Molecular Investigation of Human-Specific Neurobiology: From Cortical Evolution to Neurodevelopmental Disorders

by

Elizabeth A. Werren

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics and Anthropology)
in the University of Michigan
2022

Doctoral Committee:

Associate Professor Stephanie Bielas, Co-Chair
Associate Professor Abigail Bigham, University of California, Los Angeles, Co-Chair
Professor Jacinta Beehner
Associate Professor Sue Hammoud
Professor Jeffrey M. Kidd
Associate Professor Jacob Mueller

Elizabeth A. Werren

werrene@umich.edu

ORCID iD: 0000-0002-0151-2564

DEDICATION

This dissertation is dedicated to all individuals with neurodevelopmental disorders and their families.

I also dedicate my work to my late mother, Angela Werren, who devoted her life to helping and advocating for the neurodiverse community, and who taught me how to navigate and embrace my own learning disorder. Her unconditional support empowered me to pursue doctoral studies, to unapologetically follow my interests throughout this process, and to find the perseverance to never give up.

ACKNOWLEDGEMENTS

I want to thank my family for their continual love and support throughout this journey: my late mother, Angie Werren, and my father, Dwight Werren, for raising me to be strong and independent; my sister, Chelsea Prajapati, and my brother-in-law, Seeran Prajapati; my grandparents, Judy and Bill Eicher, and the late Sue and Herb Werren; Aunt Melanie, Uncle Mark, Uncle Dan, and Aunt Debbie; my cousins, Kenzie, Macie, and Maddie Tonn, and Andy and Caitie Eicher.

I want to thank all members of the Human Genetics and Anthropology graduate programs who have given me their support and guidance over the years. I am forever grateful to Dr. Aaron Sandel, Dr. John Mitani, Dr. Abbey Breidenstein, Dr. Obed Garcia, and Dr. Máire Malone who provided exceptional mentorship and helped shape my confidence early on in graduate school. I especially want to recognize Dr. Jordan Dalton, Paloma Contreras, and Andrew Bernard for being incredible friends and for always uplifting me. A special thanks to my committee members, Dr. Sue Hammoud, Dr. Jacob Mueller, Dr. Jeffrey Kidd, and Dr. Jacinta Beehner for imparting their wisdom and advice, as well as Dr. Stephanie Moon. I want to give a special thanks to administrative personnel Darlinda Flanigan, Molly Martin, and Jeff Holden for always offering tremendous help and kindness. I would like to express my appreciation for all the national and international travel opportunities for research and conferences made possible by Rackham Graduate School, the Anthropology and Human Genetics departments, and the researchers who facilitated these experiences, including Laura Diakiw, Dr. George Perry, Dr. Rindra

Rakotoarivony, and Dr. Mark Wilson. I also thank the Antonellis, Kalantry, Hammoud, Martin, and Camper labs for research help and use of equipment and reagents.

I would like to acknowledge all past and present members of the Bielas and Bigham labs. A special thanks to Paloma Contreras, Dr. Obed Garcia, Dr. Abbey Breidenstein, Dr. Ainash Childebayeva, and Kinsey Vear from the Bigham Lab. Special thanks to Dr. Anshika Srivastava, Dr. Yao-chang Tsan, Dr. Samma Regan, Dr. Khoa Le, Dr. Brian Mcgrath, Delia Perillo, Alba Guxholli, Charles Ryan, and Emily Peirent in the Bielas lab.

I am immensely grateful to have such wonderful friendships in this world from my favorite humans who have elevated me throughout this process. To Rachael Fenton Acra, Zachary David Lawson, and Massey Lee Pierce, whose friendships have stood the test of time. To Taylor Lovett and Britt Berkley who have been there for me in the hardest moments of my life and whose mutual love for sifakas is truly unmatched. Your love and friendship mean everything to me!

I want to recognize my favorite Ann Arbor coaches whose fitness classes have kept me going: Carson Crandell, Amanda Topping, Rachel Cohen, Marilyn A., and Sandra Verissimo. I also want to thank all the baristas and servers at my favorite study and dining spots (Espresso Royale, Sweetwaters, Detroit Filling Station, Chapala).

I am especially appreciative of the mentorship and guidance received from Dr. Stephanie Bielas and Dr. Abigail Bigham. You each have taught me different ways to navigate the world of science successfully and as a woman. You exemplify the characteristics that I admire, and hope to acquire, as a scientist. Dr. Bielas, you have significantly shaped my scientific interests and career path. I want to thank you for letting me join your lab with no background in neuroscience and little molecular biology skills. Nevertheless, you believed in me and challenged me to succeed every step of the way. I am forever grateful and lucky to have both of you in my life as an advisor, mentor, and friend.

Last, but not least, I want to thank my partner, Julien Weinstein, who has been my biggest cheerleader, my rock, and at countless times my personal chef and cleaning staff. Your dependability, emotional support, and endless help (and cheesy humor!) were fundamental in helping me accomplish my academic goals. You are my best friend, and I am so lucky to have you in my life. I also want to thank his family for their constant support, and my fur babies, Stella and Otto Werren, who were the best dissertation writing companions!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The human brain evolves under a unique set of adaptive pressures, distinguishing aspects of form and function from the brains of other primate species. An outstanding question in evolutionary biology is the biological tradeoff of human brain adaptations and neurodevelopmental/neuropsychiatric disorders (NND). Two ways to investigate proximate mechanisms of human evolutionary neurobiology are to study human-specific genetic variation and the genetic basis of neuropathology. This dissertation uses an evolutionary medicine approach to investigate three cases of functional divergence of genetic and molecular variation that mediate human-specific features of neural development implicated in NND.

Segmental duplications (SD) are a rich source of highly plastic genomic variation important for human-specific adaptive evolution and a major cause of NND. As result of SD, the neuroblastoma breakpoint gene family (*NBPF*) sequence has undergone significant expansion in humans, yet functions of expressed *NBPF* proteins remain unknown. Structural rearrangements of *NBPF* are associated with autism and schizophrenia, and are affected in 1q21.1 deletion/duplication syndrome (1q21DDS, OMIM:612474;612475). In this thesis, we characterize NBPF subcellular localization and expression profiles in human and chimpanzee models of corticogenesis. To investigate dosage effects on cortical development, *NBPF*-depleted and overexpression in cerebral organoids were generated. These findings support the hypothesis that NBPF dosage impacts proliferative dynamics in human corticogenesis, implicating *NBPF*s in the etiology of 1q21DDS.

Genetic variation in members of the CUB and Sushi Multiple Domains (*CSMD*) gene family of complement pathway regulators is associated with neuropsychiatric disorders. Unlike young *NBPF* duplicates, *CSMD* genes are fixed across species. However, evolved functions in synaptic pruning, specifically of CSMD1, contribute to species-specific synaptic plasticity and neuronal circuitry—features linked to human-specific cognitive adaptations. Expanding on these findings, we identify novel biallelic variants in *CSMD1* in individuals with overlapping features of neurodevelopmental disorders, namely microcephaly, intellectual disability, polymicrogyria, and epilepsy. Using *CSMD1* knockout human cortical organoids, we identify pathogenic mechanisms of NPC over-proliferation and premature differentiation. Our novel data expand CSMD1-associated functions to a variety of steps in corticogenesis, including neural proliferation, differentiation, neuronal migration, and synaptogenesis. Together, human genetics and functional findings implicate *CSMD1* as a novel genetic basis of NND.

Proper mRNA maturation and nuclear export regulate eukaryotic gene expression that underlies cellular and species diversity. mRNA processing is coordinated by the multimeric THO subcomplex of the TREX complex. THO component THOC6 is the genetic basis of autosomal recessive *THOC6* Intellectual Disability Syndrome (TIDS; OMIM:613680). Although TREX is considered conserved, its structural form is not, facilitated by THOC6 acquisition in metazoans. Consequently, functional divergence of mRNA processing and export coordination by TREX has co-evolved with transcriptomic complexity across species. In this dissertation, we generate mouse and human models of TIDS to investigate THOC6-dependent functions. We identify novel THOC6-associated mRNA splicing functions in neural tissue. We observe dysregulation of key signaling pathways in cortical organoids that dictate the transition from proliferative to neurogenic divisions, resulting in delayed differentiation and increased apoptosis—hallmarks of microcephaly. Lastly, THO is known to preferentially accumulate at repetitive regions and protect against transcription-induced instability. Our findings have further implications for THO processing/splicing of highly repetitive loci (e.g., functional human-specific duplicates) that shape human-specific features of corticogenesis.

Together, this thesis extends the contribution of evolutionary medicine research and models of molecular neofunctionalization to our understanding of the intersection of evolutionary and pathogenic mechanisms of human neural development.

Chapter 1
Introduction and Background

1.1 EVOLUTIONARY MEDICINE

The complex relationship between our evolutionary biology and socio-cultural-environmental factors shapes human health at both the individual and collective levels. Germline and somatic genetic variation, including genetic adaptation to past environmental pressures, interacts with socio-cultural practices, current environment, pathogen exposure, diet, and lifestyle variables to modulate modern disease risk and severity. The evolutionary history of our species has led to the evolution of diverse traits that facilitated our colonization of expansive and extreme environments across the globe. Nevertheless, evolution deals in tradeoffs and our species also acquired unique disease susceptibilities compared to our closest living relatives, including cardiovascular disease (Varki et al. 2009), neurodegenerative disease (Finch 2010), neurodevelopmental syndromes (Dennis and Eichler 2016; Dennis et al. 2017; Marques-Bonet et al. 2009) and neuropsychiatric disorders (Crow 2000; Crow 2007). The high burden of diseases and disorders affecting the brain has been referred to as the ‘Achilles heel’ of human brain size and cognitive adaptations (Crow 2000; Crow 2007). Efforts to understand how human health has been shaped by (and continues to be shaped by) evolution and apply this knowledge to inform medicine is known as *evolutionary medicine.*

One line of research in evolutionary medicine studies the genetic diversity within and between human populations that arises from local adaptation and contributes to variation in modern disease resistance and susceptibility. Such areas of active investigation include the co-evolution of pathogens and human hosts (D'Aeth et al. 2021; Huddleston et al. 2020) and evolutionary mismatch whereby adaptive alleles that evolved in a past environment become deleterious or increase disease risk in an individual’s current environment (Benton et al. 2021; Manus 2018). However, the historical bias of studying genetic variation in individuals of European ancestry has hindered our understanding of these questions. Increasing the number of genomes sequenced from geographically distributed, diverse human populations will enhance knowledge of the extent of genetic diversity that characterizes the variation, evolution, and disease of our species.

A second evolutionary medicine approach leverages comparative phylogenetics and functional studies to delineate the molecular mechanisms underlying phenotypic traits that differ between species to inform understandings of human health. This includes research on compelling adaptations of non-human species a potential to benefit medical advancement, such as longevity in bats (Seim et al. 2013) or reduced cancer risk in elephants (Vazquez and Lynch 2021). Alternatively, one can investigate the biology of *human-specific* adaptive variants to inform human disease susceptibilities. The term *human-specific* in this dissertation refers to the suite of genetic and phenotypic changes that occurred in the *Homo* lineage following divergence with the *Pan* lineage and that are shared among extant humans but may or may not have been common in other extinct *Homo* populations (i.e., Neanderthals, Denisovans, *Homo erectus*). This interdisciplinary thesis utilizes the perspectives of evolutionary medicine to investigate human-specific biology and seeks to integrate concepts and methods from the fields of anthropology and human genetics.

1.2 EVOLUTIONARY DEVELOPMENTAL BIOLOGY OF THE CEREBRAL CORTEX

Development is tightly regulated and constrained via a complex network of gene regulation (Davidson and Erwin 2006). Many of the molecular pathways that mediate development are highly conserved across organisms (i.e., Hox gene regulation of body axis patterning (Lemons and McGinnis 2006)). Despite limited access to primary human tissue, this high level of developmental conservation often allows for the use of model organisms to study human development. Yet not all aspects of developmental programs are conserved. Changes in morphology between species arise from changes in development. This is especially the case for the human cerebral cortex. The focus of this dissertation is on functional investigation of the genetic basis of human-specific features of cortical development and neurodevelopmental disorders using human *in vitro* model systems.

Primates are a fascinating group of mammals. They exhibit complex social organization and communication systems, inhabit diverse ecological niches with elaborate dietary adaptations, and navigate the physical world with distinct locomotor abilities, grasping hands and feet, and binocular vision. Perhaps one of the most scrutinized adaptations of primates as an order is their large brains relative to their body size. Cellular scaling rules that dictate the size of primate brains differ from those that orchestrate rodent brains (Herculano-Houzel et al. 2007; Herculano-Houzel et al. 2006). In rodents, the brain scales *hypermetrically* with neuron number whereby the average neuronal size is larger in a larger rodent brain compared to a smaller rodent brain (Herculano-Houzel et al. 2006). Further, the average size of a non-neuronal cell remains constant with brain size increase and neuronal density decreases as the glia-to-neuron ratio increases (Herculano-Houzel et al. 2006). Conversely, primate brains do not hyperscale as neuron number increases; thus, neuronal size stays constant, and the density of neurons remains stable (Herculano-Houzel et al. 2007). Consequently, primate brains have more neurons as well as higher neuronal densities in all brain structures compared to a rodent brain of equal size (Herculano-Houzel et al. 2007). Given this difference between the primate and rodent orders, research that seeks to investigate the molecular basis of size expansion of the human brain must utilize model systems that obey the same developmental scaling rules.

Compared to other mammalian species except cetaceans, the primate cerebral cortex—the outer layer of the cerebrum composed of grey matter—has expanded disproportionately relative to the rest of the brain (Aboitiz and Montiel 2012; Northcutt and Kaas 1995; Sherwood et al. 2012). For context, the percentage of the total brain that is comprised of cerebral cortex in mammals ranges from 40% in mouse to 80% in humans (Azevedo et al. 2009; Herculano-Houzel 2012; Hofman 1988). Comparative studies of the cerebral cortex across extant primate species have been a major focus of evolutionary inquiry (Figure 1.1) (Herculano-Houzel et al. 2015). Of note, compared to chimpanzees, the human cortex is triple the size with twice as many cells (Mora-Bermúdez et al. 2016).

Cortical expansion is widely considered to be a fundamental evolutionary substrate that facilitated changes in cognition in humans.

Since the split of the last common ancestor of hominins (*Homo*) and chimpanzees/bonobos (*Pan*) roughly 4-6 million years ago, human brain evolution has been characterized by shifts in relative cranial capacity (Carlson et al. 2011; Falk et al. 2000) as well as expansion of specific brain regions implicated in complex cognition (Semendeferi and Damasio 2000). Paleoneurology discoveries from fossil endocasts suggest that the majority of size expansions of the human cerebral cortex occurred in the last three million years (Holloway et al. 2004). Though hotly debated, *ultimate* explanations for cortical expansion have included cultural innovations in tool technologies, dietary shifts and cooking, increases in population sizes, and migration into extreme environments (Ambrose 2001; Stout et al. 2011). While the overall cerebral cortex has expanded in humans, the prefrontal cortex has enlarged *exceptionally* compared to that of other primates (Smaers et al. 2017). In other words, the size of the prefrontal cortex in humans is larger than expected by general allometric growth patterns of primate brains. The prefrontal cortex is of particular interest in human evolution for its association with language, complex decision making, and imagination (Forbes and Grafman 2010; Krienen et al. 2010; Miller 2000; Powell et al. 2010). Exceptional prefrontal expansion also is observed in other great apes, but it is the greatest in humans (Smaers et al. 2017). Small changes in the timing of development often translate to morphological differences in adulthood. Given this deviation from allometric predictions, ape and human prefrontal enlargements reflect derived heterochronic shifts in growth during neural development.

In terms of life history, humans and non-human apes share a prolonged post-natal growth of the brain compared to other primates (Leigh 2004). However, neural development is protracted in humans relative to all other primates, from gestation extending into childhood and adolescence (Finlay and Darlington 1995). Human brain evolution is thus characterized by a special type of heterochrony wherein all stages of brain development are prolonged compared to the ancestral state, a process called *sequential hypermorphosis* (McKinney 2002; McNamara 2002; Vrba 1998). This >20-year developmental window is longer than the lifespan of many non-human primates (Silbereis et al. 2016). The substantially slower development of the human brain, especially of the association cortex, is necessary due to its extraordinary complexity. This longer developmental time course contributes to enhanced plasticity of the human brain involving greater influence of environmental factors on socio-emotional and socio-cognitive development.

Developmental shifts that alter cortical size are therefore coupled with changes in cortical organization (Markov et al. 2013; Rakic 2008) and neural circuitry (Sousa et al. 2017) that further distinguish adult human brains from those of other primates (Geschwind and Rakic 2013; Sherwood et al. 2012). However, human cortical size and architecture are not direct correlates of human cognitive abilities (Sherwood et al. 2012), suggesting the genetic basis of derived cortical expansion can be studied independently of the primary evolutionary targets critical for enhancing cognition. These findings support a model of human brain evolution wherein cortical expansion is a single adaptation of a multifaceted process, of which all components are required for the development of human cognitive and behavioral traits (Geschwind and Rakic 2013; Markov et al. 2013; Rakic 2008; Sherwood et al. 2012; Sousa et al. 2017).

1.3 CELLULAR NEUROBIOLOGY

Neurobiology has demonstrated that the size of the adult brain is determined by the proliferative capacity of neural progenitor cells (NPCs) that generate neurons (Rakic 1988). Neuron number and density provide the substrate for neural connectivity and network complexification, the basis for interspecific differences in cortical areas, cytoarchitecture, and neural circuitry. The *proximate* mechanisms of increased neuronal production that underlie evolutionary cortical expansion include 1) more NPCs with greater proliferative capacity, and 2) lengthening of the neurogenic period resulting in an increase in late-born, or upper layer, neurons (Florio and Huttner 2014; Lewitus et al. 2014; Lui et al. 2011; Rakic 2009; Stepien et al. 2020; Stepien et al. 2021; Taverna et al. 2014). Thus, cortical size differences in mammalian adult brains stem from alterations in the timing and rate of corticogenesis.

Mature neurons that propagate the six-layer mammalian cortex are born from NPCs that divide to produce immature post-mitotic neurons that migrate up the basal processes of NPCs to the cortical plate where they secure their fate (Kriegstein and Alvarez-Buylla 2009). This columnar differentiation lineage that determines neuron number in the adult brain is known as the ‘radial unit’ hypothesis (Rakic 1988). The cerebral cortex develops from two primary germinal zones that are shared among mammals: the ventricular zone (VZ) and the subventricular zone (SVZ) (Angevine et al.

