla gorilla	4 | 4	13 | 11
Chimpanzee*	Pan troglodytes	4 | 4	8 | 8
Bonobo*	Pan paniscus	4 | 4	9 | 9
Sumatran orangutan*	Pongo abelii	8 | 9	23 | 23
Bornean orangutan*	Pongo pygmaeus	5 | 7	17 | 21
Gibbon*	Symphalangus syndactylus	11 | 11	31 | 30
Macaque**	Macaca mulatta	4 | 4	29 | 20
Gelada	Theropithecus gelada	3 | 3	31 | 24
Owl monkey**	Aotus nancymaae	4 | 4	8 | 8
Marmoset**	Callithrix jacchus	2 | 2	2 | 2
Mouse Lemur	Microcebus murinus	0 | 0	0 | 0
*Genomes sequenced as part of the Primate T2T Consortium (Makova et al. 2023).			
**Genomes sequenced and assembled as part of Mao et al. 2024.			
***Genomes sequenced as part of human T2T project and Human Pangenome Reference Consortium (Nurk et al. 2022; Liao et al. 2023)			

Supplemental Table S6: TBC1D3 blastn in mouse lemur assembly

Mouse Lemur Contig	Contig Length (Kbp)	Alignment Start	Alignment End	Percent Identity	matches	mismatches	deletion events
h2tg000134l	1703	47882	48001	98.3	117	0	1
h2tg000138l	3866	3827725	3827818	98.9	92	0	0
h2tg000004l	72074	17977134	17977257	99.3	122	0	0
h2tg000112l	13746	5126011	5126079	100	68	0	0
h2tg000045l	58162	18906962	18907040	100	78	0	0
h2tg000044l	56769	10440427	10440488	98.4	61	0	1
h2tg000123l	38913	25474620	25474704	96.5	82	0	2
h2tg000038l	10979	5998913	5999012	98	98	0	1
h2tg000044l	56769	35690062	35690115	98.1	53	0	1

h2tg000049l	19634	15036293	15036413	96.6	115	0	2
h2tg000100l	6749	1090755	1090813	98.3	58	0	1
h2tg000317l	734	477484	477689	97	194	0	5
h2tg000317l	734	476107	476201	97.9	94	0	2
h2tg000138l	3866	3827725	3827818	98.9	92	0	0

Supplemental Table S7: Primate Iso-Seq TBC1D3 Mapping Data

Sample	Species	Reference	TBC1D3 reference mappings	Total Reads	Reads mapped to TBC1D3
Human Fetal Brain	human	T2T Version 2	14	2.33E+05	33138
Human IPSC	human	T2T Version 2	14	2.64E+08	11366
gorilla_hap1_testis_1	gorilla	Jim_GGO Hap1	15	4.68E+06	557
gorilla_hap2_testis_1	gorilla	Jim_GGO Hap2	15	4.68E+06	557
chimp_hap1_testis_2	chimpanzee	AG18354_PTR_Hap1	8	3.00E+06	1461
chimp_hap2_testis_2	chimpanzee	AG18354_PTR_Hap2	8	3.00E+06	1461
bonobo_hap1_testis_2	bonobo	PR00251_PPA_Hap1	9	2.23E+06	915
bonobo_hap2_testis_2	bonobo	PR00251_PPA_Hap2	9	2.23E+06	915
sum_orang_hap1_testis_1	Sumatran orangutan	AG06213_PAB_Hap1	28	3.11E+06	2032
sum_orang_hap2_testis_2	Sumatran orangutan	AG06213_PAB_Hap2	25	3.11E+06	2032
born_orang_hap1_testis_1	Bornean orangutan	AG05252_PPY_Hap1	20	4.20E+06	6836
born_orang_hap2_testis_1	Bornean orangutan	AG05252_PPY_Hap2	22	4.20E+06	6836
marmoset_hap1_frontal_cortex	marmoset	CJ1700_CJA Hap1	1	4.85E+05	0
marmoset_hap2_frontal_cortex	marmoset	CJ1700_CJA Hap2	0	4.85E+05	0

Supplemental Table S8: absREL branch-site model multiple test correction

Branch	B	LRT	uncorrected p-value	corrected p-value	omega distribution	omega 1	omega 2
1	0	9.2694	0.0101	0.0034		0 (94%)	52.6 (6.1%)
2	0	0.1584	0.4153	0.4153		1.55 (100%)	0
3	0	0.3256	0.3686	0.7372		1000 (100%)	0

Supplemental Table S9: *TBC1D3* Flanking Nucleotide Diversity and Tajima's D

Chrom	Start	End	PI	TD	Callable	Sedef	Gene	Percentile Pi	Percentile Tajima's D	Region
chr17	36980000	37000000	0.00132334	-0.6399256	1	0.09375		72.83%	56.20%	TBC1D3_exp1_upstream
chr17	36990000	37010000	0.00102332	-0.9237548	1	0.04665		50.82%	37.16%	TBC1D3_exp1_upstream
chr17	37000000	37020000	0.0006507	-1.1064736	1	0	CCL18	16.46%	25.65%	TBC1D3_exp1_upstream
chr17	37010000	37030000	0.00075038	-1.3290654	1	0.174	CCL18	24.98%	14.47%	TBC1D3_exp1_upstream
chr17	37020000	37040000	0.0011296	-1.3076028	1	0.674	CCL3	59.76%	15.37%	TBC1D3_exp1_upstream
chr17	37450000	37470000	0.00045802	-1.2486995	1	0		4.83%	18.07%	TBC1D3_exp1_downstream
chr17	37460000	37480000	0.00046682	-1.7870201	1	0	ZNHIT3	5.19%	2.60%	TBC1D3_exp1_downstream
chr17	37470000	37490000	0.00050652	-1.7946004	1	0	ZNHIT3,MYO19	7.02%	2.52%	TBC1D3_exp1_downstream
chr17	38860000	38880000	0.00110097	-0.42	1	1		57.46%	69.31%	TBC1D3_exp2_upstream

				305 28						
chr 17	3887 0000	3889 0000	0.001 31639	-0.01 597 06	1	1		72.41%	86.60%	TBC1D3_exp2_upstream
chr 17	3888 0000	3890 0000	0.000 94435	-0.59 937 75	1	0.88 045		43.51%	58.77%	TBC1D3_exp2_upstream
chr 17	3889 0000	3891 0000	0.000 80273	-1.18 806 71	1	0.66 825		29.88%	21.16%	TBC1D3_exp2_upstream
chr 17	3912 0000	3914 0000	0.000 41514	-0.35 250 36	1	1		3.28%	73.09%	TBC1D3_exp2_downstream_last_gc
chr 17	3913 0000	3915 0000	0.000 47718	0.25 023 104	1	1		5.63%	92.80%	TBC1D3_exp2_downstream_last_gc
chr 17	3914 0000	3916 0000	0.000 70144	0.22 350 642	1	1		20.65%	92.32%	TBC1D3_exp2_downstream_last_gc
chr 17	3916 0000	3918 0000	0.000 54192	-0.71 405 11	1	0.91 875	MRPL 45	8.91%	51.35%	TBC1D3_exp2_downstream_last_SD

Supplemental Table S10: Primate genome assembly statistics

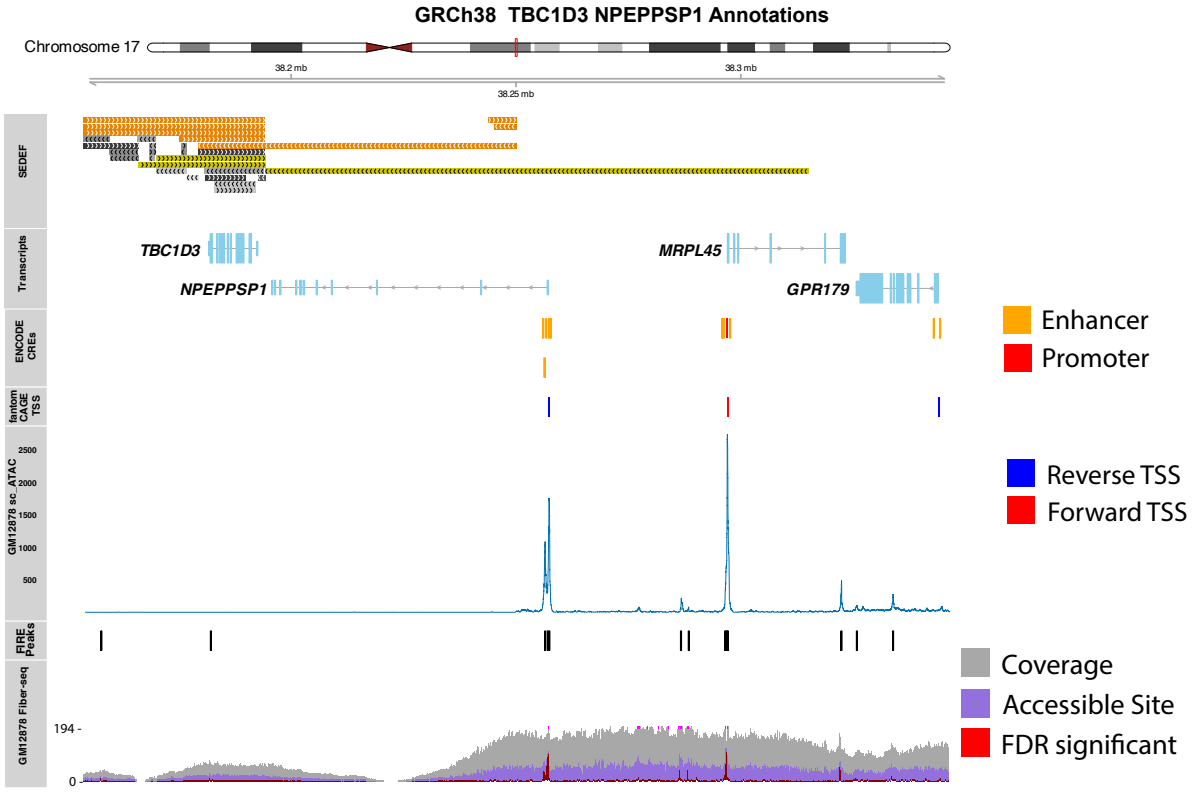
Species	Scientific Name	HiFi Coverage	ONT Coverage	Assembly Method	Assembly N50 (Hap1 | Hap2) (Mbp)
Gelada	Theropithecus gelada	50.24	NA	Hifiasm 0.15.2	109.0 | 93.6
Mouse Lemur	Microcebus murinus	30.25	NA	Hifiasm 0.15.2	29.4 | 26.7
Human	Homo sapiens	30	120	NA	150.62
Gorilla*	Gorilla gorilla	107	165	Verkko 1.4	151.43 | 150.80
Chimpanzee*	Pan troglodytes	69	127	Verkko 1.4	147.43 | 140.84
Bonobo*	Pan paniscus	131	169	Verkko 1.4	147.03 | 147.48
Sumatran orangutan*	Pongo abelii	102	91.9	Verkko 1.4	143.46 | 140.60
Bornean orangutan*	Pongo pygmaeus	62.6	200	Verkko 1.4	139.77 | 139.21
Gibbon*	Symphalangus syndactylus	83.2	111	Verkko 1.4	144.67 | 145.50

Macaque**	Macaca mulatta	38.91	NA	Hifiasm 0.15.2	18.81 | 19.01
Owl Monkey**	Aotus nancymaae	36.57	NA	Hifiasm 0.15.2	55.92 | 44.99
Marmoset**	Callithrix jacchus	39.08	NA	Hifiasm 0.15.2	103.97 | 87.06
*Genomes sequenced as part of the Primate T2T Consortium (Makova et al. 2023).					
**Genomes sequenced and assembled as part of Mao et al. 2024.					

Supplemental Table S11: TBC1D3 Probes for Iso-Seq Capture

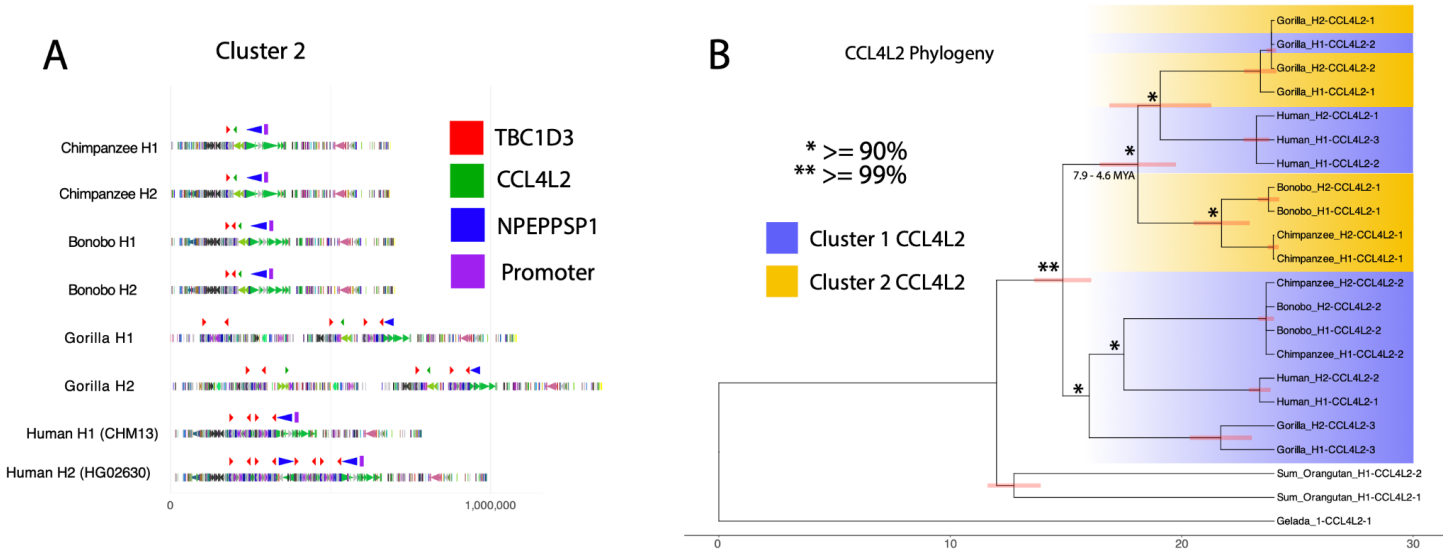
Target	Sequence
TBC1D3_target_1_1	ATGGACGTGGTAGAGGTCGCGGGCAGTTGGTGGGCACAAGAGCGAGAGGACATCATTATGAAATACGAAAAGGGACACCGAGCTGGGCTGCCAGAGGACAAGGGGCCTAAGCCTTTTCGA
TBC1D3_target_1_2	AGCTACAACAACAACGTCGATCATTTGGGGATTGTACATGAGACGGAGCTGCCTCCTCTGACTGCGCGGGAGGCGAAGCAAATTCGGCGGGAGATCAGCCGAAAGAGCAAGTGGGTGGAT
TBC1D3_target_1_3	ATGCTGGGAGACTGGGAGAAATACAAAAGCAGCAGAAAGCTCATAGATCGAGCGTACAAGGGAATGCCCATGAACATCCGGGGCCCGATGTGGTCAGTCCTCCTGAACACTGAGGAAATG
TBC1D3_target_1_4	AAGTTGAAAAACCCCGGAAGATACCAGATCATGAAGGAGAAGGGCAAGAGGTCATCTGAGCACATCCAGCGCATCGACCGGGACGTAAGCGGGACATTAAGGAAGCATATATTCTTCAGG
TBC1D3_target_1_5	GATCGATACGGAACCAAGCAGCGGGAACTACTCCACATCCTCCTGGCATATGAGGAGTACAACCCGGAGGTGGGCTACTGCAGGGACCTGAGCCACATCGCCGCCTTGTTCCTCCTCTAT
TBC1D3_target_1_6	CTTCCTGAGGAGGATGCATTCTGGGCACTGGTGCAGCTGCTGGCCAGTGAGAGGCACTCCCTGCAGGGATTTCACAGCCCAAATGGCGGGACCGTCCAGGGGCTCCAAGACCAACAGGAG
TBC1D3_target_1_7	CATGTGGTAGCCACGTCACAACCCAAGACCATGGGGCATCAGGACAAGAAAGATCTATGTGGGCAGTGTTCCCCGTTAGGCTGCCTCATCCGGATATTGATTGACGGGATCTCTCTCGGG
TBC1D3_target_1_8	CTCACCCTGCGCCTGTGGGACGTGTATCTGGTAGAAGGCGAACAGGCGTTGATGCCGATAACAAGAATCGCCTTTAAGGTTCAGCAGAAGCGCCTCACGAAGACGTCCAGGTGTGGCCCG
TBC1D3_target_1_9	TGGGCACGTTTTTGCAACCGGTTCGTTGATACCTGGGCCAGGGATGAGGACACTGTGCTCAAGCATCTTAGGGCCTCTATGAAGAAACTAACAAGAAAGAAGGGGGACCTGCCACCCCCA
TBC1D3_target_1_10	GCCAAACCCGAGCAAGGGTCGTCGGCATCCAGGCCTGTGCCGGCTTCACGTGGCGGGAAGACCCTCTGCAAGGGGGACAGGCAGGCCCCTCCAGGCCCACCAGCCCGGTTCCCGCGGCCC
TBC1D3_target_1_11	ATTTGGTCAGCTTCCCCGCCACGGGCACCTCGTTCTTCCACACCCTGTCCTGGTGGGGCTGTCCGGGAAGACACCTACCCTGTGGGCACTCAGGGTGTGCCCAGCCCGGCCCTGGCTCAG
TBC1D3_target_1_12	GGAGGACCTCAGGGTTCCTGGAGATTCCTGCAGTGGAACTCCATGCCCCGCCTCCCAACGGACCTGGACGTAGAGGGCCCTTGGTTCCGCCATTATGATTTCAGGCAGAGCTGCTGGGTC
TBC1D3_target_1_13	CGTGCCATATCCCAGGAGGACCAGCTGGCCCCCTGCTGGCAGGCTGAACACCCTGCGGAGCGGGTGAGATCGGCTTTCGCTGCACCCAGCACTGATTCCGACCAGGGCACCCCCTTC