1970; Kornack and Rakic 1998). In primates and other gyrencephalic mammals, the SVZ is further distinguished by an inner and outer region (Dehay et al. 2015; Smart et al. 2002). The outer region is referred to as the outer subventricular zone (OSVZ). Notably, this zone is absent in most rodents, including mice, which is an important consideration when using certain model organisms for studying human neural development (Figure 1.2A). Primates have an expansion of the OSVZ, which is more pronounced in great apes and further enlarged in humans (Fernández et al. 2016). The OSVZ is established later in development than the VZ, and each zone is comprised of morphologically and proliferatively distinct NPCs (Betizeau et al. 2013; Lui et al. 2011; Taverna et al. 2014).

The developing VZ and SVZ harbor the cell bodies of two major classes of NPCs, often classified as apical progenitors and basal progenitors (Borrell and Reillo 2012; Gotz and Huttner 2005; Lui et al. 2011; Taverna et al. 2014). Apical progenitors, including neuroepithelial cells, ventricular radial glia (vRG), and intermediate progenitors (IPs), are labeled as such given their contact with the ventricular zone, division at the apical surface, and maintenance of apical cellular polarity (Taverna et al. 2014). Conversely, basal progenitors (including outer radial glia (oRG) and outer IPs (oIPs)) are named for their contact with the basal surface, lack of ventricular contact, and shift in cell polarity towards the basal lamina (Taverna et al. 2014). vRG cells divide symmetrically to generate two equal daughter cells allowing for self-renewal that builds the VZ proliferative pool. Over time vRG switch from symmetric to asymmetric division via a diagonal shift in the orientation of the spindle relative to the apical-basal axis of cell polarity, producing one vRG that maintains proliferative capacity and one progeny that will differentiate into an IP, immature neuron, or oRG (Figure 1.2B). In lissencephalic (unfolded cortex) species, such as the mouse, over 90% of oIPs are neurogenic, i.e., divide once to give rise to two post-mitotic neurons, whereas oIPs of gyrencepahlic species undergo symmetric proliferative divisions prior to neurogenic divisions (Attardo et al. 2008; Betizeau et al. 2013; Hansen et al. 2010; Haubensak et al. 2004; Lui et al. 2011). Similarly, oRG represent less than 10% of basal progenitors in mouse and are mostly neurogenic (Shitamukai et al. 2011). By contrast, oRG constitute more than half of the basal progenitors in primates and exhibit increased proliferative capacity, especially human oRG (Betizeau et al. 2013; Fietz et al. 2010; Hansen et al. 2010; Reillo et al. 2011). Changes in the abundance and duration of symmetric and asymmetric divisions therefore determines the intricate balance of proliferation versus differentiation that builds the brain (Rakic 1995), and the evolutionary expansion of the cortex is primarily attributed to augmented proliferative capacity of basal progenitors (LaMonica et al. 2012; Namba and Huttner 2017).

In humans, both the VZ and OSVZ proliferative window are substantially protracted relative to other species, contributing to both a lateral and radial expansion of the cortex (Otani et al. 2016). OSVZ expansion is proposed to have played a prominent evolutionary role in ‘scaling a phylogenetically ancestral primate brain to the complexity of the human brain’ (Dehay et al. 2015). Unlike in other species such as mice where the majority neurons are generated by VZ-NPCs, the OSVZ-NPCs produce most neurons in primates (Betizeau et al. 2013; Dehay et al. 2015).

Comparison of NPCs between human, chimpanzee, and macaque determined that human NPCs protract the duration of short cell cycle proliferation (Mora-Bermúdez et al. 2016; Otani et al. 2016; Sousa et al. 2017). Human apical progenitors have a similar cell cycle length to chimpanzees (46.5 hours versus to 43.8 hours, respectively), but there is a significant 5-hour increase of S-phase length in human apical progenitors (Mora-Bermúdez et al. 2016). Further, human apical progenitor cells have a lengthened prometaphase-metaphase compared to chimpanzee apical progenitors which is not observed in non-neural cells (Mora-Bermúdez et al. 2016). While the differences in human and chimpanzee NPC cell cycle kinetics are relatively small, they nevertheless provide insight into the cumulative effect of mitotic changes during neural development acting to prolong both proliferative and neurogenic phases and subsequently translate into morphological differences.

Structural reorganization of the connectome is predicted to have occurred with the allometric expansion of the human cerebral cortex (Sousa et al. 2017). While there are no areas of the cortex that are unique to humans, there are notable human-specific features with respect to neuronal organization and connectivity. First, humans have larger excitatory projection neurons, referred to as *pyramidal* neurons, with increased spine density and dendritic arborization compared to other primates which may facilitate greater integrative connectivity (Elston et al. 2011). This may be linked to the maturation of pyramidal neurons in the cortex, which is prolonged in great apes compared to macaques (Cupp and Uemura 1980; Petanjek et al. 2011; Sedmak et al. 2018), and is further protracted in humans (Bianchi et al. 2013; Teffer et al. 2013). In addition, the human cortex, primarily the fronto-insular and anterior cingulate cortices, also harbors larger and more von Economo neurons than other primates (Allman et al. 2011). Interestingly, density of von Economo neurons has been affected in neuropsychiatric disorders (Brüne et al. 2010).

Upper cortical layers 2 and 3 are thicker with more neurons in the human cortex compared to other non-human primates (Hutsler et al. 2005; Marin-Padilla 1978). Consequently, humans show an increase in cortical short-range projections that develop from pyramidal neurons of layers 2 and 3, which connect nearby regions and are positively correlated to the level of gyrification, consistent with the enhanced gyrification observed in the human cortex (Catani et al. 2012; Hofman 2012; van den Heuvel et al. 2016). Unlike chimpanzee primary and association cortices that exhibit similar levels of neuropil—used as a proxy for measuring connectivity given its composition of dendrites, axons, synapses, and glia cell processes—human association areas, particularly in the prefrontal cortex, have a higher proportion of neuropil compared to human primary areas (Spocter et al. 2012). Neurons in human frontal and temporal cortices are also more spread out compared to that of other primates (Schenker et al. 2008; Semendeferi et al. 2011).

Comparative studies have also identified differences among primates in long-distance projection systems. Humans have increased temporal projections of perisylvian axonal tracts of the arcuate fasciculus compared to chimpanzees and which are absent in macaques, a feature that has implications for language evolution (Rilling et al. 2008). Additionally, the primary white matter tract that connects the lateral frontal and parietal areas of the cortex, known as the *superior longitudinal fasciculus*, shows differential organization in humans compared to chimpanzees which may underlie differences in spatial attention (Hecht et al. 2015).

Lastly, cortical myelination maps are comparable between humans and other primates; however, humans have unique timing and pattern with a larger total axon surface in the brain coupled with less overall myelination (Glasser et al. 2014; Miller et al. 2012). These features have functional implications for conduction velocity. Humans exhibit reorganized long-range corticopetal, intracortical, and corticofugal projections compared to other primates, especially of the prefrontal and temporal association cortices (Sousa et al. 2017). Together, changes in local circuitry and long-range projections accompanied cortical expansion, perhaps independently, during human brain evolution.

1.4 MODELING HUMAN-SPECIFIC NEURAL DEVELOPMENT

Research using primary post-mortem adult and developing brain tissue from humans and non-human primates has advanced our knowledge about developmental differences between primates, such as several of the discoveries discussed above. However, access and availability to this tissue is extremely limited and the potential for hypothesis testing is constrained. To circumvent this hurdle, recent advances in somatic cell reprogramming have allowed for the generation of cell culture models that enable researchers to mimic key aspects of organogenesis utilizing cells with a human genetic background.

Pluripotent stem cells can either be embryonically obtained (embryonic stem cells (ESCs)) or generated from chemically inducing reprogrammable adult cells, such as fibroblasts, to a pluripotent stem cell-like state (induced pluripotent stem cells (iPSCs)) (Chin et al. 2009; Zhou and Ding 2010). ESCs and iPSCs can then be differentiated in culture to various cell fates (Park et al. 2008). Recent advancements in stem cell technology allow for the differentiation of ESCs/iPSCs into 3D structures called ‘organoids’ that recapitulate *in vivo* organs, dependent on a standardized combination of nutrients, growth factors, inhibitors, and other small molecules to specify organogenesis (Figure 1.3) (Yin et al. 2016); (Lancaster et al. 2013). Significant insights into human brain development have come from cortical organoid models, especially when the genetic, molecular, and/or cellular features of inquiry are not conserved or phenocopied in rodent models (Kanton et al. 2019; Li et al. 2017; Pollen et al. 2019). Cortical organoids serve as the primary model system of human neural development and pathogenesis used in this dissertation.

1.5 RESEARCH QUESTION AND CHAPTER INTRODUCTIONS

To understand the evolution of human-specific traits and their relationship to modern disease susceptibilities, we must first identify the genetic and molecular changes underlying phenotypic differences between humans and our closest living relatives. Modification of conserved developmental programs that govern morphology occurs via two principal mechanisms: direct changes to genes and gene function, or changes in gene regulation (transcription, splicing, translation). The overarching question that drives the research projects of this dissertation is: *What is the genetic basis of human-specific features of neural development and neurodevelopmental disorders?*

CHAPTER 2. Selection mapping can help pinpoint candidate alleles in the human genome underling adaptive traits that are either population-specific due to local adaptation (microevolution) or fixed on the species level (macroevolution). Chapter 2 reviews human population genetics and comparative phylogenetic methods for genetic variant discovery and notable findings that have propelled the investigation of human adaptation, with an emphasis on the current landscape of the genetics of human-specific features of neural development and cortical expansion. Several genes associated with neural development have been identified that show accelerated evolution, divergent expression, and/or duplicated sequence across primates and specific to humans (Boyd et al. 2015; Burki and Kaessmann 2004; Dumas et al. 2012; Enard et al. 2002; Fiddes et al. 2018; Florio et al. 2015; Kamm et al. 2013; Keeney et al. 2014; Montgomery et al. 2011; Pollard et al. 2006; Zimmer and Montgomery 2015).

While major advances have been made in the study of human cortical expansion, the specific adaptive targets that were favored by natural selection and the causal genetic changes that promote these features are not fully characterized. Two main approaches have been taken to address how NPC number, particularly basal progenitors, expands during human neural development. The first is deciphering the functions of human-specific genes or genetic variants with preferential expression in human NPCs that are prime candidates for human cortical expansion (**Chapter 3**) (Fiddes et al. 2018; Kronenberg et al. 2018; Suzuki et al. 2018; Xing et al. 2021). The second is investigating the etiology of cortical malformations such as *microcephaly* (**Chapters 4 & 5**), which is a smaller than expected brain size for an individual’s age/sex mean and often accompanied by intellectual disability (National Birth Defects Prevention Network, nbdpn.org). Roughly 18 genes have been identified to cause primary microcephaly when disrupted due to their role in cortical growth (Jayaraman et al. 2018). However, with rising numbers of exome and genome sequencing of clinical cohorts, variant discovery has implicated a growing list of genes in neurodevelopmental disorders with microcephaly.

CHAPTER 3. Many human-specific structural variants, primarily gene duplicates, that arose from segmental duplications are candidates for human-specific adaptions in brain development. While these structural were likely adaptive during human evolution, the tradeoff of increased repetitive sequence has predisposed the human genome to detrimental rearrangements at these loci that cause neuropsychiatric diseases and developmental disorders. Fifteen human-specific genes show preferential expression in cortical NPCs (Florio et al. 2018). Functional work has only implicated two of these duplicate genes in cortical evolution: *ARHGAP11B* and *NOTCH2NL* (Fiddes et al. 2018; Kronenberg et al. 2018; Suzuki et al. 2018). However, many others are promising candidates, such as the neuroblastoma breakpoint gene family (*NBPF*). *NBPF* dosage has been implicated in primate cortical expansion, autism, and schizophrenia (Davis et al. 2019; Davis et al. 2014; Dumas et al. 2012; Keeney et al. 2014; Quick et al. 2016). *NBPF* sequence has undergone significant copy number expansion in the human genome, leading to an accumulation of the encoded DUF1220 (Olduvai) protein domains (O'Bleness et al. 2014; O'Bleness et al. 2012). Yet, the molecular functions of *NBPF* proteins and DUF1220 domains remain unknown. Chapter 3 investigates the function of *NBPF*s in cortical development using a human and chimpanzee comparative cell culture approach of iPSC-derived NPCs and organoid model systems.

CHAPTERS 4 & 5: Studying the disruption of conserved developmental programs due to pathogenic variants in genes or regulatory elements that cause human NND can shed light to human-specific features. This often occurs when *in vivo* model organisms fail to phenocopy human disease. Delineating evolutionary differences between functional models can help pinpoint shared and derived mechanisms of pathogenicity, and potentially uncover human-specific expressivity of disease phenotypes. **Chapter 4** identifies and utilizes novel pathogenic biallelic variants in the gene *CSMD1* to investigate previously uncharacterized functions of complement pathway regulation in human neural development. Modeling loss of CSMD1 in human cortical organoid identifies proliferation defects, premature differentiation, and disrupted cytomorphology—pathogenic mechanisms consistent with NND clinical manifestations in individuals with *CSMD1* biallelic variants and defects not phenocopied by *Csmd1* knockout mouse. **Chapter 5** investigates and proposes a loss-of-function genetic mechanism of the autosomal recessive *THOC6* intellectual disability syndrome (TIDS). We generate the first *Thoc6* knockout mouse and demonstrate embryonic lethality of *Thoc6* loss, which differs from human TIDS wherein affected individuals survive into adulthood. In addition to NND clinical presentations (intellectual disability and microcephaly), affected individuals present with multi-organ abnormalities including distinctive facial dysmorphology, renal and cardiac anomalies, and gonadal dysfunction, raising the question of why the effects of pathogenic variants are restricted to, or amplified in, specific cell types and tissues in affected individuals. We investigate THOC6-dependent TREX functions in neural tissue and identify mRNA processing/splicing defects that contribute to cellular dependencies and species differences in TIDS pathology.

Together, these projects provide a better understanding of the genetic contributions of human cortical evolution that allow species-specific differences in cortical development to be defined. This improves the utility of mammalian models to make translationally important discoveries in NND.

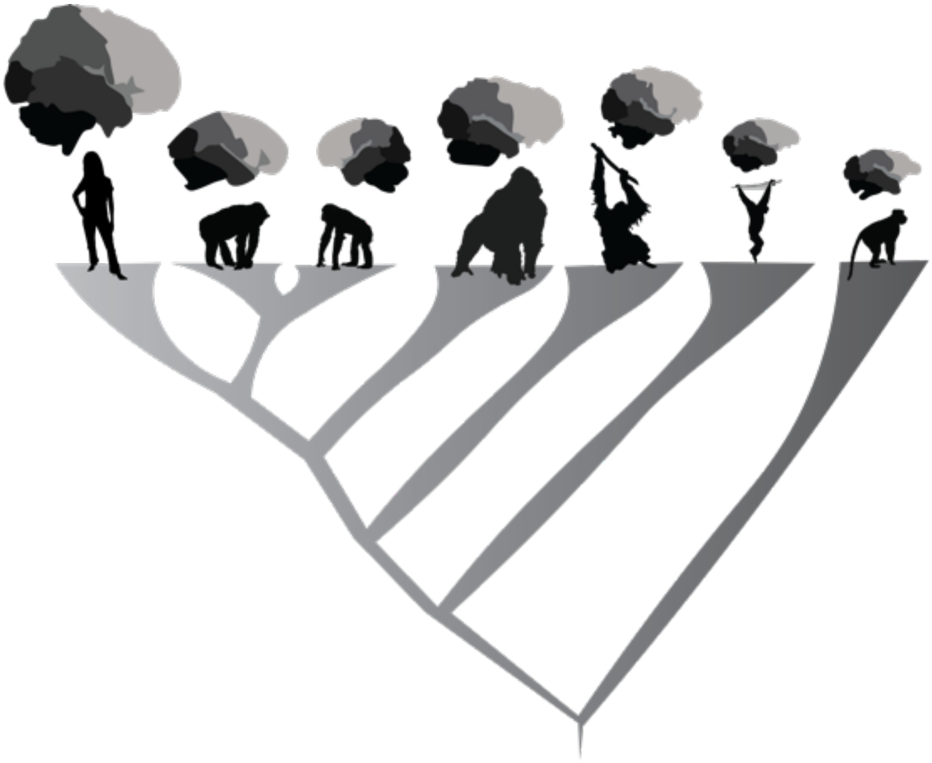


Figure 1.1. Primate phylogeny showing cortical organization and size differences.

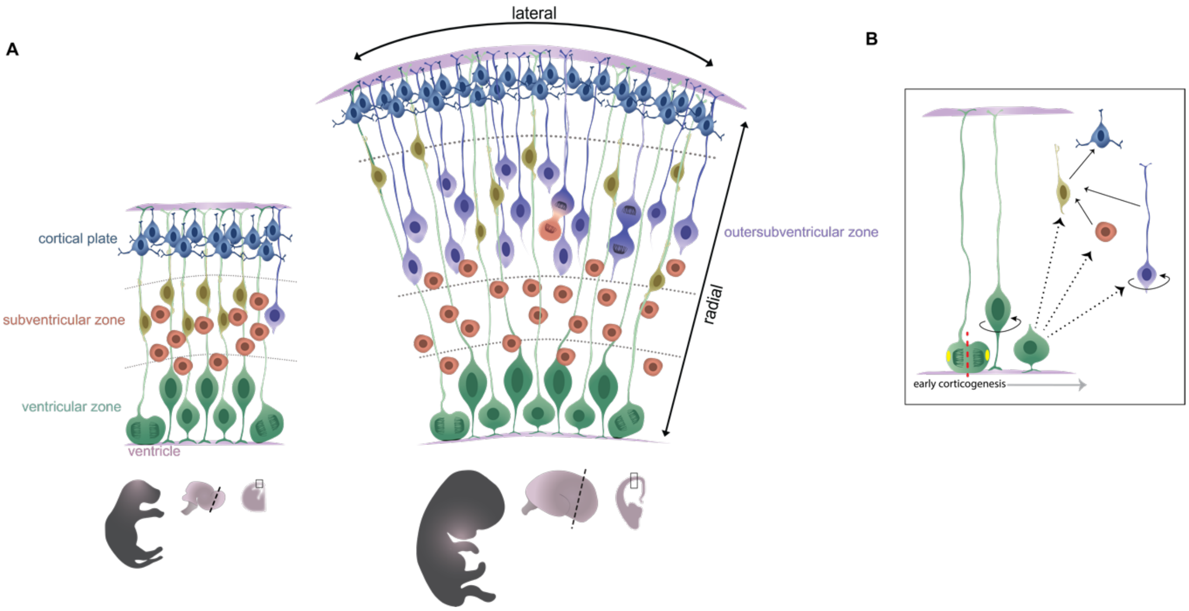


Figure 1.2. Schematic of mammalian corticogenesis.
(A) mouse versus human cortical development. (B) proliferative and neurogenic capacities of vRG.