APPENDIX B. SUPPLEMENT FOR CHAPTER 3



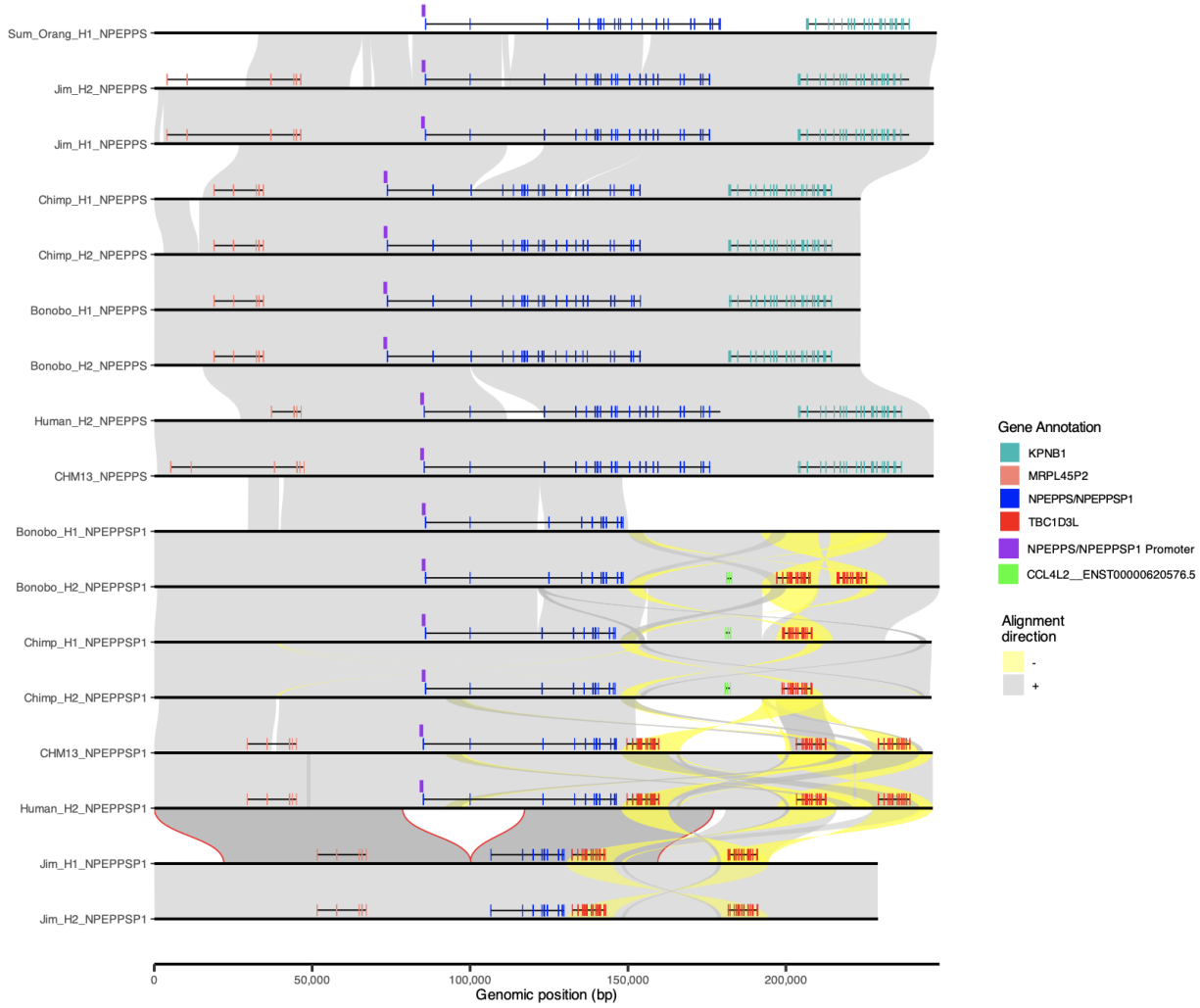
Supplementary Figure 1: GRCh38 chromatin accessible sites

Regulatory elements identified by the ENCODE project and scATAC are compared to Fiber-seq data of the GM2878 cell line (Vollger et al. 2025). TSS: transcription start site; CRE: cis-regulatory element.



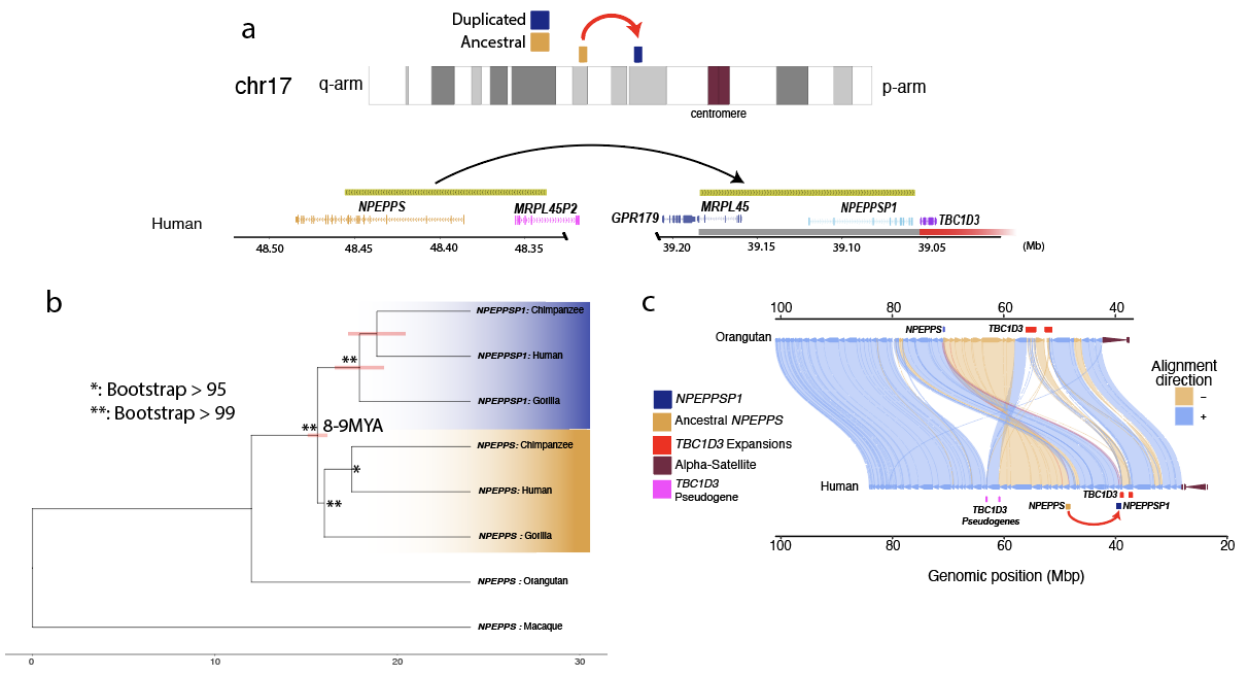
Supplementary Figure 2. *CCL4L2* structure and phylogeny

(A) Genomic architecture of *TBC1D3* Cluster 2 in African apes. *CCL4L2* obstructs *NPEPPSP1* from *TBC1D3* and co-opts fusion expression in the *Pan* genera. This structure is fixed in both chimpanzee and bonobo. **(B)** Maximum likelihood phylogeny of *CCL4L2* sequence. Timing estimates of the *CCl4L2* obstruction event are between 4.6 and 7.9 MYA, and human and gorilla *CCL4L2* copies are more similar to one another than *Pan*.