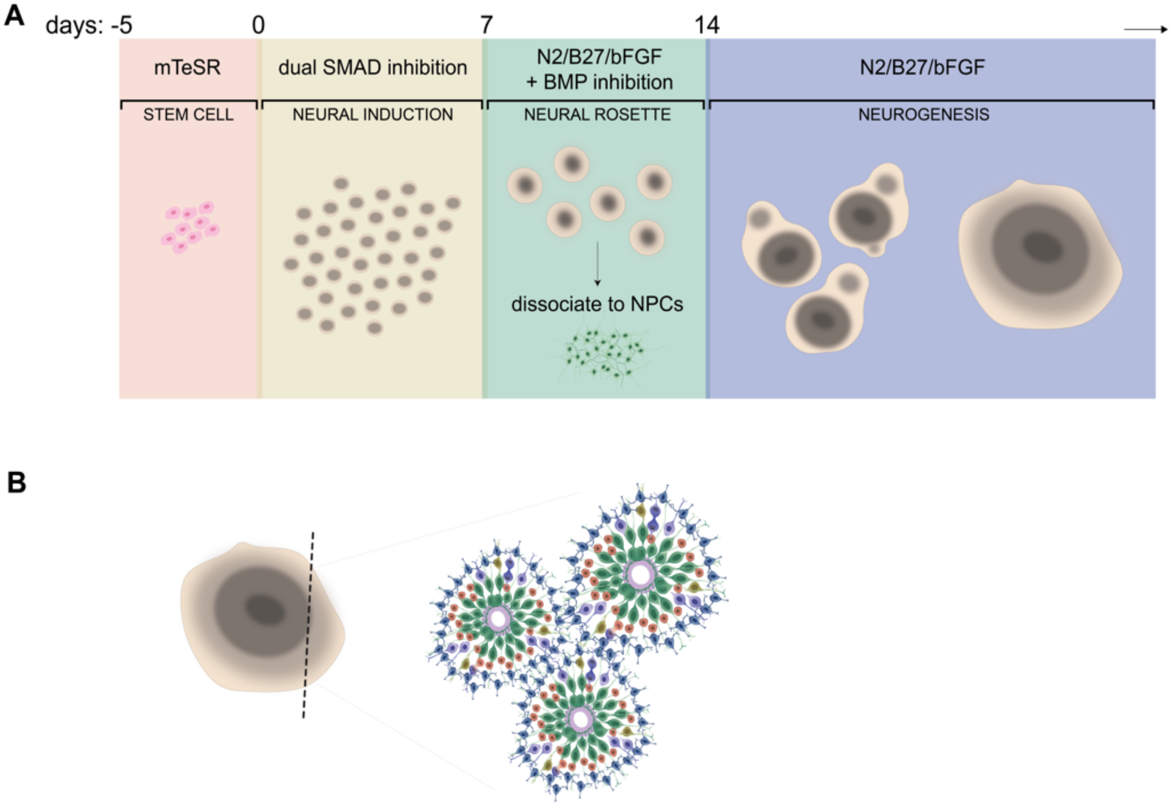


Figure 1.3. Cerebral organoid differentiation.
(A) differentiation protocol beginning with confluent PSCs that undergo neural induction followed by neural differentiation. (B) Cartoon of cortical organoid cross section showing neural rosettes—structures that mimic the developing ventricular zone and subsequent radial unit differentiation from NPC to neurons.

REFERENCES

Aboitiz F, Montiel JF (2012) From tetrapods to primates: conserved developmental mechanisms in diverging ecological adaptations. Prog Brain Res 195: 3-24. doi: 10.1016/B978-0-444-53860-4.00001-5

Allman JM, Tetreault NA, Hakeem AY, Manaye KF, Semendeferi K, Erwin JM, Park S, Goubert V, Hof PR (2011) The von Economo neurons in the frontoinsular and anterior cingulate cortex. Ann N Y Acad Sci 1225: 59-71. doi: 10.1111/j.1749-6632.2011.06011.x

Ambrose SH (2001) Paleolithic technology and human evolution. Science 291: 1748-53. doi: 10.1126/science.1059487

Angevine JB, Bodian D, Coulomb AJ, Edds Jr. MV, Hamburger V, Jacobson M, Lyser KM, Prestige MC, Sidman RL, Varon S, Weiss PA (1970) Embryonic vertebrate central nervous system: revised terminology. The Boulder Committee. Anat Rec 166: 257-61. doi: 10.1002/ar.1091660214

Attardo A, Calegari F, Haubensak W, Wilsch-Bräuninger M, Huttner WB (2008) Live imaging at the onset of cortical neurogenesis reveals differential appearance of the neuronal phenotype in apical versus basal progenitor progeny. PLoS One 3: e2388. doi: 10.1371/journal.pone.0002388

Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, Leite RE, Jacob Filho W, Lent R, Herculano-Houzel S (2009) Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J Comp Neurol 513: 532-41. doi: 10.1002/cne.21974

Benton ML, Abraham A, LaBella AL, Abbot P, Rokas A, Capra JA (2021) The influence of evolutionary history on human health and disease. Nat Rev Genet 22: 269-283. doi: 10.1038/s41576-020-00305-9

Betizeau M, Cortay V, Patti D, Pfister S, Gautier E, Bellemin-Ménard A, Afanassieff M, Huissoud C, Douglas RJ, Kennedy H, Dehay C (2013) Precursor diversity and complexity of lineage relationships in the outer subventricular zone of the primate. Neuron 80: 442-57. doi: 10.1016/j.neuron.2013.09.032

Bianchi S, Stimpson CD, Duka T, Larsen MD, Janssen WG, Collins Z, Bauernfeind AL, Schapiro SJ, Baze WB, McArthur MJ, Hopkins WD, Wildman DE, Lipovich L, Kuzawa CW, Jacobs B, Hof PR, Sherwood CC (2013) Synaptogenesis and development of pyramidal neuron dendritic morphology in the chimpanzee neocortex resembles humans. Proc Natl Acad Sci U S A 110 Suppl 2: 10395-401. doi: 10.1073/pnas.1301224110

Borrell V, Reillo I (2012) Emerging roles of neural stem cells in cerebral cortex development and evolution. Dev Neurobiol 72: 955-71. doi: 10.1002/dneu.22013

Boyd JL, Skove SL, Rouanet JP, Pilaz LJ, Bepler T, Gordân R, Wray GA, Silver DL (2015) Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. Curr Biol 25: 772-779. doi: 10.1016/j.cub.2015.01.041

Brüne M, Schöbel A, Karau R, Benali A, Faustmann PM, Juckel G, Petrasch-Parwez E (2010) Von Economo neuron density in the anterior cingulate cortex is reduced in early onset schizophrenia. Acta Neuropathol 119: 771-8. doi: 10.1007/s00401-010-0673-2

Burki F, Kaessmann H (2004) Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. Nat Genet 36: 1061-3. doi: 10.1038/ng1431

Carlson KJ, Stout D, Jashashvili T, de Ruiter DJ, Tafforeau P, Carlson K, Berger LR (2011) The endocast of MH1, Australopithecus sediba. Science 333: 1402-7. doi: 10.1126/science.1203922

Catani M, Dell'acqua F, Vergani F, Malik F, Hodge H, Roy P, Valabregue R, Thiebaut de Schotten M (2012) Short frontal lobe connections of the human brain. Cortex 48: 273-91. doi: 10.1016/j.cortex.2011.12.001

Chin MH, Mason MJ, Xie W, Volinia S, Singer M, Peterson C, Ambartsumyan G, Aimiuwu O, Richter L, Zhang J, Khvorostov I, Ott V, Grunstein M, Lavon N, Benvenisty N, Croce CM, Clark AT, Baxter T, Pyle AD, Teitell MA, Pelegrini M, Plath K, Lowry WE (2009) Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. Cell Stem Cell 5: 111-23. doi: 10.1016/j.stem.2009.06.008

Crow TJ (2000) Schizophrenia as the price that homo sapiens pays for language: a resolution of the central paradox in the origin of the species. Brain Res Brain Res Rev 31: 118-29. doi: 10.1016/s0165-0173(99)00029-6

Crow TJ (2007) Nuclear schizophrenic symptoms as the key to the evolution of the human brain. Evolution of Nervous Systems 549-567.

Cupp CJ, Uemura E (1980) Age-related changes in prefrontal cortex of Macaca mulatta: quantitative analysis of dendritic branching patterns. Exp Neurol 69: 143-63. doi: 10.1016/0014-4886(80)90150-8

D'Aeth JC, van der Linden MP, McGee L, de Lencastre H, Turner P, Song JH, Lo SW, Gladstone RA, Sá-Leão R, Ko KS, Hanage WP, Breiman RF, Beall B, Bentley SD, Croucher NJ, Consortium G (2021) The role of interspecies recombination in the evolution of antibiotic-resistant pneumococci. Elife 10. doi: 10.7554/eLife.67113

Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. Science 311: 796-800. doi: 10.1126/science.1113832

Davis JM, Heft I, Scherer SW, Sikela JM (2019) A Third Linear Association Between Olduvai (DUF1220) Copy Number and Severity of the Classic Symptoms of Inherited Autism. Am J Psychiatry 176: 643-650. doi: 10.1176/appi.ajp.2018.18080993

Davis JM, Searles VB, Anderson N, Keeney J, Dumas L, Sikela JM (2014) DUF1220 dosage is linearly associated with increasing severity of the three primary symptoms of autism. PLoS Genet 10: e1004241. doi: 10.1371/journal.pgen.1004241

Dehay C, Kennedy H, Kosik KS (2015) The outer subventricular zone and primate-specific cortical complexification. Neuron 85: 683-94. doi: 10.1016/j.neuron.2014.12.060

Dennis MY, Eichler EE (2016) Human adaptation and evolution by segmental duplication. Curr Opin Genet Dev 41: 44-52. doi: 10.1016/j.gde.2016.08.001

Dennis MY, Harshman L, Nelson BJ, 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 HAF, Nuttle X, Hoekzema K, Lindsay-Graves TA, Wilson RK, Eichler EE (2017) The evolution and population diversity of human-specific segmental duplications. Nat Ecol Evol 1: 69. doi: 10.1038/s41559-016-0069

Dumas LJ, O'Bleness MS, Davis JM, Dickens CM, Anderson N, Keeney JG, Jackson J, Sikela M, Raznahan A, Giedd J, Rapoport J, Nagamani SS, Erez A, Brunetti-Pierri N, Sugalski R, Lupski JR, Fingerlin T, Cheung SW, Sikela JM (2012) DUF1220-domain copy number implicated in human brain-size pathology and evolution. Am J Hum Genet 91: 444-54. doi: 10.1016/j.ajhg.2012.07.016

Elston GN, Benavides-Piccione R, Elston A, Manger PR, Defelipe J (2011) Pyramidal cells in prefrontal cortex of primates: marked differences in neuronal structure among species. Front Neuroanat 5: 2. doi: 10.3389/fnana.2011.00002

Enard W, Khaitovich P, Klose J, Zöllner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Pääbo S (2002) Intra- and interspecific variation in primate gene expression patterns. Science 296: 340-3. doi: 10.1126/science.1068996

Falk D, Redmond JC, Guyer J, Conroy C, Recheis W, Weber GW, Seidler H (2000) Early hominid brain evolution: a new look at old endocasts. J Hum Evol 38: 695-717. doi: 10.1006/jhev.1999.0378

Fernández V, Llinares-Benadero C, Borrell V (2016) Cerebral cortex expansion and folding: what have we learned? EMBO J 35: 1021-44. doi: 10.15252/embj.201593701

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, Lorig-Roach R, Field AR, Haeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor MC, Zwolinski S, Katzman S, Kriegstein A, Eichler EE, Salama SR, Jacobs FMJ, Haussler D (2018) Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell 173: 1356-1369.e22. doi: 10.1016/j.cell.2018.03.051

Fietz SA, Kelava I, Vogt J, Wilsch-Bräuninger M, Stenzel D, Fish JL, Corbeil D, Riehn A, Distler W, Nitsch R, Huttner WB (2010) OSVZ progenitors of human and ferret neocortex are epithelial-like and expand by integrin signaling. Nat Neurosci 13: 690-9. doi: 10.1038/nn.2553

Finch CE (2010) Evolution in health and medicine Sackler colloquium: Evolution of the human lifespan and diseases of aging: roles of infection, inflammation, and nutrition. Proc Natl Acad Sci U S A 107 Suppl 1: 1718-24. doi: 10.1073/pnas.0909606106

Finlay BL, Darlington RB (1995) Linked regularities in the development and evolution of mammalian brains. Science 268: 1578-84.

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, Guhr E, Klemroth S, Prüfer K, Kelso J, Naumann R, Nüsslein I, Dahl A, Lachmann R, Pääbo S, Huttner WB (2015) Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science 347: 1465-70. doi: 10.1126/science.aaa1975

Florio M, Heide M, Pinson A, Brandl H, Albert M, Winkler S, Wimberger P, Huttner WB, Hiller M (2018) Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. Elife 7. doi: 10.7554/eLife.32332

Florio M, Huttner WB (2014) Neural progenitors, neurogenesis and the evolution of the neocortex. Development 141: 2182-94. doi: 10.1242/dev.090571

Forbes CE, Grafman J (2010) The role of the human prefrontal cortex in social cognition and moral judgment. Annu Rev Neurosci 33: 299-324. doi: 10.1146/annurev-neuro-060909-153230

Geschwind DH, Rakic P (2013) Cortical evolution: judge the brain by its cover. Neuron 80: 633-47. doi: 10.1016/j.neuron.2013.10.045

Glasser MF, Goyal MS, Preuss TM, Raichle ME, Van Essen DC (2014) Trends and properties of human cerebral cortex: correlations with cortical myelin content. Neuroimage 93 Pt 2: 165-75. doi: 10.1016/j.neuroimage.2013.03.060

Gotz M, Huttner WB (2005) The cell biology of neurogenesis. Nat Rev Mol Cell Biol 6: 777-88.

Hansen DV, Lui JH, Parker PR, Kriegstein AR (2010) Neurogenic radial glia in the outer subventricular zone of human neocortex. Nature 464: 554-561. doi: nature08845 [pii]

10.1038/nature08845

Haubensak W, Attardo A, Denk W, Huttner WB (2004) Neurons arise in the basal neuroepithelium of the early mammalian telencephalon: a major site of neurogenesis. Proc Natl Acad Sci U S A 101: 3196-201. doi: 10.1073/pnas.0308600100

Hecht EE, Gutman DA, Bradley BA, Preuss TM, Stout D (2015) Virtual dissection and comparative connectivity of the superior longitudinal fasciculus in chimpanzees and humans. Neuroimage 108: 124-37. doi: 10.1016/j.neuroimage.2014.12.039

Herculano-Houzel S (2012) Neuronal scaling rules for primate brains: the primate advantage. Prog Brain Res 195: 325-40. doi: 10.1016/B978-0-444-53860-4.00015-5

Herculano-Houzel S, Catania K, Manger PR, Kaas JH (2015) Mammalian Brains Are Made of These: A Dataset of the Numbers and Densities of Neuronal and Nonneuronal Cells in the Brain of Glires, Primates, Scandentia, Eulipotyphlans, Afrotherians and Artiodactyls, and Their Relationship with Body Mass. Brain Behav Evol 86: 145-63. doi: 10.1159/000437413

Herculano-Houzel S, Collins CE, Wong P, Kaas JH (2007) Cellular scaling rules for primate brains. Proc Natl Acad Sci U S A 104: 3562-7. doi: 10.1073/pnas.0611396104

Herculano-Houzel S, Mota B, Lent R (2006) Cellular scaling rules for rodent brains. Proc Natl Acad Sci U S A 103: 12138-43. doi: 10.1073/pnas.0604911103

Hofman MA (1988) Size and shape of the cerebral cortex in mammals. II. The cortical volume. Brain Behav Evol 32: 17-26. doi: 10.1159/000116529

Hofman MA (2012) Design principles of the human brain: an evolutionary perspective. Prog Brain Res 195: 373-90. doi: 10.1016/B978-0-444-53860-4.00018-0

Holloway R, Broadfield D, Yuan M (2004) THE HUMAN FOSSIL RECORD, Volume Three: Brain Endocasts-The Paleoneurological Evidence. vol 3. John Wiley & Sons, Hoboken, NJ

Huddleston J, Barnes JR, Rowe T, Xu X, Kondor R, Wentworth DE, Whittaker L, Ermetal B, Daniels RS, McCauley JW, Fujisaki S, Nakamura K, Kishida N, Watanabe S, Hasegawa H, Barr I, Subbarao K, Barrat-Charlaix P, Neher RA, Bedford T (2020) Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza A/H3N2 evolution. Elife 9. doi: 10.7554/eLife.60067

Hutsler JJ, Lee DG, Porter KK (2005) Comparative analysis of cortical layering and supragranular layer enlargement in rodent carnivore and primate species. Brain Res 1052: 71-81. doi: 10.1016/j.brainres.2005.06.015

Jayaraman D, Bae BI, Walsh CA (2018) The Genetics of Primary Microcephaly. Annu Rev Genomics Hum Genet 19: 177-200. doi: 10.1146/annurev-genom-083117-021441

Kamm GB, López-Leal R, Lorenzo JR, Franchini LF (2013) A fast-evolving human NPAS3 enhancer gained reporter expression in the developing forebrain of transgenic mice. Philos Trans R Soc Lond B Biol Sci 368: 20130019. doi: 10.1098/rstb.2013.0019

Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, Guijarro P, Sidow L, Fleck JS, Han D, Qian Z, Heide M, Huttner WB, Khaitovich P, Pääbo S, Treutlein B, Camp JG (2019) Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature 574: 418-422. doi: 10.1038/s41586-019-1654-9

Keeney JG, Dumas L, Sikela JM (2014) The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. Front Hum Neurosci 8: 427. doi: 10.3389/fnhum.2014.00427

Kety SS, Schmidt CF (1948) THE NITROUS OXIDE METHOD FOR THE QUANTITATIVE DETERMINATION OF CEREBRAL BLOOD FLOW IN MAN: THEORY, PROCEDURE AND NORMAL VALUES. J Clin Invest 27: 476-83. doi: 10.1172/JCI101994

Khaitovich P, Lockstone HE, Wayland MT, Tsang TM, Jayatilaka SD, Guo AJ, Zhou J, Somel M, Harris LW, Holmes E, Pääbo S, Bahn S (2008) Metabolic changes in schizophrenia and human brain evolution. Genome Biol 9: R124. doi: 10.1186/gb-2008-9-8-r124

Kornack DR, Rakic P (1998) Changes in cell-cycle kinetics during the development and evolution of primate neocortex. Proc Natl Acad Sci U S A 95: 1242-6.

Kriegstein A, Alvarez-Buylla A (2009) The glial nature of embryonic and adult neural stem cells. Annu Rev Neurosci 32: 149-84. doi: 10.1146/annurev.neuro.051508.135600

Krienen FM, Tu PC, Buckner RL (2010) Clan mentality: evidence that the medial prefrontal cortex responds to close others. J Neurosci 30: 13906-15. doi: 10.1523/JNEUROSCI.2180-10.2010

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, Munson KM, Hastie AR, Diekhans M, Hormozdiari F, Lorusso N, Hoekzema K, Qiu R, Clark K, Raja A, Welch AE, Sorensen M, Baker C, Fulton RS, Armstrong J, Graves-Lindsay TA, Denli AM, Hoppe ER, Hsieh P, Hill CM, Pang AWC, Lee J, Lam ET, Dutcher SK, Gage FH, Warren WC, Shendure J, Haussler D, Schneider VA, Cao H, Ventura M, Wilson RK, Paten B, Pollen A, Eichler EE (2018) High-resolution comparative analysis of great ape genomes. Science 360. doi: 10.1126/science.aar6343

LaMonica BE, Lui JH, Wang X, Kriegstein AR (2012) OSVZ progenitors in the human cortex: an updated perspective on neurodevelopmental disease. Curr Opin Neurobiol 22: 747-53. doi: 10.1016/j.conb.2012.03.006

Lancaster MA, Renner M, Martin CA, Wenzel D, Bicknell LS, Hurles ME, Homfray T, Penninger JM, Jackson AP, Knoblich JA (2013) Cerebral organoids model human brain development and microcephaly. Nature 501: 373-9. doi: 10.1038/nature12517

Leigh SR (2004) Brain growth, life history, and cognition in primate and human evolution. Am J Primatol 62: 139-64. doi: 10.1002/ajp.20012

Lemons D, McGinnis W (2006) Genomic evolution of Hox gene clusters. Science 313: 1918-22. doi: 10.1126/science.1132040

Lewitus E, Kelava I, Kalinka AT, Tomancak P, Huttner WB (2014) An adaptive threshold in mammalian neocortical evolution. PLoS Biol 12: e1002000. doi: 10.1371/journal.pbio.1002000

Li Y, Muffat J, Omer A, Bosch I, Lancaster MA, Sur M, Gehrke L, Knoblich JA, Jaenisch R (2017) Induction of Expansion and Folding in Human Cerebral Organoids. Cell Stem Cell 20: 385-396.e3. doi: 10.1016/j.stem.2016.11.017

Lui JH, Hansen DV, Kriegstein AR (2011) Development and evolution of the human neocortex. Cell 146: 18-36. doi: 10.1016/j.cell.2011.06.030

Manus MB (2018) Evolutionary mismatch. Evol Med Public Health 2018: 190-191. doi: 10.1093/emph/eoy023

Marin-Padilla M (1978) Dual origin of the mammalian neocortex and evolution of the cortical plate. Anat Embryol (Berl) 152: 109-26. doi: 10.1007/BF00315920

Markov NT, Ercsey-Ravasz M, Van Essen DC, Knoblauch K, Toroczkai Z, Kennedy H (2013) Cortical high-density counterstream architectures. Science 342: 1238406. doi: 10.1126/science.1238406

Marques-Bonet T, Girirajan S, Eichler EE (2009) The origins and impact of primate segmental duplications. Trends Genet 25: 443-54. doi: 10.1016/j.tig.2009.08.002

McKinney M (2002) Brain evolution by stretching the global mitotic clock of development. *Human Evolution Through Developmental Change*. The Johns Hopkins University Press, Baltimore, MD

McNamara K (2002) Sequential hypermorphosis. Stretching ontogeny to the limit. *Human Evolution Through Developmental Change*. Johns Hopkins University Press

Miller DJ, Duka T, Stimpson CD, Schapiro SJ, Baze WB, McArthur MJ, Fobbs AJ, Sousa AM, Sestan N, Wildman DE, Lipovich L, Kuzawa CW, Hof PR, Sherwood CC (2012) Prolonged myelination in human neocortical evolution. Proc Natl Acad Sci U S A 109: 16480-5. doi: 10.1073/pnas.1117943109

Miller EK (2000) The prefrontal cortex and cognitive control. Nat Rev Neurosci 1: 59-65. doi: 10.1038/35036228

Montgomery SH, Capellini I, Venditti C, Barton RA, Mundy NI (2011) Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. Mol Biol Evol 28: 625-38. doi: 10.1093/molbev/msq237

Mora-Bermúdez F, Badsha F, Kanton S, Camp JG, Vernot B, Köhler K, Voigt B, Okita K, Maricic T, He Z, Lachmann R, Pääbo S, Treutlein B, Huttner WB (2016) Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife 5. doi: 10.7554/eLife.18683

Namba T, Huttner WB (2017) Neural progenitor cells and their role in the development and evolutionary expansion of the neocortex. Wiley Interdiscip Rev Dev Biol 6. doi: 10.1002/wdev.256

Northcutt RG, Kaas JH (1995) The emergence and evolution of mammalian neocortex. Trends Neurosci 18: 373-9. doi: 10.1016/0166-2236(95)93932-n

O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, Mak AC, Lai YY, Lin C, Chu C, Graves T, Kwok PY, Wilson RK, Sikela JM (2014) Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genomics 15: 387. doi: 10.1186/1471-2164-15-387

O'Bleness MS, Dickens CM, Dumas LJ, Kehrer-Sawatzki H, Wyckoff GJ, Sikela JM (2012) Evolutionary history and genome organization of DUF1220 protein domains. G3 (Bethesda) 2: 977-86. doi: 10.1534/g3.112.003061

Otani T, Marchetto MC, Gage FH, Simons BD, Livesey FJ (2016) 2D and 3D Stem Cell Models of Primate Cortical Development Identify Species-Specific Differences in Progenitor Behavior Contributing to Brain Size. Cell Stem Cell 18: 467-80. doi: 10.1016/j.stem.2016.03.003

Park IH, Lerou PH, Zhao R, Huo H, Daley GQ (2008) Generation of human-induced pluripotent stem cells. Nat Protoc 3: 1180-6.

Petanjek Z, Judaš M, Šimic G, Rasin MR, Uylings HB, Rakic P, Kostovic I (2011) Extraordinary neoteny of synaptic spines in the human prefrontal cortex. Proc Natl Acad Sci U S A 108: 13281-6. doi: 10.1073/pnas.1105108108

Pollard KS, Salama SR, King B, Kern AD, Dreszer T, Katzman S, Siepel A, Pedersen JS, Bejerano G, Baertsch R, Rosenbloom KR, Kent J, Haussler D (2006) Forces shaping the fastest evolving regions in the human genome. PLoS Genet 2: e168. doi: 10.1371/journal.pgen.0020168

Pollen AA, Bhaduri A, Andrews MG, Nowakowski TJ, Meyerson OS, Mostajo-Radji MA, Di Lullo E, Alvarado B, Bedolli M, Dougherty ML, Fiddes IT, Kronenberg ZN, Shuga J, Leyrat AA, West JA, Bershteyn M, Lowe CB, Pavlovic BJ, Salama SR, Haussler D, Eichler EE, Kriegstein AR (2019) Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. Cell 176: 743-756.e17. doi: 10.1016/j.cell.2019.01.017

Pontzer H, Brown MH, Raichlen DA, Dunsworth H, Hare B, Walker K, Luke A, Dugas LR, Durazo-Arvizu R, Schoeller D, Plange-Rhule J, Bovet P, Forrester TE, Lambert EV, Thompson ME, Shumaker RW, Ross SR (2016) Metabolic acceleration and the evolution of human brain size and life history. Nature 533: 390-2. doi: 10.1038/nature17654

Powell JL, Lewis PA, Dunbar RI, García-Fiñana M, Roberts N (2010) Orbital prefrontal cortex volume correlates with social cognitive competence. Neuropsychologia 48: 3554-62. doi: 10.1016/j.neuropsychologia.2010.08.004

Quick VB, Davis JM, Olincy A, Sikela JM (2016) DUF1220 copy number is associated with schizophrenia risk and severity: implications for understanding autism and schizophrenia as related diseases. Transl Psychiatry 6: e735. doi: 10.1038/tp.2016.11

Rakic P (1988) Specification of cerebral cortical areas. Science 241: 170-6. doi: 10.1126/science.3291116

Rakic P (1995) A small step for the cell, a giant leap for mankind: a hypothesis of neocortical expansion during evolution. Trends Neurosci 18: 383-8. doi: 10.1016/0166-2236(95)93934-p

Rakic P (2008) Confusing cortical columns. Proc Natl Acad Sci U S A 105: 12099-100. doi: 10.1073/pnas.0807271105

Rakic P (2009) Evolution of the neocortex: a perspective from developmental biology. Nat Rev Neurosci 10: 724-35. doi: 10.1038/nrn2719

Reillo I, de Juan Romero C, García-Cabezas M, Borrell V (2011) A role for intermediate radial glia in the tangential expansion of the mammalian cerebral cortex. Cereb Cortex 21: 1674-94. doi: 10.1093/cercor/bhq238

Rilling JK, Glasser MF, Preuss TM, Ma X, Zhao T, Hu X, Behrens TE (2008) The evolution of the arcuate fasciculus revealed with comparative DTI. Nat Neurosci 11: 426-8. doi: 10.1038/nn2072

Schenker NM, Buxhoeveden DP, Blackmon WL, Amunts K, Zilles K, Semendeferi K (2008) A comparative quantitative analysis of cytoarchitecture and minicolumnar organization in Broca's area in humans and great apes. J Comp Neurol 510: 117-28. doi: 10.1002/cne.21792

Sedmak D, Hrvoj-Mihić B, Džaja D, Habek N, Uylings HB, Petanjek Z (2018) Biphasic dendritic growth of dorsolateral prefrontal cortex associative neurons and early cognitive development. Croat Med J 59: 189-202.

Seim I, Fang X, Xiong Z, Lobanov AV, Huang Z, Ma S, Feng Y, Turanov AA, Zhu Y, Lenz TL, Gerashchenko MV, Fan D, Hee Yim S, Yao X, Jordan D, Xiong Y, Ma Y, Lyapunov AN, Chen G, Kulakova OI, Sun Y, Lee SG, Bronson RT, Moskalev AA, Sunyaev SR, Zhang G, Krogh A, Wang J, Gladyshev VN (2013) Genome analysis reveals insights into physiology and longevity of the Brandt's bat Myotis brandtii. Nat Commun 4: 2212. doi: 10.1038/ncomms3212

Semendeferi K, Damasio H (2000) The brain and its main anatomical subdivisions in living hominoids using magnetic resonance imaging. J Hum Evol 38: 317-32. doi: 10.1006/jhev.1999.0381

Semendeferi K, Teffer K, Buxhoeveden DP, Park MS, Bludau S, Amunts K, Travis K, Buckwalter J (2011) Spatial organization of neurons in the frontal pole sets humans apart from great apes. Cereb Cortex 21: 1485-97. doi: 10.1093/cercor/bhq191

Sherwood CC, Bauernfeind AL, Bianchi S, Raghanti MA, Hof PR (2012) Human brain evolution writ large and small. Prog Brain Res 195: 237-54. doi: 10.1016/B978-0-444-53860-4.00011-8

Shitamukai A, Konno D, Matsuzaki F (2011) Oblique radial glial divisions in the developing mouse neocortex induce self-renewing progenitors outside the germinal zone that resemble primate outer subventricular zone progenitors. J Neurosci 31: 3683-95. doi: 10.1523/JNEUROSCI.4773-10.2011

31/10/3683 [pii]

Silbereis JC, Pochareddy S, Zhu Y, Li M, Sestan N (2016) The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. Neuron 89: 248-68. doi: 10.1016/j.neuron.2015.12.008

Smaers JB, Gómez-Robles A, Parks AN, Sherwood CC (2017) Exceptional Evolutionary Expansion of Prefrontal Cortex in Great Apes and Humans. Curr Biol 27: 1549. doi: 10.1016/j.cub.2017.05.015

Smart IH, Dehay C, Giroud P, Berland M, Kennedy H (2002) Unique morphological features of the proliferative zones and postmitotic compartments of the neural epithelium giving rise to striate and extrastriate cortex in the monkey. Cereb Cortex 12: 37-53. doi: 10.1093/cercor/12.1.37

Sousa AMM, Zhu Y, Raghanti MA, Kitchen RR, Onorati M, Tebbenkamp ATN, Stutz B, Meyer KA, Li M, Kawasawa YI, Liu F, Perez RG, Mele M, Carvalho T, Skarica M, Gulden FO, Pletikos M, Shibata A, Stephenson AR, Edler MK, Ely JJ, Elsworth JD, Horvath TL, Hof PR, Hyde TM, Kleinman JE, Weinberger DR, Reimers M, Lifton RP, Mane SM, Noonan JP, State MW, Lein ES, Knowles JA, Marques-Bonet T, Sherwood CC, Gerstein MB, Sestan N (2017) Molecular and cellular reorganization of neural circuits in the human lineage. Science 358: 1027-1032. doi: 10.1126/science.aan3456

Spocter MA, Hopkins WD, Barks SK, Bianchi S, Hehmeyer AE, Anderson SM, Stimpson CD, Fobbs AJ, Hof PR, Sherwood CC (2012) Neuropil distribution in the cerebral cortex differs between humans and chimpanzees. J Comp Neurol 520: 2917-29. doi: 10.1002/cne.23074

Stepien BK, Naumann R, Holtz A, Helppi J, Huttner WB, Vaid S (2020) Lengthening Neurogenic Period during Neocortical Development Causes a Hallmark of Neocortex Expansion. Curr Biol 30: 4227-4237.e5. doi: 10.1016/j.cub.2020.08.046

Stepien BK, Vaid S, Huttner WB (2021) Length of the Neurogenic Period-A Key Determinant for the Generation of Upper-Layer Neurons During Neocortex Development and Evolution. Front Cell Dev Biol 9: 676911. doi: 10.3389/fcell.2021.676911

Stout D, Passingham R, Frith C, Apel J, Chaminade T (2011) Technology, expertise and social cognition in human evolution. Eur J Neurosci 33: 1328-38. doi: 10.1111/j.1460-9568.2011.07619.x

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F, Detours V, Vanderhaeghen P (2018) Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. Cell 173: 1370-1384.e16. doi: 10.1016/j.cell.2018.03.067

Taverna E, Götz M, Huttner WB (2014) The cell biology of neurogenesis: toward an understanding of the development and evolution of the neocortex. Annu Rev Cell Dev Biol 30: 465-502. doi: 10.1146/annurev-cellbio-101011-155801

Teffer K, Buxhoeveden DP, Stimpson CD, Fobbs AJ, Schapiro SJ, Baze WB, McArthur MJ, Hopkins WD, Hof PR, Sherwood CC, Semendeferi K (2013) Developmental changes in the spatial organization of neurons in the neocortex of humans and common chimpanzees. J Comp Neurol 521: 4249-59. doi: 10.1002/cne.23412

van den Heuvel MP, Bullmore ET, Sporns O (2016) Comparative Connectomics. Trends Cogn Sci 20: 345-361. doi: 10.1016/j.tics.2016.03.001

Varki N, Anderson D, Herndon JG, Pham T, Gregg CJ, Cheriyan M, Murphy J, Strobert E, Fritz J, Else JG, Varki A (2009) Heart disease is common in humans and chimpanzees, but is caused by different pathological processes. Evol Appl 2: 101-12. doi: 10.1111/j.1752-4571.2008.00064.x

Vazquez JM, Lynch VJ (2021) Pervasive duplication of tumor suppressors in Afrotherians during the evolution of large bodies and reduced cancer risk. Elife 10. doi: 10.7554/eLife.65041

Vrba E (1998) Multiphasic growth models and the evolution of prolonged growth exemplified by human brain evolution. *Journal of Theoretical Biology* 190: 227-239.

Wrangham RW, Jones JH, Laden G, Pilbeam D, Conklin-Brittain N (1999) The Raw and the Stolen. Cooking and the Ecology of Human Origins. Curr Anthropol 40: 567-594.