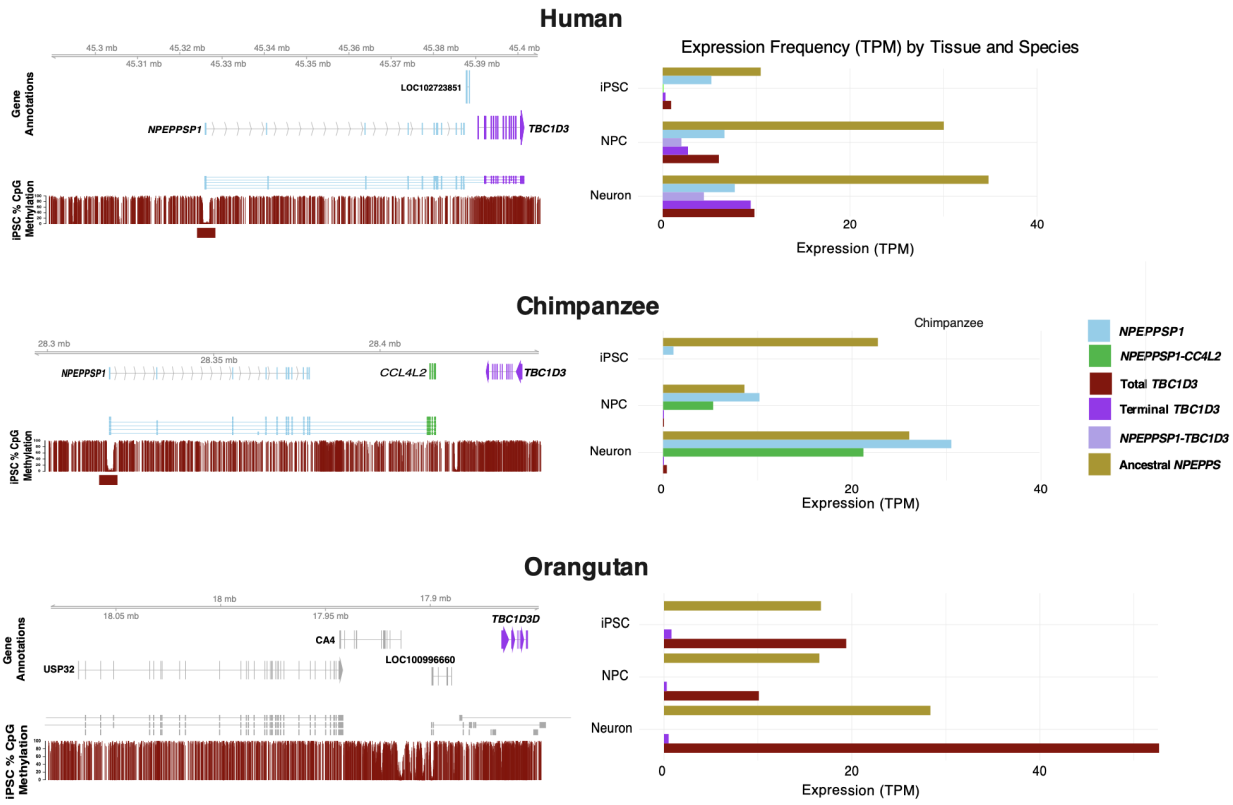
Supplementary Figure 3. Gorilla deletion

Alignments of the *NPEPPS* (ancestral) and *NPEPPSP1-TBC1D3* (derived) loci are illustrated, showing the orthologous *NPEPPSP1* promoter inherited across African apes but lost in the gorilla lineage.



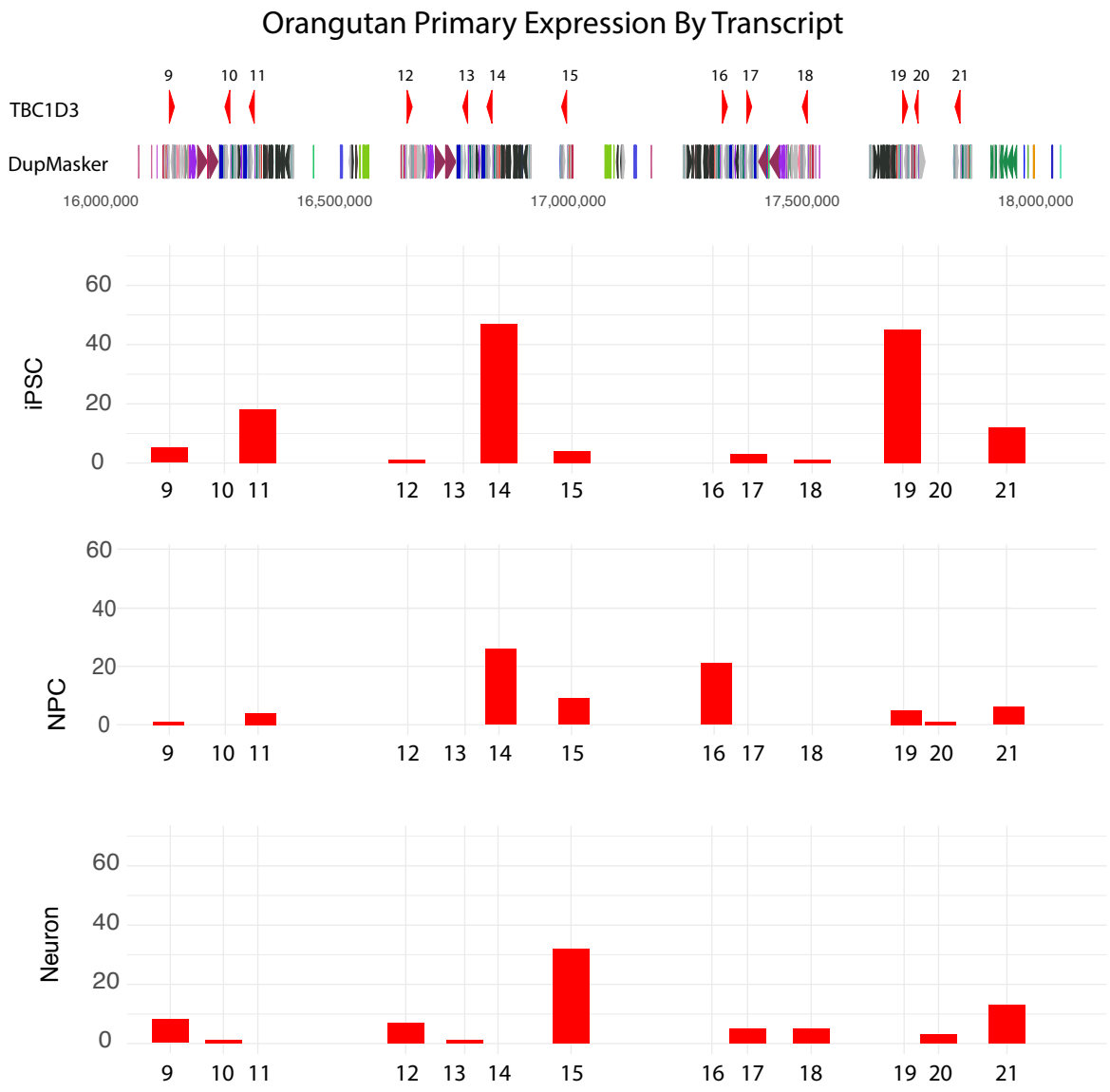
Supplementary Figure 4. Evolutionary origin of *NPEPPSP1* regulatory sequence

(A) *NPEPPS* SD. A ~119 kbp duplication of the N-terminus of *NPEPPS* and C-terminus *MRPL45P1* relocated ~9.15 Mbp to the *TBC1D3* Cluster 2 region. **(B)** *NPEPPS1* SD evolution. A phylogeny of 15 kbp of the *NPEPPS* duplication most proximal to *TBC1D3* and macaque (25 MYA divergence) predicts that the duplication occurred ~8-9 MYA. **(C)** SVbyEye illustrates the repositioning of *NPEPPSP1* to *TBC1D3*.