Xing L, Kubik-Zahorodna A, Namba T, Pinson A, Florio M, Prochazka J, Sarov M, Sedlacek R, Huttner WB (2021) Expression of human-specific ARHGAP11B in mice leads to neocortex expansion and increased memory flexibility. EMBO J 40: e107093. doi: 10.15252/embj.2020107093

Yin X, Mead BE, Safaee H, Langer R, Karp JM, Levy O (2016) Engineering Stem Cell Organoids. Cell Stem Cell 18: 25-38. doi: 10.1016/j.stem.2015.12.005

Zhou H, Ding S (2010) Evolution of induced pluripotent stem cell technology. Curr Opin Hematol 17: 276-80. doi: 10.1097/MOH.0b013e328339f2ee

Zimmer F, Montgomery SH (2015) Phylogenetic Analysis Supports a Link between DUF1220 Domain Number and Primate Brain Expansion. Genome Biol Evol 7: 2083-8. doi: 10.1093/gbe/evv122

Chapter 2 [1]
Identifying Adaptive Genetic Variation in the Human Genome

2.1 INTRODUCTION

The evolution of modern humans involved extensive migration to diverse geographic regions with shifting environmental conditions. Changes in climate, diet, and novel exposures to pathogens coincided with transitions in sociocultural organization (Laland et al. 2010). During these shifts, changes in the genome that initially arose as stochastic events within individuals of populations were subjected to evolutionary forces over generational time. These forces of random genetic drift, gene flow, and natural selection have together shaped the genetic diversity of our species. Evidence supports that a large portion of this variation has evolved neutrally (Harris 2018; Kimura 1968; Kimura 1983). Nonetheless, environmental and cultural factors jointly have imposed selective pressures on the human genome, contributing to extant patterns of human genetic diversity (Feldman and Cavalli-Sforza 1976). Species-specific and local adaptation as result of natural selection contributes to the phenotypic variation observed across human populations (Bamshad and Wooding 2003; Vitti et al. 2013). Deciphering

[1] This chapter represents a published review. Werren EA, Garcia O, Bigham AB. Identifying Adaptive Alleles in the Human Genome: from Selection Mapping to Functional Validation. *Human Genetics.* 2021. 140(2):241-276. Epub. 2020.

which genetic changes from natural selection underlie adaptive traits in humans is an ongoing area of investigation in evolutionary biology. Additionally, current environments may no longer be suitable for formerly beneficial alleles, resulting in a gene-environment mismatch underlying human disease (Gluckman et al. 2009; Nesse et al. 2010; Nesse and Stearns 2008; Williams and Nesse 1991). Determining the genetic basis of human adaptation therefore has broad implications for understanding human biology, health, and disease.

2.2 SELECTION MAPPING STATISTICAL DISCOVERY METHODS

Natural selection can be described in three general categories: positive directional selection, balancing selection, and negative purifying selection (Figure 2.1A). Each mode of natural selection will influence the alleles in a population in different ways (Gillespie 1994). Positive selection favors the fixation of an advantageous allele that increases the evolutionary fitness of its carrier. Balancing selection acts to maintain frequencies of multiple alleles at a locus for fitness benefit. Purifying selection, also known as background selection, removes deleterious alleles that reduce the fitness of the carrier from the population (Charlesworth et al. 1993). Purifying selection is essential for the evolutionary conservation of biological function (Cvijovic et al. 2018). The effects of strong purifying selection on genomic diversity have been characterized in several organisms (Begun and Aquadro 1992; Cvijovic et al. 2018; McVicker et al. 2009). Signals of evolutionary conservation due to negative purifying selection and accelerated evolution driven by positive and balancing selection have been leveraged to detect functionally significant regions of the genome (e.g., Thomas et al. 2003; Nielsen et al. 2005; Sawyer et al. 2005). Here we focus on the latter two forms of natural selection, positive and balancing, as these forms lead to the fixation or maintenance of adaptive alleles in a population over time (Maynard-Smith and Haigh 1974). Furthermore, these evolutionarily favored genetic changes can be identified through a process called selection mapping and functionally assessed for relevance to proposed adaptive biologies.

Selection mapping refers to the process of using population genetics methods to detect genomic regions, genes, or variants that have been acted on by natural selection (Wisser et al. 2008). It is founded on the principle that phenotypic change in a population is accompanied by changes in allele frequencies (Fumagalli et al. 2010; Vitti et al. 2013; Wisser et al. 2008). Species- and population-specific changes in allele frequency due to natural selection can leave distinct statistical signatures at the target loci relative to neutrally evolving loci (Fu 1995; Fu and Li 1993). Inferred signals of selection can be leveraged to identify fitness-related genomic regions, genes, and/or variants. Larger detected genomic regions can be further fine mapped to pinpoint putative functional alleles contributing to adaptive phenotypes (Akbari et al. 2018; Schaid et al. 2018; Szpak et al. 2018). Indeed, selection mapping has been widely used by evolutionary and population geneticists to identify candidate functional alleles underlying adaptive phenotypic change in plant and animal populations, including humans (Akey et al. 2002; Matsumoto et al. 2017; Wisser et al. 2008).

This review will discuss the application of selection mapping approaches to identify genetic variation contributing to adaptive phenotypes in human populations. We begin by describing the search for candidate adaptive alleles, from the detection of selection signatures (namely positive and balancing) through genotypic-phenotypic association and fine mapping variants. We provide a review of the discovered candidate adaptive alleles in human populations followed by a summary of the molecular approaches for functional validation of their biological role. While the bulk of genotyping and sequencing efforts have focused on European populations (Popejoy and Fullerton 2016; Sirugo et al. 2019), we detail several cases of selection mapping in non-European populations. Nevertheless, we acknowledge the impact of this major Eurocentric bias in the current map of natural selection in humans, and we conclude with comments on future directions and inclusivity considerations for population geneticists, human geneticists, and molecular anthropologists.

Genetic variants that define the alleles subjected to natural selection, due to effects on biological fitness, are distributed across functional coding and non-coding genomic elements (Bamshad and Wooding 2003; Jha et al. 2015; Wooding et al. 2002; Zhao et al. 2000). The types of genetic variants targeted by natural selection include nucleotide substitution, insertion, deletion, duplication, inversion, and translocation (Figure 2.1B). Natural selection can act on either *de novo* genetic variant(s) or standing genetic variation, leaving behind distinct signatures in the genome (Barrett and Schluter 2008; Hermisson and Pennings 2005; Przeworski et al. 2005). For selection on *de novo* variation, a novel variant arises in a population in which a selective pressure is present (Peter et al. 2012). For selection to act, this new variant will confer either an adaptive advantage or disadvantage to the existing pressure. Selection on standing variation may occur when a population faces an environmental change or range expansion that results in a new selective pressure (Przeworski et al. 2005). Thus, standing variation is preexisting variation in the ancestral population at the time the environmental conditions for selection appear (Barrett and Schluter 2008). Depending on the mode of natural selection, genomic diversity will be shaped differently in a population over time.

Distinguishing selection on de novo variants from standing variation

The structure of haplotypes—sets of alleles (or genetic variants) that are inherited together on a single chromosome in an individual—in a population offer clues into how natural selection has acted on the genome (Bamshad and Wooding 2003; Nordborg and Tavare 2002; Sabeti et al. 2002b). In particular, the heterogeneity of haplotype backgrounds that comprise the allele under natural selection, as well as the changes in allele frequency, help to distinguish between signals of selection from standing variation versus *de novo* variation (Barrett and Schluter 2008; Peter et al. 2012). If a selective pressure is present before a favorable variant originates, as in the case of selection on a *de novo* variant, then the new (or derived) allele will come to exist on a single, distinct haplotype (Figure 2.1A) in the study population and will be absent in the ancestral population, where only the ancestral allele is found (Nielsen et al. 2007; Peter et al. 2012). If the derived allele reaches fixation rapidly from positive selection, haplotype homozygosity will be strong around the selected allele before mutation and recombination act to break it down (Nielsen et al. 2007; Peter et al. 2012). In other words, the linkage disequilibrium (LD)—the non-random association of alleles in haplotypes—will be high resulting from genetic hitchhiking of neutral alleles linked to the adaptive allele during the rise in frequency (Maynard-Smith and Haigh 1974; Przeworski 2002). This is referred to as a 'hard' selective sweep. Effective population size and the effect size of the adaptive allele contribute to its fixation time, with shorter time to fixation in cases of small effective population size and large effect alleles (Pritchard et al. 2010). Once a derived allele has swept to high frequency, recombination acts to break up linkage with neutral sites resulting in an excess of both intermediate and high frequency alleles (Barrett and Schluter 2008; Fay and Wu 2000). However, if the selective sweep is incomplete, i.e., the beneficial allele is in the midst of being swept to fixation in the population, heterozygosity increases, and this is called a 'soft' sweep (Pritchard et al. 2010).

Selection also acts on standing variation. When selection favors variation already present in the population, the pre-adaptive alleles may exist on several distinct haplotypes, resulting in a greater number of linked neutral alleles at intermediate frequencies post-selection (Figure 2.1A) (Barrett and Schluter 2008; Peter et al. 2012; Przeworski et al. 2005). In other words, alleles that are maintained within the population as neutral or weakly functional before becoming selectively advantageous will have a longer population history than *de novo* alleles (Barrett and Schluter 2008). This protracted population history increases the likelihood that recombination has broken up haplotype blocks containing the standing variant, placing it on different backgrounds (Barrett and Schluter 2008; Peter et al. 2012; Przeworski et al. 2005). Selection on standing variants will integrate more polymorphisms than selective sweeps of *de novo* variants, and this distribution across haplotypes weakens the statistical signature of selection (Barrett and Schluter 2008; Hermisson and Pennings 2005; Peter et al. 2012; Przeworski et al. 2005). Variation around the fixed adaptive allele will therefore be lower in a population following selection on *de novo* variation compared to selection on standing variation (Barrett and Schluter 2008). Given the predicted effects of selection on alleles and haplotypes between populations, the type of sequence information acquired (SNP microarray, targeted sequencing, whole exome, whole genome, etc.) and selection of study populations will determine which methods (Table 1) can be applied for detecting natural selection in the genome.

Haplotype tests of selection

Several extended haplotype tests for identifying selective sweeps exist and include the Long-Range Haplotype (LRH), integrated Haplotype Score (iHS), number of segregating sites by length statistic (nSL), and Cross Population Extended-Haplotype Heterozygosity (XP-EHH) (Table 1) (Sabeti et al. 2002b; Sabeti et al. 2007; Voight et al. 2006). The LRH test takes into consideration both the length of extended haplotype homozygosity (EHH) and the frequency of the haplotype in the population (Sabeti et al. 2002b). This test performs well when identifying haplotypes that arose quickly and have yet to be broken down by recombination, such as for hard sweeps. A modification of this test, the whole genome long-range haplotype (WGLRH) test, has been developed to use genome-wide data to detect selection by combining LRH with derived and ancestral allele status (Zhang et al. 2006). The iHS test quantifies the strength of selection acting at a locus in a single population by calculating the integral of EHH decay at varying distances away from a core allele (Figure 2.2) (Voight et al. 2006). It considers the ratio of haplotype homozygosity for haplotypes carrying the derived allele relative to those carrying the ancestral allele at the candidate locus. By doing so, iHS is sensitive to variation in recombination rates across the genome as the two alleles serve as controls for each other, canceling out heterogeneity in the estimated genetic map due to recombination coldspots and hotspots. This statistic is more powered to detect incomplete, or soft, selective sweeps (Pickrell et al. 2009; Voight et al. 2006). Like iHS, nSL uses the ratio of haplotype homozygosity for derived and ancestral alleles to identify positive selection in a single population (Ferrer-Admetlla et al. 2014). nSL differs from iHS in that it quantifies EHH length for a given haplotype pair based on the number of SNPs present in that region on all other haplotypes in the sample population (Ferrer-Admetlla et al. 2014; Voight et al. 2006). In doing so, a genetic map is not needed to calculate nSL, thereby further increasing sensitivity to varying recombination and mutation rates, with greater power for detecting both hard and incomplete/soft selective sweeps. XP-EHH compares haplotypes between populations to identify derived selected alleles that have recently reached or nearly reached fixation in one population yet remain polymorphic across all human populations (Sabeti et al. 2007). It performs well for detecting sweeps that have reached, or have almost reached, fixation (Pickrell et al. 2009; Sabeti et al. 2007; Voight et al. 2006). As such, it has weaker power to detect older selective sweeps that have been subjected to LD degradation by recombination over time.

Confounding variables in haplotype-based tests

Population-specific recombination rate and copy-number variants (CNVs) may confound the results of haplotype-based tests of selection and should be considered (Sabeti et al. 2007). Certain regions of the genome exhibit variable recombination rates between populations (Abecasis et al. 2010). A long-range haplotype signal at these loci could result from a slower recombination rate in one population relative to other populations, instead of a spike in allele frequency from natural selection (Sabeti et al. 2002b). To account for this, the level of EHH decay of non-selected alleles in the region surrounding the candidate allele can be compared to the EHH decay of those same non-selected alleles in other populations (Sabeti et al. 2002b). If the EHH decay is similar between populations, variable recombination rate can be ruled out as the source of the long-range signal, bolstering claims of selection (Sabeti et al. 2002b). CNVs can result in the increase of nearly identical sequence at a locus that will reduce heterozygosity (Lucas et al. 2019). This makes it challenging to distinguish a selective sweep from a region with high copy number variation (Iskow et al. 2012). Likewise, CNVs that lie close to a selected locus may falsely extend the signal (Sabeti et al. 2007). To account for potential confounding from CNVs, recent innovations in long-read sequencing technologies that avoid PCR amplification steps offer a platform to more accurately map and quantify CNVs (Pollard et al. 2018). Further, hybrid-sequencing approaches that combine long-read with short-read technologies that offer higher resolution and lower genotyping error rates, can afford both CNV accuracy and resolution (Kronenberg et al. 2018).

Allele frequency tests

The joint distribution of allele frequencies in a given population is referred to as the allele frequency spectrum (AFS) or the site frequency spectrum (SFS) (Ewens 1972). The standard neutral model SFS can be predicted theoretically under the assumptions of no selection, no recombination, and that all new genetic variants are unique (Ewens 1972). Selection can act to shift the AFS (Figure 2.2) (Fu 1995). Positive selection results in an increase in high frequency alleles and a near absence of intermediate frequency alleles whereas balancing selection results in an excess of alleles at intermediate frequency (Tajima 1989). Demographic events impact the SFS in the same manner as selective events (Marth et al. 2004). Similar to balancing selection, population subdivision reduces the number of highly shared alleles in the population as a whole, thereby increasing the number of intermediate frequency alleles (Tajima 1989). Conversely, population expansion increases the number of singletons and high frequency alleles, much like directional positive selection (Tajima 1989). However, demographic changes are expected to influence all loci across the genome in a similar manner whereas selection tends to be locus-specific (Biswas and Akey 2006; Przeworski et al. 2000). Therefore, to distinguish demography from selection, the SFS at one locus can be compared to the genome-wide average in each population (Akey et al. 2004; Biswas and Akey 2006).

SFS summary statistics use sequence data to test for departures from neutrality and are useful for identifying loci that have undergone positive or balancing selection (Table 1) (Fay and Wu 2000; Fu and Li 1993; Nielsen et al. 2007; Przeworski et al. 2005; Tajima 1989). Tajima's *D* measures the difference between the number of segregating sites (Θ_W) and the average number of nucleotide differences (π) (Tajima 1989). Under a neutral model, Θ_W and π are equal, and Tajima's *D* is zero. Positive values indicate an excess of intermediate frequency alleles suggesting either balancing selection or population subdivision (Nielsen et al. 2007; Przeworski et al. 2005; Tajima 1989). Negative values indicate an excess of rare alleles and may suggest positive selection or rapid population growth (Nielsen et al. 2007; Przeworski et al. 2005; Tajima 1989). Fu and Li's *D* and *F* statistics compare population mutation rate estimates obtained from derived singleton variants, to Θ_W and π, respectively. Both summary statistics rest on the premise that older variants tend to exist on internal branches, or roots of a phylogenetic tree, while newer variants exist on the external branches (Fu and Li 1993). Therefore, recent positive selection and purifying selection are predicted to result in an excess of variants that lie on the external branches resulting in negative values of the statistics. Balancing selection is predicted to result in positive values of *D* and *F*. One benefit to Fu and Li's *D* is that it is more sensitive to selective sweeps acting on derived alleles compared to Tajima's *D*. However, selection acting on standing variation can favor ancestral alleles, limiting the ability of these statistics to distinguish positive selection on standing variation. Fay and Wu's *H* statistic measures the amount of high and intermediate frequency derived variants to detect positive selection (Fay and Wu 2000). A negative value of *H* results from an excess of high frequency derived alleles while a positive value implies a lack of intermediate and high frequency derived alleles compared to neutral expectations (Fay and Wu 2000). While SFS summary statistics are popular statistical tests for identifying deviations from neutral evolution within a population, other methods have been developed for comparisons between populations.

Population-specific natural selection can result in allele frequency changes in one population compared to a neutrally evolving population. Wright's fixation index statistic, F_{ST}, measures the amount of genetic differentiation between two or more populations (Table 1) (Wright 1950). Positive selection tends to increase the population differentiation estimates at a selected locus whereas balancing selection tends to decrease F_{ST} (Akey et al. 2002; Andolfatto 2001; Cavalli-Sforza 1966; Wright 1950). Genetic drift also may increase the level of differentiation between populations. However, while positive selection is confined to one or few loci, genetic drift is expected to influence heterogeneity at many unlinked loci across the genome.

Phylogenetic-based tests

Methods that expand on allele frequency differences to explore phylogenetic relationships have been developed to identify population-specific selective events (Table 1, Figure 2.2). The locus-specific branch length (LSBL) approach incorporates F_{ST} distances to analyze population-specific allele frequency shifts (Shriver et al. 2004). This method compares F_{ST} between three closely related populations in order to express distances between populations in terms of branch lengths (Shriver et al. 2004). Likewise, the population branch statistic (PBS) uses phylogenetic theory to evaluate changes in allele frequency since population divergence (Yi et al. 2010). Both LSBL and PBS test the null hypothesis that two closely related groups, A and B, have similar branch lengths when compared to an outlier group C. If population A experienced positive selection at a locus, the effect will be a longer branch length than expected. This tree structure will look very different from that of the tree using genome-averaged allelic differentiation values for the three populations.

A widely used metric that measures substitutions to draw inferences about positive or purifying selection is the d*N*/d*S* ratio, or *ω*, which quantifies the ratio of nonsynonymous to synonymous nucleotide changes (Kimura 1977; Miyata and Yasunaga 1980). Briefly, a high ratio greater than one indicates an excess of nonsynonymous changes suggesting positive selection whereas low estimates under one suggests purifying selection. Values at or near one indicate neutrality. Although initially intended for use in conjunction with phylogenetics analyses on sequences from distantly-related lineages, major issues have arisen when attempting to interpret ω for within species variation or closely-related lineages due to the effects of recombination and large population size as both can falsely inflate signals of positive selection (Anisimova et al. 2003; Mugal et al. 2014; Schierup and Hein 2000; Shriner et al. 2003). Expanding on ω, the McDonald-Kreitman test (MKT) uses both polymorphic and divergence information at tested sites from two organisms to test the null hypothesis of neutrality where the dN/dS ratio within a species is equal to the proportion of nonsynonymous to synonymous variants between species (pN/pS); when the dN/dS ratio is greater than the pN/pS ratio, positive selection is suggested whereas if dN/dS is less than pN/pS then purifying selection is hypothesized (McDonald and Kreitman 1991). Given the MKT assumes strong neutrality at segregating sites, weaker negative selective pressures from slightly deleterious variants as well as population size can confound estimates (Charlesworth and Eyre-Walker 2008). Extensions of this metric have been developed to try and correct for errors of MKT, including the Fay, Wyckoff and Wu correction ($_{\text{FWW}}$MKT), the extended MKT (eMKT), and the asymptotic MKT (aMKT) (Fay et al. 2001; Mackay et al. 2012; Messer and Petrov 2013).

Several phylogenetic methods have expanded on dN/dS and offer different information to infer the strength of positive selection in sequencing data. Examples include: 1. Fixed Effects Likelihood (FEL), which uses a maximum likelihood model to infer substitution rates per site for a phylogeny (Kosakovsky Pond and Frost 2005), 2. the Genetic Algorithm for Recombination Detection (GARD) that infers phylogenetic histories for detected recombination events that can be used when making selection inferences (Kosakovsky Pond et al. 2006), 3. Mixed Effects Model of Evolution (MEME) which uses a mixed-effects maximum likelihood model to test for episodic positive selection or diversifying selection on a site-level under a proportion of branches (Murrell et al. 2012), 4. Fast, Unconstrained Bayesian AppRoximation (FUBAR) which uses a Bayesian approach to infer substitution rates per site for a phylogeny (Murrell et al. 2013), 5. the adaptive Branch-Site Random Effects Likelihood (aBSREL), which tests for positive selection on a proportion of branches but not at specific sites (Smith et al. 2015), and 6. the Branch-Site Unrestricted Statistical Test for Episodic Diversification (BUSTED) that tests for positive selection on a gene-level (Murrell et al. 2015). Collectively, these methods can be accessed via the Hypothesis Testing using Phylogenies (HyPhy) site (stevenweaver.github.io/hyphy-site/methods/selection-methods/#).

Composite tests

Composite tests for selection combine a subset of the independent tests for selection described above to increase resolution of the selective signature. Composite methods thus decrease the probability that the signal detected is a false positive and are an effective way to fine-map causal variants. These methods include the composite likelihood ratio (CLR), cross population composite likelihood ratio (XP-CLR), 3.population composite likelihood ratio (3P-CLR), the composite of multiple signals (CMS), and the SWeep Inference Framework (controlling for correlation) (SWIF(r)) (Table 1). CLR identifies an excess of derived alleles across several sites in a single population (Kim and Stephan 2002; Nielsen et al. 2005) whereas XP-CLR does so in two populations (Chen et al. 2010; Vitti et al. 2013). By integrating multiple populations, XP-CLR models allele frequency differences across a chromosome as predicted by the genetic distance from the selected allele (Chen et al. 2010). XP-CLR has greater power to detect selection from standing variation as well as 'ancient' selective sweeps—sweeps that arose following the divergence of two populations, yet ended several thousand generations before the present (Chen et al. 2010). 3P-CLR is a modification of CLR used to distinguish selection events following the divergence of two populations, as well as selection sweeps shared between the two populations prior to the divergence from a third outgroup population (Racimo 2016). The CMS method combines F_{ST}, iHS, and XP-EHH for derived, high frequency alleles in one population compared to other populations (Grossman et al.

2010). This method has been shown to significantly reduce the number of genomic candidates of recent selection for local adaptation in human populations (Grossman et al. 2013; Grossman et al. 2010). Similar to CMS, SWIF(r) combines scores from F_{ST}, iHS, and XP-EHH (Sugden et al. 2018). However, this method also incorporates a machine learning mechanism whereby SWIF(r) is trained on simulations from user-selected demographic models (Sugden et al. 2018). By doing so, the probabilistic statistic does not require comparison with a genome-wide distribution, as is the case for CMS, and potentially increases power to localize candidate adaptive loci in both soft and hard selective sweeps (Sugden et al. 2018).

Polygenic adaptation

Polygenic adaptation occurs when multiple adaptive variants across many loci are favored by selection. Selection for polygenic traits only produces slight changes in frequencies of each of the causal alleles (Figure 2.1A) (Pritchard et al. 2010). Consequently, the response of a single allele will be statistically negligible, and the signal of selection only can be identified in the cumulative frequency shift of all alleles associated with the phenotype in the population. In other words, polygenic selection for adaptive traits will dampen the signature of selection at each independent locus (Berg and Coop 2014; Haasl and Payseur 2016; Latta 1998), reducing the power of detection by summary statistics (Latta 1998; Le Corre and Kremer 2003). One strategy to increase sensitivity to detection of selection on multi-allelic traits is to apply regression analysis to extended haplotype homozygosity measures (Wiener and Pong-Wong 2011). The regression approach has been proven to be most effective at detecting polygenic selection when allelic diversity is high (Wiener and Pong-Wong 2011).

Limitations for selection mapping on ancient DNA

Selection mapping also has been applied in the burgeoning field of archaeogenetics that seeks to use ancient DNA (aDNA) to resolve questions of human evolution, including the timing of selection events (Knapp and Hofreiter 2010). Hominin comparative genomics has implications for addressing if candidate human-specific genetic variation identified from comparative primate genomics evolved recently in modern humans, or is shared with earlier humans, potentially offering insight into human phenotypic diversity in the past. However, while some have performed aDNA selection scans (Fehren-Schmitz and Georges 2016; Lewandowska et al. 2018; Lindo et al. 2018; Mathieson et al. 2015), applying a selection mapping approach with aDNA sequences is often very difficult given the highly degraded, short fragment nature of recovered samples (Knapp and Hofreiter 2010). This results in low sequencing coverage, often with only one mapped read at each nucleotide site (Gunther and Nettelblad 2019; Haak et al. 2015; Monroy Kuhn et al. 2018). Given this low sequencing coverage, diploid calls cannot be made, resulting in sequences that cannot be confidently phased and thus are unsuitable for haplotype-based selection tests (Gunther and Nettelblad 2019; Haak et al. 2015; Monroy Kuhn et al. 2018). Therefore, aDNA selection scans are limited to allele frequency-based tests (Gunther and Nettelblad 2019; Haak et al. 2015; Monroy Kuhn et al. 2018). Technological advances in aDNA recovery that improve sample quality and yield upon collection and extraction, as well as methodological improvements of imputation and phasing, may offer more opportunities in the future for selection mapping of aDNA to inform evolutionary models of the chronology and tempo of selective events.

2.3 IDENTIFYING CANDIDATE ADAPTIVE VARIATION IN THE HUMAN GENOME

The suite of statistical selection methods provides a means for detecting genomic loci with evidence of positive or balancing selection that can be further dissected to assess their contribution to biological adaptations. Study designs utilizing selection detection methods to hunt down functional adaptive alleles in the genome, and to inform evolutionary hypotheses of human adaptation, can be thought of in two categories: forward genetics or reverse genetics. Those pursuing forward genetics approaches start with a putative adaptive phenotype and uncover the genetic basis (Bigham and Lee 2014; Vitti et al. 2013). In this way, forward genetics seeks to test a specific hypothesis. A classic example of hypothesis-driven, forward genetics using selection mapping to understand human adaptation is the discovery of a causal variant in *HBB* underlying sickle cell anemia that was subsequently found to be under balancing selection in populations living in malaria-endemic regions (Allison 1954; Vitti et al. 2013).

Reverse genetics, on the other hand, seeks to understand the biological function of observed genetic variation, and therefore is hypothesis-generating (Figure 2.2) (Hardy et al. 2010). With advancements in genotyping arrays and genome sequencing technology, there have been a burst of whole-genome scans for signals of adaptive evolution in human populations (Crawford et al. 2017b; Eaaswarkhanth et al. 2020; Grossman et al. 2013; Ilardo et al. 2018; Karlsson et al. 2013). Often, these scans identify several candidate genes and/or regulatory regions, or detect signals spanning large segments of the genome (Vitti et al. 2013). To pinpoint the causal variant(s) that underlies a putative adaptive phenotype and understand its function, genetic variation must first be linked to biological phenotypes.

Identifying putative adaptive phenotypes associated with selection signatures

A common strategy for linking genetic variation to phenotypic variation is to perform genotype-phenotype association analyses. Association analysis can be targeted, such as candidate gene association studies, or genome-wide, such as genome-wide association studies (GWAS) (Rodriguez-Murillo and Greenberg 2008). Population-based association mapping is the process of linking genotypes to phenotypes within a population of unrelated individuals by leveraging LD (Singh and Singh 2015). Study designs include case-control association using a logistic regression and linear regression for quantitative traits (Tak and Farnham 2015). These approaches can help detect genomic regions harboring functional variants, but need to be followed up with further statistical and functional testing to identify the causal variants driving the association (Schaid et al. 2018).

The power of association mapping is dependent upon genotyped markers that lie in regions of strong, detectable LD in the population of interest (Risch and Merikangas 1996; Tak and Farnham 2015). Sample size, effect size, number of causal variants, and allelic heterogeneity all influence the statistical power of detecting an association (Schaid et al. 2018). A major issue with association mapping is that neutral variants linked to the causal variant can be identified as false positives (Tak and Farnham 2015). Also, association mapping is less powered to detect rare functional variants and those of small effect size (Xu et al. 2017). Better phenotypic measurements can sometimes improve the genetic effect size as can increased SNP density and sample size (Schaid et al. 2018). Furthermore, when phenotypic variation correlates with genetic relatedness, or when the study sample has unidentified population structure, markers may produce a false signal of association (Kang et al. 2008). Study designs can control for population structure to reduce the number of false associations (Kang et al. 2008). Given that association mapping can become cumbersome when several causal alleles exist or the relevant phenotypic variation is unavailable, a more efficient strategy to find and prioritize candidate alleles may be preferred. For example, some association methods combine expression data with genetic and phenotypic variables, such as with expression Quantitative Trait Locus (eQTL) mapping (Hormozdiari et al. 2016) and Backward Three-way Association Mapping (BTAM) (Lee et al. 2017).

Integrating selection mapping with association analysis can be a powerful tool to refine loci and localize candidate adaptive alleles (Guo et al. 2018; Johnsson 2018). For example, testing genetic variation in the candidate pigmentation gene, *SLC24A5,* that is under positive selection revealed that the rs1426654-A allele was able to explain ~22-32% of total skin color variation in a cohort of individuals from across South Asia (Basu Mallick et al. 2013). Furthermore, selection mapping can be performed before or after genotype-phenotype association analysis, or the two can be statistically combined. However, most combined selection-association mapping studies identify several candidate genes or larger candidate genomic regions. For instance, studies in Bangladeshi individuals from the Ganges River Delta show evidence that cholera has exerted strong selective pressures at ~305 candidate genes (Chowdhury et al. 2011; Faruque et al. 2003; Karlsson et al. 2013). Likewise, for putative polygenic adaptations, extensive lists of candidate loci driving overlapping selection and association signals have been compiled, such as for high animal-fat diet adaptation in indigenous Central Siberian populations (lipid and fatty acid metabolism genes *FADS1, FADS2, HADHA, HADHB,* *MAN1B1, PLD2,* and *AGPAT1*) (Hsieh et al. 2017), or pygmy stature adaptation in African rain forest hunter-gatherers (~60 Mb region comprising genes *DOCK3*, *MAPKAP3*, and *CISH)* (Jarvis et al. 2012). While these data represent key initial steps to understanding human genetic adaptation, they nevertheless substantiate the need for further strategies to map causal alleles. These datasets thus serve as rich sources for the fine-mapping and prioritization of candidate adaptive variants for functional testing.

Fine mapping and functional annotation to prioritize candidate adaptive alleles

'Fine mapping' is a strategy to localize causal variants from association and/or selection mapping hits through statistical, bioinformatic, and/or functional approaches (Schaid et al. 2018). Common statistical fine mapping methods, developed for GWAS hits, include heuristic methods (analyze pairwise correlations of each surrounding SNPs with the identified lead SNP), penalized regression models (use cross-validation to randomly partition into training sets to select SNPs to use in the regression), and Bayesian methods (probabilistic modeling) (Schaid et al. 2018). Fine mapping methods include CMS (described above), 'selection detection by conditional coalescent tree' (SCCT) (Wang et al. 2014), 'integrated selection of allele favored by evolution' (iSAFE, exploits the shoulders of sweeps with high performance up to 5 Mbp regions) (Akbari et al. 2018), and Fine-Mapping of Adaptive Variation (FineMAV, combines population differentiation, derived allele frequency, and molecular functional annotation) (Szpak et al. 2018).

Functional annotation of variants present within candidate genes or non-coding regulatory regions can help rank and prioritize variants for functional investigation. It can be used to inform statistical fine mapping or be assessed following fine mapping analyses (Schaid et al. 2018). Annotation criteria include mutation type (substitution, insertion, deletion, duplication, inversion, translocation), effect on coding sequence (missense, nonsense, frameshift, splice), and non-coding regulatory element type (promoter, enhancer, silencer, insulator) (Ip et al. 2019; Zhou and Zhao 2018). Variants also can be prioritized based on where they localize in the functional element, such as within a particular protein domain, splicing site, transcription factor binding site, among others (Ip et al. 2019; Zhou and Zhao 2018).

Functional annotations can be made from searching published literature, mining publicly available functional annotation databases, and using free bioinformatics tools. CAUSALdb is a free database that offers uniform fine-mapping of GWAS summary data pulled from 3,000 studies using three state-of-the-art tools (including PAINTOR, CAVIARBF and FINEMAP), combined with functional annotation, offering users a browser view of variant-, gene-, and trait-level causal relations (mulinlab.tmu.edu.cn/causaldb, (Wang et al. 2020). Some tools have been curated for improving variant prioritization based on annotation data and functional prediction, such as Variant Prioritization Ordering Tool (VPOT) (Ip et al. 2019). The Encyclopedia of DNA Elements (ENCODE) Project portal (encodeproject.org) is a great resource for genome-wide annotations of functional elements, including a web interface called SCREEN, or Search for Candidate Cis-Regulatory Elements by ENCODE (screen.wenglab.org). The UCSC Genome Browser is another excellent resource to simultaneously visualize functional annotations across a genomic interval from several databases, with annotation tracks for cross-species conservation scores, structural variants, phenotype-associated variants from the Online Mendelian Inheritance in Man (OMIM, omim.org), ClinVar (ncbi.nlm.nih.gov/clinvar/), UniProt (uniprot.org), and GWAS studies, to name a few. The browser also includes tracks for common SNPs from catalogs such as dbSNP (ncbi.nlm.nih.gov/projects/SNP) and the Genome Aggregation Database (GnomAD) (gnomad.broadinstitute.org). The GnomAD browser, developed by the BROAD Institute, compiles whole genome and exome sequencing data from disease-specific and population genetic cohorts, totaling 213,158 unrelated individuals, offering variant-level annotations (gnomad.broadinstitute.org). The Genotype-Tissue Expression (GTEx) Portal (gtexportal.org) provides extensive information on the relationship between genetic variants and human gene regulation/expression by using global RNA expression data across several tissue types and eQTL analysis. The WashU Epigenome Browser (epigenomegateway.wustl.edu) is a great visualization tool of high throughput epigenomic and expression data from diverse adult and embryonic cell types, including annotated regions of open chromatin from DNAse hypersensitivity assays as well as histone modifications associated with gene repression or activation.

Insight into the molecular mechanism of an adaptive change can be gained from determining where candidate variants localize with respect to protein structure and protein sequence. Some popular biomolecular structure databases include Protein Data Bank (PDB, rcsb.org/pdb), Structure Classification of Proteins (SCOP, scop.berkeley.edu), and 'Class, Architecture, Topology' (CATH, cathdb.info). Three-dimensional structural analysis can be helpful to model the variant effect on protein structure, which can be achieved using free bioinformatics tools, such as PyMol (pymol.org/2/) and Missense3D (sbg.bio.ic.ac.uk/~missense3d/). PDB offers a suite of other available molecular graphics software (rcsb.org/pages/thirdparty/molecular_graphics). Useful resources to analyze protein function and domain information include PFam (pfam.xfam.org), Prosite

(prosite.expasy.org), and PRINTS (130.88.97.239/PRINTS/index.php). Lastly, phenotype-drive mutagenesis screens, using different genome editing tools, are a rich source of annotated variants with visible phenotypes. Several high-throughput mutagenesis screens have been performed in diverse human cells, and available databases include GenomeRNAi (genomernai.org) and GenomeCRISPR (omictools.com/genomecrispr-tool). Functional annotation and bioinformatic modeling of genetic variants across diverse criteria using different online resources is a cost- and time-effective way to prioritize candidate variants for additional allele frequency tests and for functional investigation. Together these data can build a case for putative adaptive function in human evolution.

Major challenges remain for fine-mapping and variant prioritization despite the growing number of functional annotation resources and tools. This is particularly the case when results from integrated selection and association mapping studies detect extensive and/or numerous genomic loci; or when several causal variants of small effect contribute to the statistical signature of selection (Akbari et al. 2018; Schaid et al. 2018; Szpak et al. 2018; van de Bunt et al. 2015). While it remains an ongoing challenge to pinpoint the functional allele(s) driving the selection signal, we discuss successful examples in human populations below including infectious disease resistance and/or susceptibility, climatic adaptations, and dietary adaptations (Figure 2.3 and Table 2).

2.3.2 Candidate Adaptive Alleles within Human Populations

Infectious Disease Adaptations: Malaria

Searching for signatures of selection to identify susceptibility alleles for infectious disease is an appealing strategy as pathogens are hypothesized to have exerted strong selective pressures on evolving modern humans (Fumagalli et al. 2011; Haldane 1932; Karlsson et al. 2014). Many genes subject to local positive natural selection (e.g., *DARC*) or balancing natural selection (e.g., *G6PD, SLC4A1*) are associated with susceptibility to infectious disease (Bamshad et al. 2002; Hughes and Yeager 1998; Sabeti et al. 2002a). Malaria, perhaps the most extensively studied human infectious disease, imparts one of the strongest selective pressures on modern humans (Evans and Wellems 2002). Co-evolution with the malaria parasite has led homozygous deleterious alleles that are associated with specific blood disorders to confer a selective advantage in the heterozygous state (Kwiatkowski 2005; Luzzatto 2012). The most well-known adaptive allele that confers resistance to *Plasmodium falciparum* malaria is the sickle cell allele (HbS) of the beta-globin locus, *HBB*, wherein individuals heterozygous for the HbS allele are protected from malaria invasion due to the malformation of a portion of the erythrocytes, yet do not present with severe sickle cell anemia (Allison 1954; Haldane 1949; Piel et al. 2010). The HbS allele of SNP rs334, caused by a missense variant c.20A>T(p.Glu7Val) in *HBB*, may function to reduce malarial virulence by reducing surface expression of *P. falciparum* erythrocyte membrane protein-1 (PfEMP-1) on infected erythrocyte cells, thereby impairing their cytoadherence to microvascular endothelial cells, a property necessary for the parasite to escape bloodstream removal by the spleen (Cholera et al. 2008).