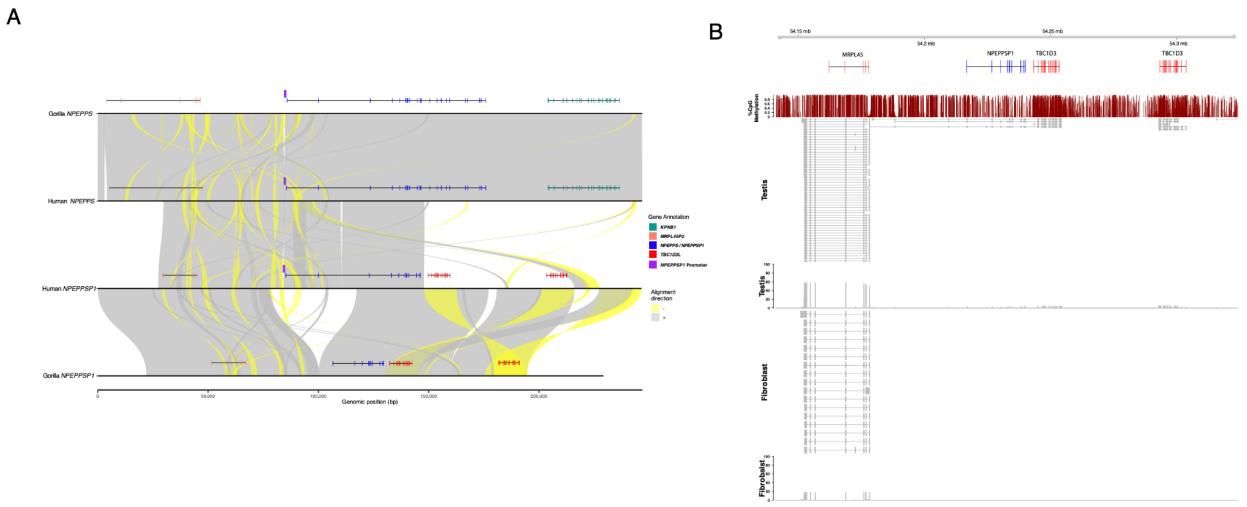
Supplementary Figure 5. Comparative expression with *NPEPPS* and total *TBC1D3* in a neuronal developmental cell culture model of great apes

Gene annotation and methylation of *NPEPPSP1-TBC1D3* can be observed on the left for human (top), chimpanzee (middle), and orangutan (bottom). On the right, expression of *NPEPPSP1*, terminal *TBC1D3*, and their fusion are compared to global *TBC1D3* and *NPEPPS* expression, normalized as transcripts per million (TPM; Methods).



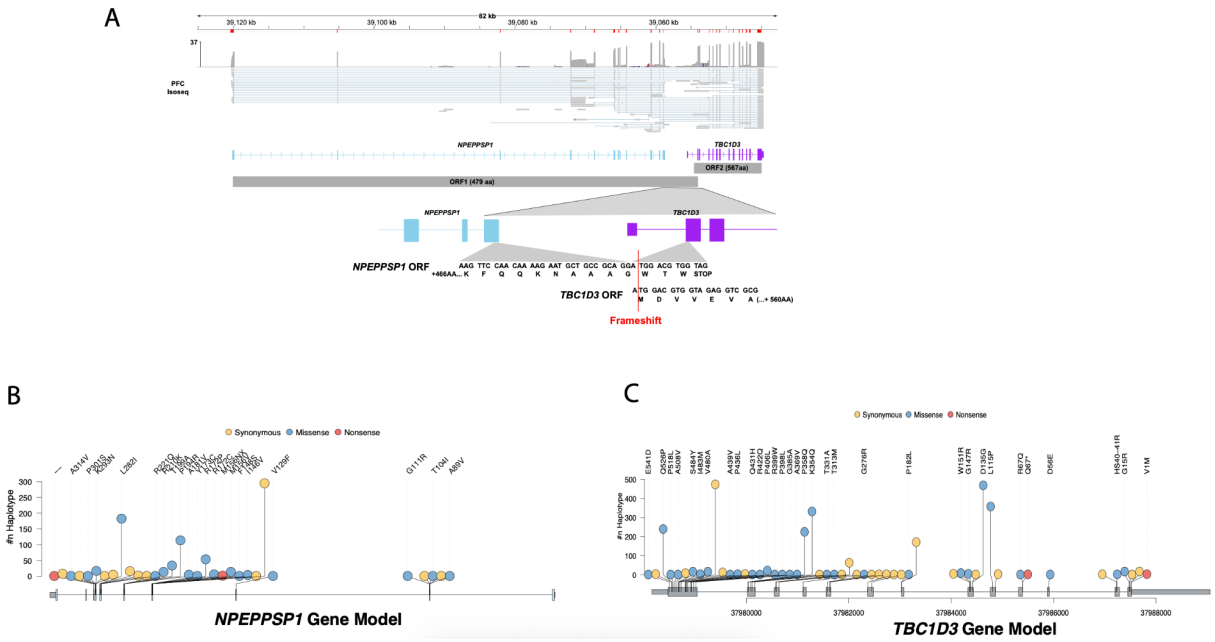
Supplementary Figure 6. Orangutan expression by *TBC1D3* transcript

Orangutan primary transcripts map to numerous internal *TBC1D3* paralog copies (*TBC1D3*-14,15,16,19).



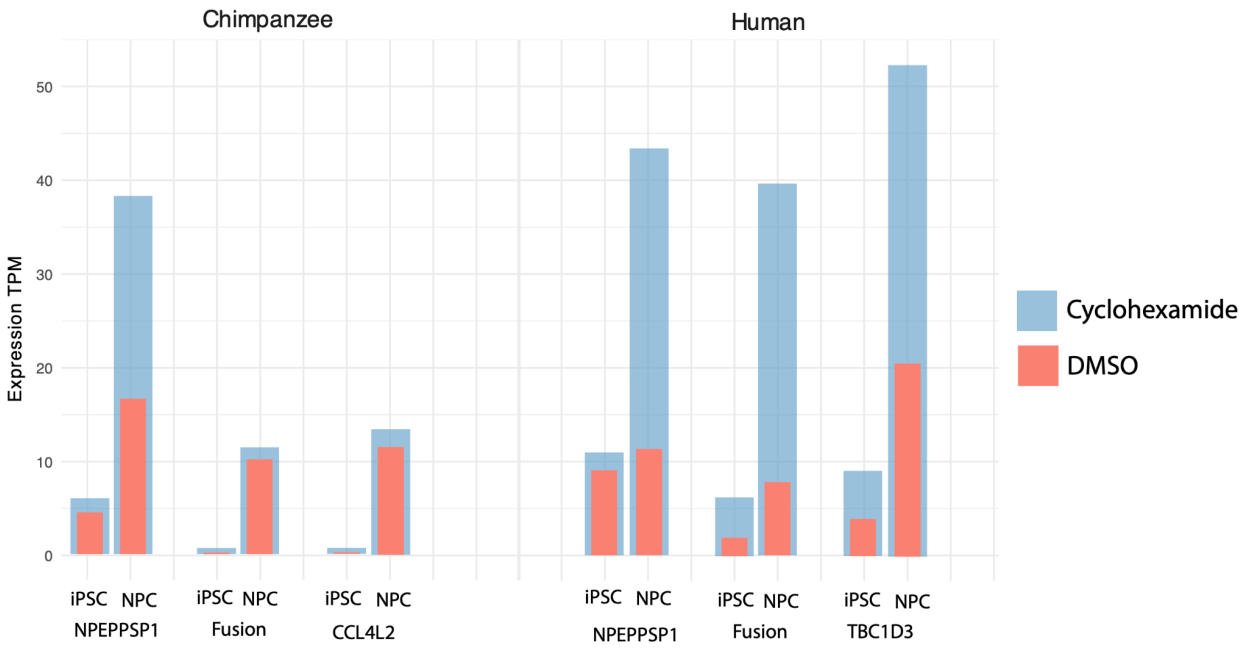
Supplementary Figure 7. Gorilla *NPEPPSP1* promoter deletion and *TBC1D3* expression

(A) Comparison of gorilla and human genome organization for *NPEPPSP1-TBC1D3* (SVbyEye) highlights a 38 kbp deletion removing the promoter and two exons of *NPEPPSP1* in both haplotypes of the gorilla. **(B)** Gorilla isoforms and expression. Mapping of gorilla Iso-Seq data to the gorilla locus shows no evidence of transcript initiation from *NPEPPSP1* promoter. Instead, abundant transcription and isoforms are observed from MRPL45, and only three transcripts from testis could be identified that include non-deleted exons from *NPEPPSP1* and *TBC1D3*.



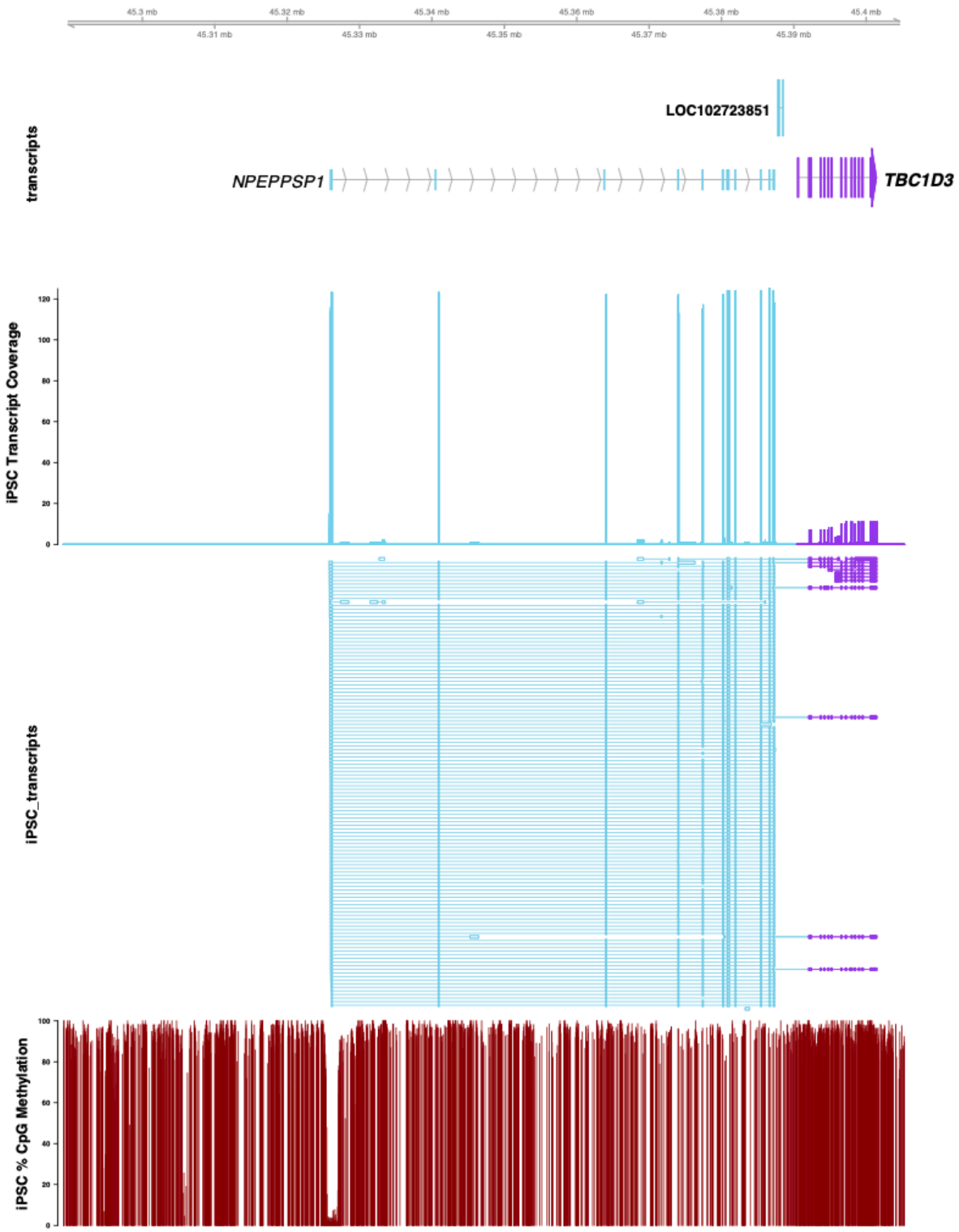
Supplementary Figure 8. *NPEPPSP1-TBC1D3* retains two ORFs

(A) Iso-Seq data from the prefrontal cortex illustrates the fusion transcript across *NPEPPSP1* and *TBC1D3*. Below are two open reading frames (ORFs) identified with ORFfinder (Methods), which span *NPEPPSP1* and *TBC1D3* in separate reading frames overlapping by 13 nucleotides. **(B)** Lollipop plots of *NPEPPSP1* illustrate synonymous (yellow), missense (blue), and nonsense (red) variants. Three missense variants were identified in 2,013 genome assemblies. **(C)** Same as (B) but with *TBC1D3* ORF2. One missense variant was identified across the gene body.



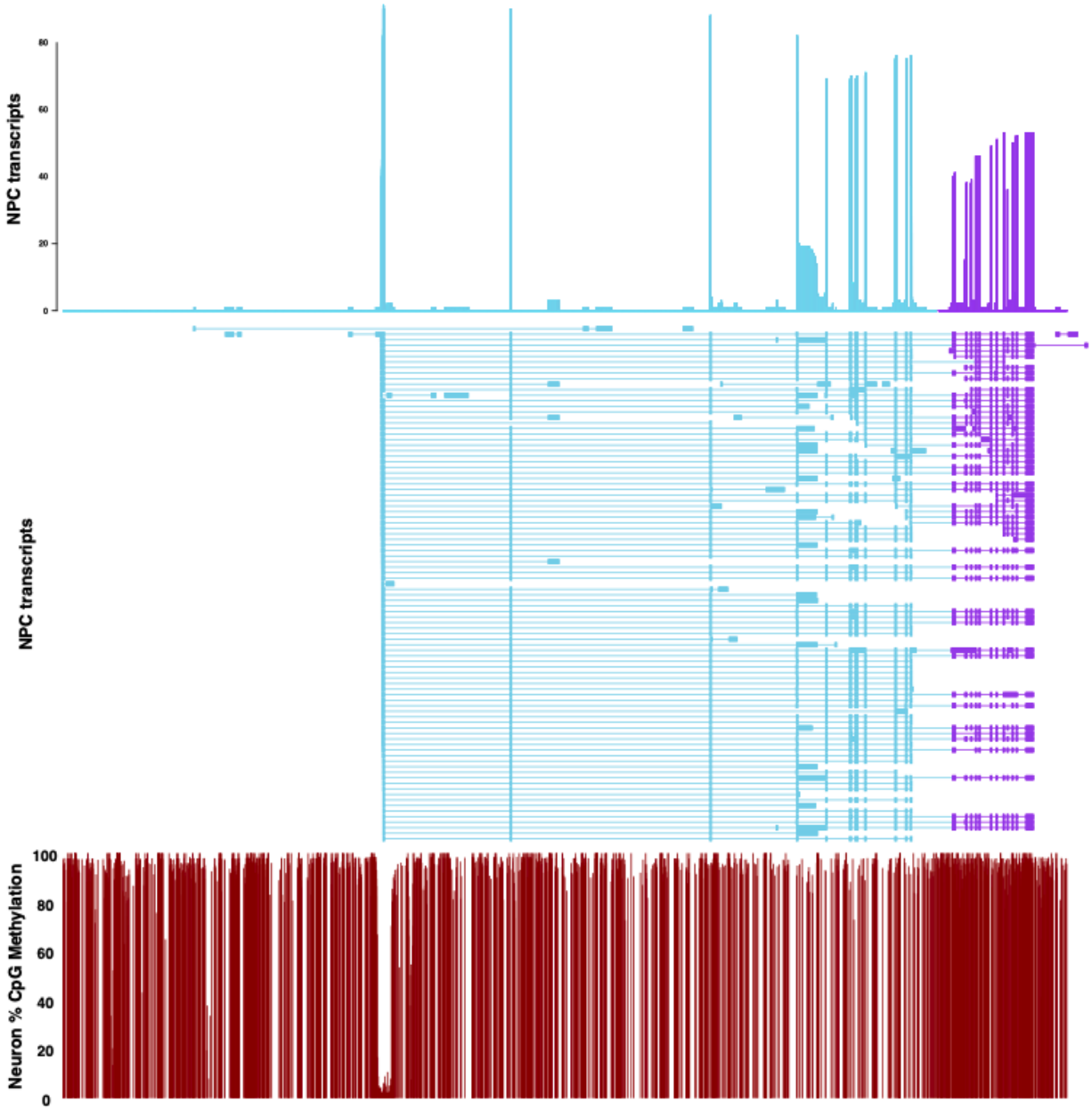
Supplementary Figure 9. Disruption of nonsense mediated decay (NMD)

iPSCs and NPCs from chimpanzee and humans were treated with either DMSO or cycloheximide, a disruptor of NMD, to investigate posttranscriptional fate of fusion genes. Notably, *NPEPPSP1* is preferentially rescued in both species in NPCs relative to *TBC1D3*, though in humans both *NPEPPS1*, *TBC1D3*, and the fusion isoform increase equally.