In addition to the well-known *HBB* example, additional candidate alleles under balancing selection have been identified that underlie malarial resistance adaptations. Hypomorphic alleles at the glucose-6-phosphate dehydrogenase locus (*G6PD)* on the X chromosome that cause G6PD enzyme deficiency are the genetic basis for a number of hemopathologies in humans (Gregg and Prchal 2018). Among these alleles, the rs137852331-G (or "*G6PD A*") allele and the rs1050828-A (or "*G6PD A-")* allele are under balancing selection for partial malaria resistance in sub-Saharan African populations, while the rs5030868-T (or "*G6PD Mediterranean*") allele may be under balancing selection in Southern European, Middle Eastern, and Indian populations (Kwiatkowski 2005; Saunders et al. 2002; Saunders et al. 2005; Tishkoff et al. 2001; Verrelli et al. 2002). The *G6PD A-* allele results in the missense change c.292G>A(p.Val98Met) that reduces G6PD activity by 82% compared to the ancestral allele, and is associated with a 46-58% reduction in severe malaria risk in both heterozygous females and hemizygous males (Ruwende et al. 1995). Southeast Asian Ovalocytosis (SAO) is another blood disease resulting in oval-shaped erythrocytes that is protective against severe vivax malaria (Jarolim et al. 1991). A 27 base-pair deletion (Δ27) in the *SLC4A1* gene which encodes an anion transport protein is responsible for this phenotypic trait, and while the *SLC4A1Δ27* allele is protective against malaria and typically asymptomatic in heterozygous individuals, it is lethal in its homozygous state (Jarolim et al. 1991; Paquette et al. 2015; Wilder et al. 2009). Indonesian populations exhibit an excess of high-frequency, derived variants around *SLC4A1Δ27* (Wilder et al. 2009), suggesting that natural selection has acted recently on this locus.

The Duffy antigen receptor for chemokine, *DARC* (or *ACKR1*), is another target of selection from malarial pressure (Oliveira et al. 2012). *DARC* encodes the transmembrane glycoprotein Duffy antigen (*Fy*) (Hadley and Peiper 1997), which exhibits allelic differentiation by geographic region with near fixation for the Duffy-null allele, *Fy-,* in the majority of sub-Saharan Africa (Livingstone 1984). The Duffy-null allele is due to a A>G substitution at rs2814778 in the GATA promoter of *DARC*, and individuals homozygous for Duffy-null (*Fy-/Fy-*) are resistant to erythrocyte invasion by *Plasmodium vivax* malaria (Miller et al. 1976). The selection signature around *DARC* extends ~6 kb upstream of the 5' UTR (Hamblin and Di Rienzo 2000). Levels of sequence diversity and site frequency spectrum results indicate directional positive selection at *DARC* in sub-Saharan African populations (the Mandinka, Beti, Hausa, Mbuti, and Luo) when compared to sequence data from non-African groups (Italian, Han Chinese, Pakistani) that do not fit simple models of selection (Hamblin and Di Rienzo 2000; Hamblin et al. 2002). These results support the hypothesis that the fixation of the Duffy-null allele in mainland sub-Saharan Africa was the result of selective pressures from vivax malaria. Additionally, multiple Malagasy populations in Madagascar show high frequencies of the Duffy-null allele, with evidence for recent positive selection in the Merina Highlanders (Hodgson et al. 2014).

The *GYPA* locus encodes a ligand of the glycophorin A receptor expressed on the cell surface of erythrocytes that *P. falciparum* malaria uses to gain cell entry (Camus and Hadley 1985; Sim et al. 1994; Tolia et al. 2005). From targeted selection tests, *GYPA* shows patterns of variation supporting balancing selection in both malaria-endemic populations as well as in European populations (Bigham et al. 2018; Ko et al. 2011).

Candidate adaptive alleles driving balancing selection at *GYPA* are the ancestral alleles rs7682260-G and rs7687256-C, and derived alleles rs7682260-A and rs7687256-T, which encode the M and N antigens on GYPA, respectively (Bigham et al. 2018; Ko et al. 2011). Alleles rs7682260-A and rs7687256-T result in missense changes p.Ser20Leu and p.Glu24Gly, respectively. In addition, the glycophorin locus (spanning genes *GYPE, GYPB,* and *GYPA*) is known to undergone copy number variation and gene conversion in the human population (Algady et al. 2018; Leffler et al. 2017; Suktitipat et al. 2014). A structural variant called DUP4*,* likely the result of a series of six copy number changes at the glycophorin locus, generates a GYPB-GYPA fusion protein (Leffler et al. 2017). The DUP4 allele is under positive selection in East African populations, likely conferring protection against *P. falciparum* infection by altering the parasite's ability to interact with erythrocyte surface receptors (Algady et al. 2018). Differential signals of positive selection also have been discovered across Orang Asli populations from peninsular Malaysia in the immune response interleukin genes *IL3* and *IL4*, as well as genes *HMOX1*, *LTA*, *TNF*, *CDH13, NOS2*, *IRF1*, and *FAS,* supporting the hypothesis that these groups have undergone convergent adaptation to counteract malaria (Liu et al. 2015). Fine mapping of candidate alleles within *IL3* and *IL4* genes that are under positive selection in the Che Wong population identified top candidate SNPs rs40401, rs2243250, and rs2070874 (Liu et al. 2015). The rs40401-T allele of *IL3*, which causes a missense change c.79C>T(p.Pro27Ser), is observed at high frequency in the Che Wong, as well as in Ghanaian, Nigerian, and African American populations (Liu et al. 2015). Furthermore, epidemiological studies found a 33% protective effect from recurrent malaria episodes in Ghanaians (Meyer et al. 2011). Top candidate alleles of *IL4* (rs2243250-T and rs2070874-T) lie within the promoter and are associated with higher anti-malarial antibody levels (Liu et al. 2015). The only other candidate alleles identified were in the Jakun population in the heme oxygenase regulator *HMOX1* gene (rs2071748-A and rs8139532-G) (Liu et al. 2015), both were previously found to be associated with cerebral malaria protection (Sambo et al. 2010). Altogether, the above examples from diverse populations demonstrate that malaria has been a robust evolutionary pressure resulting in localized adaptations in immune response.

Infectious Disease Adaptations: HIV-1

Selection mapping also can be used to identify risk alleles to modern pathogens that were not the past selective agent (Karlsson et al. 2014). This is particularly relevant given the high level of redundancy of cellular and molecular protective activity within the immune system (Nish and Medzhitov 2011). This evolved redundancy arms the immune system with surplus compensatory mechanisms, for example, when loss of function variants disrupt certain immune response pathways (Fischer and Rausell 2016). In this way, host immune response utilizes existing genomic and molecular variation to respond to novel pathogens. Favored host genetic variation from past selective pressures from ancient pathogens can influence host immune resistance or susceptibility phenotypes in the face of novel pathogens. One well-studied immune locus under natural selection in human populations is the gene *CCR5,* which encodes a protein expressed on the cell surface of white blood cells (Libert F 1998; Novembre et al. 2005; Stephens et al. 1998). A 32 base-pair deletion (Δ32) at this locus, *CCR5Δ32*, is at high frequencies and under selection in Europe and western Asia (Novembre et al. 2005). Interestingly, the *CCR5Δ32* allele confers resistance to infection by Human Immunodeficiency Virus-1 (HIV-1) by abrogating expression of the CCR5 chemokine receptor exploited by HIV-1 for entry into CD4+ T cells (Liu et al. 1996; Samson et al. 1996). The HIV-1 lentivirus evolved less than 100 years ago in Africa from simian immunodeficiency virus, and is therefore not the selective pressure that previously acted on the *CCR5* locus (Sharp et al. 2001). Through the reconstruction of historical events, scholars hypothesize a different pathogen, such as smallpox, drove the signature of selection observed at this locus (Galvani and Slatkin 2003). The story of the *CCR5Δ32* allele demonstrates how previously adaptive genetic variation can influence current susceptibility to modern infectious diseases. Further, it substantiates the use of selection mapping for identifying genes and/or regulatory elements that harbor resistance and/or susceptibility alleles to modern pathogens.

Infectious Disease Adaptations: West Nile Virus

West Nile Virus (WNV) is second modern pathogen for which susceptibility alleles have been identified by prioritizing genes with strong signatures of natural selection (Bigham et al. 2011; Lim et al. 2009; Yakub et al. 2005). WNV is responsible for highly variable infection outcomes across populations wherein infected individuals may be asymptomatic, present with West Nile fever, or develop potentially fatal West Nile neuroinvasive disease (Bigham et al. 2011). A case-control association study informed by selection mapping to identify host genetic susceptibility factors to WNV infection among Europeans found strong associations at intronic SNPs in *IRF3* (rs2304207-C allele) and *MX1* (rs7280422-G allele) with symptomatic WNV infection, as well as in *OAS1* (rs34137742-T allele) with severe infection (Bigham et al. 2011). This research illustrates how selection mapping can be used to prioritize candidate genes for case-control association testing to localize susceptibility alleles.

Climatic Adaptations: High altitude

High altitude, defined as regions lying above 2,500 meters, represents one of the most marginalized and harsh environments inhabited by humans (Niermeyer et al. 2001). There are three high-altitude regions where humans have resided for millennia: the Tibetan Plateau, the Andean Altiplano, and the Semien Plateau of Ethiopia (Beall 2013). Populations from these regions display physiological and genetic adaptations to overcome reduced oxygen levels (Bigham and Lee 2014; Scheinfeldt and Tishkoff 2013; Simonson et al. 2012). Genome-wide and candidate gene research has identified compelling evidence for selection on the hypoxia-inducible factor (HIF) pathway and hypoxia-related genes (Alkorta-Aranburu et al. 2012; Beall et al. 2010; Bigham et al. 2010; Scheinfeldt et al. 2012; Simonson et al. 2010; Yi et al. 2010). Candidate alleles in *EGLN1, EPAS1,* and *PRKAA1* are suggested to underlie altitude-adaptive phenotypes including hemoglobin concentration, birth weight, and maximal oxygen consumption during exercise (Beall et al. 2010; Bigham et al. 2010; Bigham et al. 2014; Brutsaert et al. 2019; Eichstaedt et al. 2017; Simonson et al. 2010; Yi et al. 2010). For example, high frequency candidate alleles in *EGLN1,* rs12097901-C and rs186996510-G (resulting in missense changes c.380G>C(p.Cys127Ser) and c.12C>G(p.Asp4Glu), respectively), are associated with hemoglobin concentration in Tibetan highlanders (Lorenzo et al. 2014). In addition to high altitude, human adaptation to hypoxic conditions may have occurred in the Bajau deep-sea divers of Southeast Asia. An intronic SNP in *PDE10A* (rs3008052) that is found within a peak positive selection signal in this population, shows significant association with enlarged spleen size (Ilardo et al. 2018). The discovery and functional investigation of adaptive alleles that mitigate hypoxia has translational implications for understanding and treating the disease biology caused by chronic hypoxia, including cases of ischemic heart disease, cerebrovascular disease, anemia, chronic obstructive pulmonary disease, and pulmonary hypertension.

Climatic Adaptations: Cold climate

Another extreme environmental pressure that challenges human survival is cold climate. Siberia is the coldest region inhabited by humans with an average annual temperature roughly 5°C. Archaeological data estimate first human arrivals into the region around 45-40 kya (Goebel 1999; Goebel et al. 2008). Although cultural adaptations can help buffer humans from the physiological stressors of extreme cold, biological adaptations, such as increased levels of thermogenic brown adipose tissue (Daanen and Van Marken Lichtenbelt 2016), are necessary to withstand continued sub-zero temperatures. Genome-wide selection scans on indigenous populations from Siberia using iHS, XP-EHH, and PBS discovered a 3 Mb region on chromosome 11, containing ~79 genes with strong positive selection signals (Cardona et al. 2014; Clemente et al. 2014). While genes involved in the regulation of energy and metabolism (*CPT1A, LRP5,* and *THADA)* as well as the smooth muscle contraction, *PRKG1,* were identified in this scan, candidate alleles driving this signal we not identified (Cardona et al. 2014). However, further selection mapping work using whole genome sequencing data from Canadian and Greenland Inuit populations identified a nonsynonymous variant (c.1436C>T (p.Pro479Leu)) at the rs80356779 SNP in *CPT1A*, which is essential for fatty acid oxidation (Clemente et al. 2014). Given the rs80356779-T allele also is observed at a very high frequency in the Northeast Siberian populations from Cardona et al., 2014, it is a top candidate causal allele driving the strong signal of positive selection from cold exposure (Clemente et al. 2014). Functional evidence shows that the *CPT1A* c.1436C>T (p.Pro479Leu) missense variant decreases fatty-acid oxidation and ketogenesis, decreases mitochondrial inhibition of malonyl-CoA on fatty acid β-oxidation, and is associated with increased high-density lipoprotein cholesterol with reduced adiposity in the Alaskan Yup'iks population (Clemente et al. 2014; Greenberg et al. 2009; Lemas et al. 2012). Together these data suggest an adaptive metabolic role based on gene-environment interaction (Clemente et al. 2014; Collins et al. 2010). Functional models seeking to recapitulate an adaptive function of this allele based on different environmental parameters of temperature and diet will be important to shed light on gene-environment interaction hypotheses. In addition, further fine mapping and functional investigation of other candidate adaptive alleles underlying cold adaptation in humans may help to uncover regulatory mechanisms of human adipogenesis, with potential to inform medical research of certain metabolic diseases, such as obesity.

In addition to metabolic adaptations, it is widely hypothesized that human body size and body proportions have adapted to temperature exposure for proper thermoregulation. Signatures of positive selection surrounding height loci have been identified (Amato et al. 2011; Voight et al. 2006; Wu et al. 2012). One interesting selection candidate for short stature is *GDF5,* which encodes the growth differentiation factor 5 protein involved in skeletal, joint, and central nervous system development (Buxton et al. 2001; Francis-West et al. 1999). Among East Asian populations, *GDF5* shows evidence of positive selection and variation in this gene is associated with both decreased stature and osteoarthritis risk (Wu et al. 2012). Using a CMS approach, a strong selection peak was observed surrounding the 3' UTR of *GDF5* (Capellini et al. 2017). Functional fine-mapping analysis of this signal through a series of transgenic mouse reporter assays and gene rescue experiments led to the discovery of a novel growth enhancer of *GDF5*, named *GROW1*. A search for candidate alleles driving the positive selection signal surrounding *GROW1* identified the regulatory rs4911178-A allele, which is associated with short height, observed at a high frequency in non-African populations, and is present in Neanderthal and Denisovan genomes (Capellini et al. 2017). These findings highlight the potential of selection mapping in unmasking adaptive variants in regulatory elements of the genome.

Climatic Adaptations: Ultraviolet radiation

Changes in exposure to ultraviolet radiation (UVR) accompanied changes in temperature as humans migrated out of Africa. The continuous spectrum of human skin color across the globe strongly maps onto the latitudinal distribution of UVR intensity (Chaplin 2004; Elias and Williams 2013; Jablonski and Chaplin 2000, 2013). Darker pigmentation is caused by an increased production of eumelanin in the dermis, which acts to absorb UV light and scatter it. In this way, melanin protects against dermal damage as well as UV-A photolysis of folate—the breakdown of which can lead to neural tube defects (Brenner and Hearing 2008; Kaidbey et al. 1979; Schmitz et al. 1995). Conversely, in regions of low UVR, lighter pigmentation is hypothesized to have been selectively favored for sufficient endogenous synthesis of Vitamin D, minimizing risk of Vitamin D deficiency (Loomis 1967).

Selection scans for candidate alleles underlying skin pigmentation adaptation have uncovered a handful of candidate genes that differ in signals between human populations, supporting evolutionary convergence (Alonso et al. 2008; de Gruijter et al. 2011; Izagirre et al. 2006; Lamason et al. 2005; Lao et al. 2007; Myles et al. 2007; Norton et al. 2007; Voight et al. 2006). For instance, evidence for positive selection in African populations has been found at pigmentation-associated genes *ASIP, MFSD12, DDB1, HERC2, TMEM138, OCA2, TP53BP1*, *DCT*, *TYRP1*, *EGFR,* and *DRD2* (Alonso et al. 2008; Crawford et al. 2017b; Izagirre et al. 2006; Lao et al. 2007), whereas *MC1R* is under strong functional constraint by purifying selection (John et al. 2003, Harding et al. 2000, Rana et al. 1999). In European populations, the strongest signals of positive selection have been found in *SLC24A5*, *TYRP1*, *KITLG*, *MATP,* and *TYR* (Lamason et al. 2005; Lao et al. 2007; Norton et al. 2007). *OCA2*, *DCT*, *KITLG*, *TYRP1*, *EGFR,* and *DRD2* have been suggested as candidates for positive selection in Asian populations (Alonso et al. 2008; Lao et al. 2007; Myles et al. 2007). In indigenous American populations, 14 candidate genes under positive selection have been identified, including *SLC24A5, MATP, OPRM1,* and *EGFR* (Quillen et al. 2012). Within these candidate genes, several candidate pigmentation alleles have been fine-mapped and functionally annotated. Identified dark pigmentation alleles include: rs6058017-G (*ASIP*); rs6510760-A, rs10424065-T, rs112332856-C, and rs56203814-T (*MFSD12*); rs11230664-C (*DDB1*); rs7948623-T (*TMEM138*); rs1800404-C (*OCA2*); rs4932620-T and rs6497271-A (*HERC2*) (Crawford et al. 2017b). Associated light pigmentation alleles include: rs1426654-A (*SLC24A5*); rs16891982-G (*MATP*), rs6058017-A (*ASIP*); rs1800414-G and rs1448484-T (*OCA2*); rs1042602-A and rs1126809-A (*TYR*) (Lamason et al. 2005; Lao et al. 2007; Norton et al. 2007).