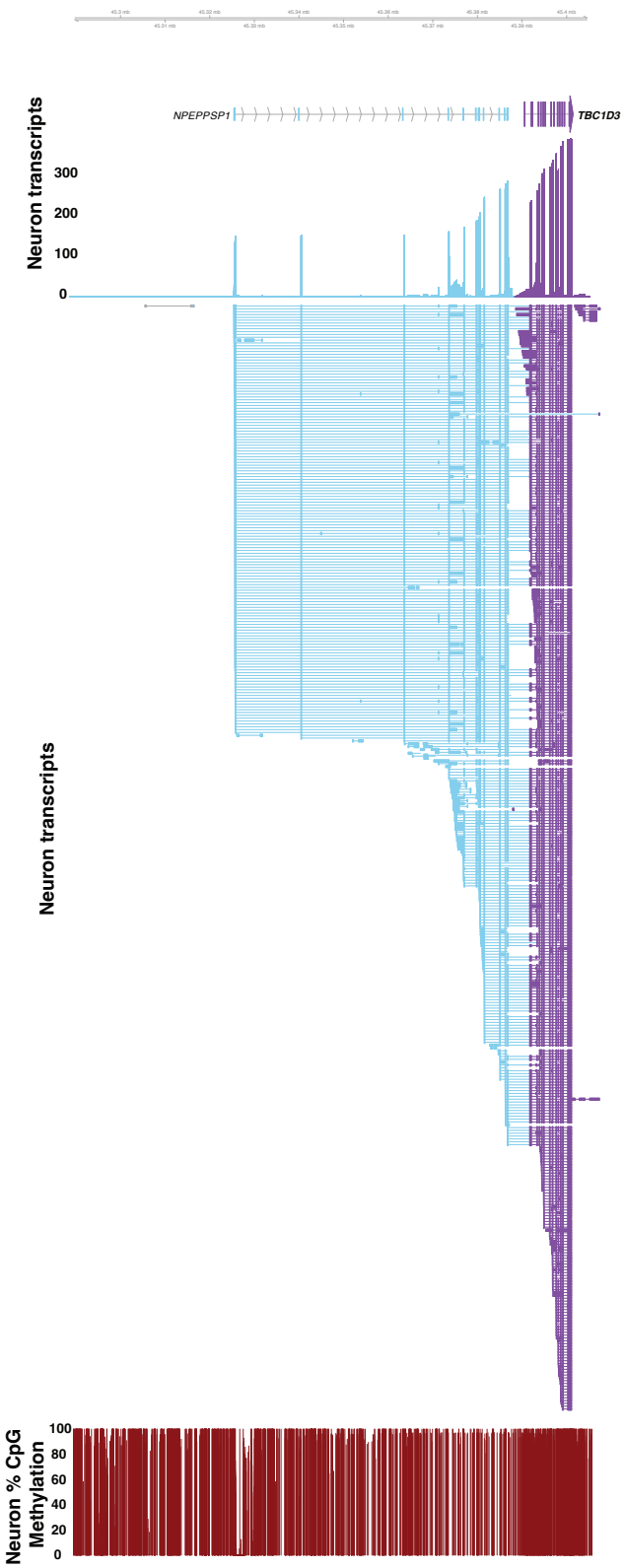
Supplementary Figure 10. Human iPSC expression and methylation

Human iPSC gDNA and full-length transcripts were mapped back to their donor-specific genome assembly (DSA) and annotated for *NPEPPSP1* (blue) or *TBC1D3* (purple) gene models. The *NPEPPSP1* promoter may be seen as a dip in methylation at *NPEPPSP1* exon 1.



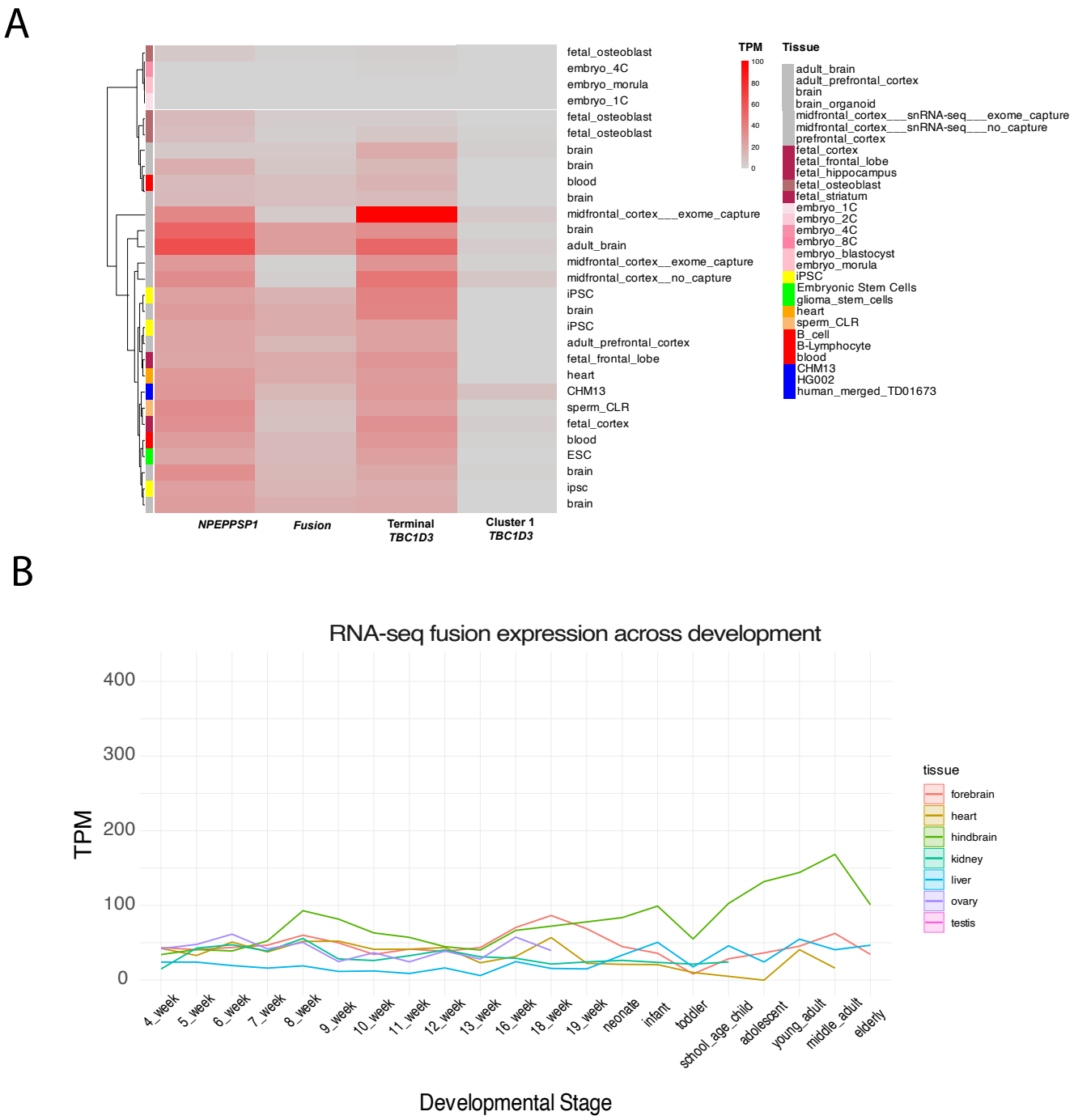
Supplementary Figure 11. Human NPC expression and methylation

Human NPC gDNA and full-length transcripts were mapped back to their DSA and annotated for *NPEPPSP1* (blue) or *TBC1D3* (purple) gene models. The *NPEPPSP1* promoter may be seen as a dip in methylation at *NPEPPSP1* exon.



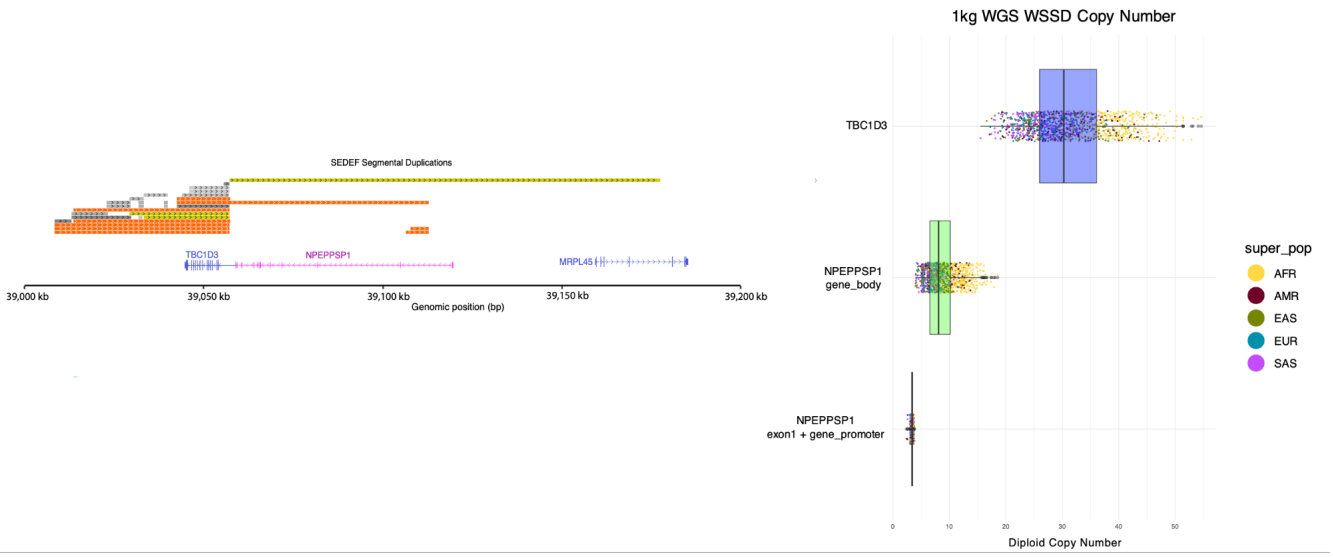
Supplementary Figure 12. Human neuron expression and methylation

Human neuron gDNA and full-length transcripts were mapped back to their DSA and annotated for *NPEPPSP1* (blue) or *TBC1D3* (purple) gene models. The *NPEPPSP1* promoter may be seen as a dip in methylation at *NPEPPSP1* exon 1. Notably, in neurons, *TBC1D3* has increased in expression relative to *NPEPPSP1*, in contrast to NPCs or iPSCs.



Supplementary Figure 13. *TBC1D3* expression

(A) Full-length transcriptome sequencing of various libraries mapped to *NPEPPSP1*, *NPEPPSP1-TBC1D3* fusion, terminal *TBC1D3*, or any Cluster 1 *TBC1D3* (control). **(B)** RNA-seq of *NPEPPSP1-TBC1D3* fusion across development.



Supplementary Figure 14. *NPEPPSP1* promoter fixed in copy number.

Left: Segmental duplication (SD) annotation of the *NPEPPSP1-TBC1D3* locus illustrates the numerous high-identity SDs across both *TBC1D3* and *NPEPPSP1* but does not include the *NPEPPSP1* exon 1 or its upstream promoter. Right: Diploid copy number estimates of *TBC1D3*, *NPEPPSP1*, and *NPEPPSP1* promoter in the 1000 Genomes Project (1KGP) identified by whole-genome shotgun sequence detection (WSSD; Sudmant et al. 2010) illustrate the fixed copy number of the *NPEPPSP1-TBC1D3* promoter relative to either *NPEPPSP1* or *TBC1D3*.

Supplementary Table 1: *NPEPPS* vs. *NPEPPSP1* Mass Spectra Peptide Observations

NPEPPSP1 Tryptic Peptides	Datasets where peptide observed	Proteins with peptide
MWLAAAAPSL	23	A6NEC2, P55786, E9PPD4, E7EWZ2, …12 total
MWLAAAAPSLARRLLFL	1 (breast cancer)	A6NEC2, P55786, E9PPD4, …7 total
ARRLLFLGPP	0	0
PADVSPINCS	2	A6NEC2
FARTPVTSTY	3	A6NEC2
CVCVYTPVGK	3 (Cytoplasmic Proteins of Untreated Cells)	A6NEC2
AGAMENWDLV	5 ()	A6NEC2
NPEPPS Tryptic Peptides	Datasets where peptide observed	Proteins with peptide
MASFSMDCSP	MSV000079835	E9PLK3, E9PJF9, E9PPZ2,
SFCVPGLWNP	MSV000083043, MSV000079835, MSV000080679, MSV000080826, MSV000080851, MSV000084248	E9PLK3, E9PP11, E9PJF9, E9PPZ2, E9PP11, E9PLK3, E9PJF9, E9PPZ2
PADVSPINYS	117	P55786, E9PLK3, E9PJF9, … (27 total)
FARTPVMSTY	7	P55786, E9PLK3, P55786, …7 total
CVRVYTPVGK	16	P55786, E9PLK3, A0A7I2V389, …8 total
AGAMENWGLV	22	

Supplementary Table 2: iPSC-NPC-iNeuron Transcript Counts

Species	Tissue	TOTAL NPEPPS	NPEPPS	NPEPPSP1	TBC1D3	Terminal TBC1D3	CCL4L2	Fusions	Total Reads
Chimp	ipsc	202	193	9	0	0	0	0	8489660
Chimp	npc	88	75	89	1	0	50	46	8710214
Chimp	neuron	1214	559	655	7	1	511	454	21456687
Human	ipsc	172	115	57	7	3	0	1	10983570
Human	npc	335	275	60	30	25	0	18	9156530
Human	neuron	937	767	170	9	208	0	98	22035016
Orang	ipsc	228	228	0	254	11	0	0	13648087
Orang	npc	151	151	0	89	3	0	0	9127847
Orang	neuron	605	605	0	1114	11	0	0	21333770

Supplementary Table 3: iPSC-NPC-iNeuron Transcript Counts (TPM)

Species	Tissue	NPEPPS TPM	TBC1D3 TPM	Terminal TBC1D3 TPM	CCL4L2 TPM	Fusion TPM
Chimp	ipsc	22.73	0	0	0	0
Chimp	npc	8.61	0.1	0	5.7	5.3
Chimp	neuron	26.05	0.3	0	23.8	21.2
Human	ipsc	10.47	0.6	0.3	0	0.1
Human	npc	30.03	3.3	2.7	0	2
Human	neuron	34.81	0.4	9.4	0	4.4
Orang	ipsc	16.71	18.6	0.8	0	0
Orang	npc	16.54	9.8	0.3	0	0
Orang	neuron	28.36	52.2	0.5	0	0

Supplementary Table 5: Human iPSC-NPC-iNeuron mappings by *TBC1D3* paralog

TBC1D3	Context	Promoter Identified	iPSC-DMSO	iPSC-CHX	NPC-DMSO	NPC-CHX	Neuron-DMSO
0	NPEPPSP1	yes	24	33	35	125	146
1	Cluster2 (NPEPPSP1 fusion)	yes	10	23	60	142	196
2	Cluster2		0	0	0	0	0
3	Cluster2		0	0	0	0	0
4	Cluster2		0	0	0	0	0
5	Cluster2		0	0	0	0	1
6	Cluster2		0	0	0	0	0
7	Cluster2		0	0	0	0	0
8	Cluster 1		0	0	0	0	0
9	Cluster 1		0	0	0	0	0
10	Cluster 1		0	0	0	0	0
11	Cluster 1		0	0	0	0	0

Supplementary Table 5: Neurospheres mappings by *TBC1D3* paralog

TBC1D3	Context	Promoter Identified	Neurospheres	Percentage-of-TBC1D3-transcripts
0	NPEPPSP1	yes	51	
1	Cluster2 (NPEPPSP1-fusion)	yes	116	56.31%
2	Cluster2		0	0.00%
3	Cluster2		0	0.00%
4	Cluster2		11	5.34%
5	Cluster 2		2	0.97%
6	Cluster2		0	0.00%
7	Cluster2		0	0.00%
8	Cluster2		8	3.88%

9	Cluster 1		15	7.28%
10	Cluster 1		4	1.94%
11	Cluster 1		12	5.83%
12	Cluster 1		4	1.94%
13	Cluster 1		34	16.50%
		Total TBC1D3 Transcripts	206	

APPENDIX C. SUPPLEMENT FOR CHAPTER 4

Supplementary Table 1: Sample description for *in vitro* model

Sample	Context
Human TBC1D3 ORF	Test human TBC1D3
Bonobo TBC1D3 ORF	Test Bonobo TBC1D3 as comparison
Human TBC1D3 with C-terminus deletion	Isolate C-terminus from human TBC1D3
Human TBC1D3 with Bonobo TBC1D3 C-terminus	Isolate human TBC1D3 protein changes from human C-terminus
Bonobo TBC1D3 with Human TBC1D3 C-terminus	Isolate human C-terminus modification from other human TBC1D3 protein changes
Empty Vector	Control
NLS-EGFP	Control
NGN2	Negative control