Recently, it has been suggested that light pigmentation alleles in modern human populations were present in the hominin lineage before modern humans evolved (Crawford et al. 2017b), supporting a model of selection on standing variation instead of convergent *de novo* variants in northern latitude groups. Significant associations with skin pigmentation variants in *SLC24A5, MFSD12, DDB1, TMEM138, OCA2,* and *HERC2* were observed among diverse African populations (Crawford et al. 2017b). The well-known light-pigmentation allele in *SLC24A5,* rs1426654-A, was introduced via gene flow from non-African groups into East Africa (Crawford et al. 2017b). The dark pigmentation alleles rs6510760-A, rs10424065-T, rs112332856-C, and rs56203814-T (*MFSD12*), rs11230664-C (*DDB1*), rs7948623-T (*TMEM138*), rs1800404-C (*OCA2*), rs4932620-T and rs6497271-A (*HERC2*) are identical by descent between African populations and populations from South Asia and Australo-Melanesia (Crawford et al. 2017b). These fascinating results demonstrate the evolutionary complexity of human pigmentation as the result of both shared ancestry and derived adaptive variation. Together, the extensive body of research on the genetics of human pigmentation variation represents one of the greatest selection mapping efforts across diverse human populations.

Dietary Adaptations

The advent of agriculture brought about new selective pressures on human populations in response to changes in diet. A prominent example of dietary adaptation is lactase persistence in diverse pastoralist populations. A 3.2 Mb region defining a strong signature of selection in these groups contains the lactase gene, *LCT* (Bersaglieri et al.

2004; Peter et al. 2012; Tishkoff et al. 2007; Voight et al. 2006). The protein encoded by this gene is responsible for breaking down lactose found in dairy products into glucose and galactose, thus enabling digestion in the intestines (Arribas et al. 2000). In northern European populations, two intronic SNPs upstream in *MCM6* (within the regulatory region of *LCT*) are under strong positive selection (rs4988235-T and rs182549-A) and regulates *LCT* expression (Enattah et al. 2002). Evidence supports recent *de novo* selection for these regulatory alleles (Bersaglieri et al. 2004; Peter et al. 2012; Tishkoff et al. 2007; Voight et al. 2006). Among African pastoralist populations from Tanzania, Kenya, and Sudan, three derived alleles (rs145946881-C, rs41525747-G, and rs41380347-G) in intron 13 of *MCM6* enhance *LCT* gene expression (Tishkoff et al. 2007). Lactase persistence thus represents another example of convergent evolution in humans from a shared cultural adaptation of adult milk consumption in northern Europe and East Africa.

Additional examples of dietary pressures driving human evolution include increased starch and fat consumption and exposure to arsenic drinking water. Copy number variation of the salivary α-amylase gene (*AMY1*) is under positive selection among populations with dietary histories of high-starch consumption (Perry et al. 2007). Greater *AMY1* copy number corresponds to increased salivary amylase protein expression, which has led to the hypothesis that increased AMY1 protein levels enhances starch digestion and/or plays a role in reducing risk of intestinal disease (Perry et al. 2007). However, conflicting evidence exists for the biochemical role of AMY1 in starch digestion, and some scholars argue AMY1 expression profiles across tissues need to be considered when hypothesizing an evolutionary role (Fernandez and Wiley 2017). Selection mapping in other populations points to additional parts of the genome involved in starch digestion. Signals of positive selection in the maltase-glucoamylase gene, *MGAM,* involved in the final stages of starch digestion, have been identified from comparative analysis of ancient Andean genomes (~7,000 years ago) and contemporary lowland and highland Peruvians (Lindo et al. 2018). The top candidate adaptive allele in *MGAM* is the C allele of intronic SNP rs77768615—a locus annotated with promoter-associated histone modifications in intestinal cells and enhancer-associated histone marks in cells of the duodenum, stomach, and intestines (Lindo et al. 2018). The findings suggest an adaptive shift to a starch-rich diet in these Peruvian populations consistent with archaeological evidence of intensive potato processing and consumption in the region (Haas et al. 2017; Rumold and Aldenderfer 2016).

Aside from food pressures, exceptionally high levels of the carcinogen arsenic have permeated the water source of the Colla population in the Puna region of Northwest Argentina for thousands of years (Eichstaedt et al. 2015). Candidate adaptive alleles rs1046778-C and rs7085104-G, located within the arsenic methyltransferase gene, *AS3MT,* under positive selection in the Colla population, are associated with decreased expression of *AS3MT* and lower excreted monomethylarsonic acid levels in urine (Eichstaedt et al. 2015; Schlebusch et al. 2013). Given that high levels of monomethylarsonic acid are associated with arsenic-related illness, this data suggests adaptive variants altering AS3MT expression underlie adaptation to high arsenic environments in the Colla (Eichstaedt et al. 2015). Altogether, natural selection in response to dietary changes further exemplifies the diversity of adaptive genetic variation that underlies human adaptation.

Cardiovascular Health

Selection mapping also is useful for the study of complex diseases, such as cardiovascular disease. As discussed earlier, when a trait is subjected to positive directional selection, linked variation rises in frequency along with the causative adaptive allele(s). In some cases, linked deleterious variants may exist on the adaptive haplotype(s). If the fitness benefits of the adaptive allele(s) outweigh the costs of the accompanied deleterious variants, those variants may persist in a population through time. In addition, beneficial (or neutral linked) alleles in a past environment may become deleterious alleles in a novel or changing environment, resulting in a gene-environment mismatch or interaction (Corbett et al. 2018). A powered phenotype-genotype association analysis of the top candidate loci under positive selection in a Mexican American cohort identified candidate risk alleles for dyslipidemia—a highly prevalent condition among Amerindian-origin populations that results in elevated lipid levels in the blood (Ko et al. 2014). Among the top associated hits were well-known obesity and cardiovascular disease genes, previously associated risk alleles rs3135506-A and rs662799-C of *APOA5,* and novel risk alleles rs139961185-A of *SIK3* and intergenic rs28680850-A (nearby *LPL*) (Ko et al. 2014), potentially pointing to a gene-environment mismatch underlying disease risk in these populations. Similarly, it is well-described that high-altitude Andean populations, unlike other groups adapted to high altitude, exhibit polycythemia—an extreme production of red blood cells that can hinder blood flow to tissues from increased viscosity (León-Velarde 2003; Monge 1942). Top candidate alleles within haplotypes showing strong signatures of positive selection in a high-altitude Andean cohort—*BRINP3* (rs11578671-A), *NOS2* (rs34913965-T), and *TBX5* (rs10744822-C)—were associated phenotypic variation in cardiovascular health (Crawford et al. 2017a). The authors propose that high-altitude adaptation in Andeans involved changes in cardiovascular tolerance to polycythemia, instead of its reduction (Crawford et al. 2017a). Together, this work emphasizes how the application of selection mapping can aid in the discovery of risk alleles for complex disease, particularly when disease expression depends upon gene-environment interactions. Subsequent functional investigation of identified human candidate alleles for both adaptive traits and disease risk is necessary to determine the molecular mechanism, build models of human evolution, and facilitate medical research.

2.3.3 Genetic Variation Underlying Human-Specific Adaptations

From obligate bipedalism and hairlessness to disproportional expansion of the prefrontal cortex and language acquisition, adaptive traits derived in the human lineage have been shaped by selective pressures since the divergence of the last common ancestor of humans and chimpanzees/bonobos. Understanding how human-specific traits develop given a comparable set of genes to other closely related primates is an outstanding question yet to be resolved. Genetic variation exclusive to the human genome with predicted functional consequence serves as a key starting point for uncovering the genetic basis of human-specific adaptations. Functionally important lineage-specific genetic variation is often found in regions of high evolutionary conservation across other species (Levchenko et al. 2018). Recent advancements in sequencing technologies coupled with the increasing availability of high-quality reference genomes for diverse primate species, comparative genomics is at the cutting edge of discovery for human lineage-specific adaptive genomic variation (Kronenberg et al. 2018).

Human-specific genetic variants affect genes involved in immune response, metabolism, olfaction, and brain development (Bitar et al. 2019). Most identified human-specific variation is non-coding regulatory and structural variation, but there are also examples of positive selection resulting in amino acid changes associated with derived human adaptive phenotypes (Haygood et al. 2010; Kronenberg et al. 2018; Levchenko et al. 2018). Distinguishing between human-specific sequences that are causative for, rather than simply associated with, phenotypic variation remains a major challenge. In some cases, human disease genetics can provide insight into functional roles of candidate adaptive loci if their disruption is linked to specific clinical phenotypes (Cooper and Kehrer-Sawatzki 2011; O'Bleness et al. 2012; Sikela 2006). The following discussion will highlight the growing literature of protein-coding and non-coding human-specific genetic features linked to adaptive functions, with an emphasis on findings implicated in the human brain (Figure 2.4).

Protein-coding substitutions

Although adaptive coding substitutions unique to humans are scarce, a few noteworthy cases have been well studied. A prominent example is the detection of selection at primary microcephaly genes in primate genomes (Evans et al. 2006; Guernsey et al. 2010; Montgomery et al. 2011; Pervaiz and Abbasi 2016). These genes were first explored for evidence of human brain evolution after pathogenic variants were discovered in human versions of these genes. The pathogenic variants altered the proliferative capacity and mitotic events of cortical neural progenitor cells during development, causing primary microcephaly. Interestingly, selection studies provided evidence for positive selection in the primate lineage in the primary microcephaly genes *CDK5RAP2*, *CENPJ*, *CEP152*, and *WDR62* (Evans et al. 2006; Guernsey et al. 2010; Montgomery et al. 2011; Pervaiz and Abbasi 2016). *MCPH1* and *ASPM* show signals of positive selection in the great ape lineage, with *ASPM* showing further evidence of positive selection in the human lineage (Evans et al. 2004a; Evans et al. 2004b; Kouprina et al. 2004; Wang and Su 2004; Zhang 2003). Another gene implicated in the regulation of neural development, *REST*, shows strong positive selection in humans prior to the divergence of modern human and Neanderthal lineages (Mozzi et al. 2017). A missense variant in *NOVA1* that evolved more recently in modern humans and that is absent in the genomes of Neanderthals and Denisovans is implicated in synaptogenesis and glutamatergic signaling in human cortical neurons via regulation of splicing (Trujillo et al. 2021). Consistent with the observation of a medial-lateral expansion of the cerebellum in ape evolution (Barton and Venditti 2014), signals of positive selection have also been detected in human *AHI1*, a gene associated with the neurodevelopmental disorder Joubert syndrome that causes malformation of the cerebellar vermis and brainstem (Cheng et al. 2012; Doering et al. 2008; Ferland et al. 2004; Harrison and Montgomery 2017).

Recent work using the branch-site-model CODEML and aBSREL tests for selection in over 3,000 gene regulatory factors in the human genome and their orthologous sequences in 26 non-human primate genomes identified five regulatory proteins under positive selection in the human genome as strong candidates for adaptation and/or speciation events (Jovanovic et al. 2021). These included four KRABzinc-finger containing proteins, *PRDM9*, *ZNF626*, *ZNF806*, and *ZNF860*, which make up the largest class of gene regulatory factors in the human genome, and *MAMLD1* (Jovanovic et al. 2021). Using early hominin genomes, the timing of selection was dated prior to divergence with the lineages that led to Neanderthals and Denisovans (Jovanovic et al. 2021). *PRDM9*, involved in meiotic recombination, and *MAMLD1*, which plays a role in sex determination and male gonad development, are candidates for driving speciation in humans (Daub et al. 2015; Fukami et al. 2006; Jovanovic et al. 2021; Schwartz et al. 2014). Less is known functionally about *ZNF626*, *ZNF806,* and *ZNF860*. *ZNF626* is highly expressed in the middle temporal gyrus involved in language processing and has been implicated in bipolar disorder and posttraumatic stress disorder (Acheson and Hagoort 2013; Jovanovic et al. 2021; Stein et al. 2016). Variants in *ZNF806* have been associated with Alzheimer's disease and tardive dystonia (Jovanovic et al. 2021; Parcerisas et al. 2014) whereas *ZNF860* is associated with cancer and early-onset type 2 diabetes (Jovanovic et al. 2021; Pan et al. 2019). In addition to natural selection on protein regulatory factors, significant discoveries have also implicated regulatory elements and non-coding regions of the genome in human adaptation.

Non-coding regulatory variation

Early work comparing human and chimpanzee proteomes documented the high similarity in amino acid sequences between the two. In 1975, Mary-Claire King and Allan Wilson hypothesized in a seminal study that most of the genetic variation underlying phenotypic differences between humans and chimpanzees is non-coding regulatory sequence (King and Wilson 1975). Since then, great efforts to discover regulatory variation contributing to morphological differences led to the characterization of several human accelerated regions (HARs)—short DNA sequences of just a couple hundred base-pairs found in noncoding, highly conserved regions across vertebrates yet exhibit high substitution rates in the human genome (Bird et al. 2007; Bush and Lahn 2008; Gittelman et al. 2015; Hubisz and Pollard 2014; Lindblad-Toh et al. 2011; Pollard et al. 2006; Prabhakar et al. 2006). Given that noncoding regions with high conservation across species implicate functional significance, several HARs have been linked to regulation of gene expression underlying human traits (Levchenko et al. 2018). Interestingly, loci associated with schizophrenia are enriched near several HARs, and variation in HARs have been implicated in autism risk (Doan et al. 2016; Xu et al. 2015).

HAR1, which encodes the functional lncRNAs *HAR1A* and *HAR1B*, has acquired human-specific substitutions predicted to influence lncRNA secondary structure stability and thus lncRNA function (Beniaminov et al. 2008; Levchenko et al. 2018; Pollard et al. 2006). Human-specific accelerated evolution of the HAR *HACNS1* created a gain-of-function developmental enhancer in limb development, proposed to contribute to human-specific adaptations in hand dexterity, limb and digit patterning, and hindlimb morphology of bipedalism (Prabhakar et al. 2008). *AUTS2, NPAS3,* and *FZD8* harbor several HARs within their noncoding genomic sequence act as enhancers, particularly in the central nervous system, wherein human-specific substitutions are hypothesized to modify transcription factor binding and influence neural development (Boyd et al. 2015; Kamm et al. 2013; Levchenko et al. 2018; Oksenberg et al. 2013). *CUX1*, *PTBP2*, and *GPC4* interact with HAR-containing enhancers (HAR426, HAR169, and HAR1325, respectively) involved in neural development and/or associated with neurodevelopmental disorders (Bird et al. 2007; Doan et al. 2016; Levchenko et al. 2018; Prabhakar et al. 2006). The rising functional literature on HARs is uncovering the extent of their role in shaping human adaptive biology and demonstrate how accelerated genetic adaptations can influence modern disease risk.

In addition to accelerated substitution changes, human-specific insertions and deletions that affect conserved regulatory sequences have been identified. For example, a 3 kilobase deletion of a conserved forebrain subventricular zone-specific enhancer of *GADD45G* has been proposed to augment cortical expansion in humans (McLean et al. 2011). Human-specific deletion of the conserved enhancer for penile spine development occurred during human evolution likely from relaxed selective pressures (McLean et al. 2011). Positive selection has also led to tandem repeat expansion and single nucleotide substitutions in the promoter of *PDYN* in humans, which is otherwise highly conserved among non-human primates, and leads to increased expression of the opioid polypeptide hormone prodynorphin in human cells (Rockman et al. 2005).

Structural variation

While King and Wilson's hypothesis continues to hold true (King and Wilson 1975), we know substantially more about the role of genomic structural variation—that concurrently affects coding and noncoding elements—in human and non-human primate evolution. Genomic structural variation is a major driving force of evolution. Novel genetic sequence or altered genomic relationships from genomic structural changes create new genes, destroy established genes, and/or alter gene expression, giving rise to phenotypic diversity (Lupiáñez et al. 2015). The advent of highly accurate long read sequencing, used in combination with BAC-based and short read sequencing, has provided high quality genome annotation and has led to unprecedented mapping of complex structural variation across human and non-human primate genomes (Kronenberg et al. 2018). According to recent comparative genomics analyses, segmental duplications are rapidly and disproportionately accumulating in ape genomes (Cheng et al. 2005; Kronenberg et al. 2018; Marques-Bonet et al. 2009; Sudmant et al. 2013). Potentially, initial expansion of segmental duplications in the ape lineage that was adaptive and maintained has made ape genomes prone to further structural events due to the increased repetitive sequence, and has acted to accelerate structural changes and acquisition of variants of large effect during evolution (Kronenberg et al. 2018). In humans, this genomic evolution has led to over 17,000 fixed human-specific structural variants, with 90 predicted to disrupt genes and 643 influencing regulatory sequence (Kronenberg et al. 2018). The most functionally studied category of human-specific structural variants have been human-specific duplicate genes.

A major source of adaptive evolution is gene duplication. However, resolving the evolutionary mechanism of the retention of duplicate genes is very difficult. Several models have been proposed that invoke either neutral evolution or positive selection, and which model best explains observed sequence variation is highly debated (Bergthorsson et al. 2007; Kondrashov and Kondrashov 2006; Kondrashov et al. 2002; Ohno 1970; Otto and Yong 2002; S 1970). The functional fate of duplicate genes is influenced by mutation events in copies, which can involve gain-of-function mutations that contribute to divergence of paralog function (neofunctionalization), loss-of-function mutations that divvy up ancestral gene functions among duplicates (subfunctionalization), mutations that result in complete loss of a functional protein (pseudogenization), or retention of ancestral function through achieving dosage-balance (Force et al. 1999; Gout et al. 2009; Gout et al. 2010; Ohno 1970; Papp et al. 2003; Qian et al. 2010; Rastogi and Liberles 2005). Duplicate genes can evolve via complete or incomplete duplication events, and both mechanisms have been characterized in the human genome. The former involves duplication of the complete genic sequence and in some cases leads to intragenic copy number expansion (*NBPF10, NBPF14, NBPF19, NBPF20*) (Astling et al. 2017; Heft et al. 2020). Incomplete duplication can result in a shortened version of the ancestral gene, 5' end truncation with loss of the ancestral promoter that can either result in acquisition of a new promoter or loss of expression (*FRMPD2B, GTF2IRD2P1, CORO1AP*), 3' end truncations where the ancestral promoter is retained (*SRGAP2B, SRGAP2C, SRGAP2D, NOTCH2NL*), dual-truncation of 5' and 3' ends (*HYDIN2*), and/or gene fusion (Dougherty et al. 2018). Of the duplicates that retain functionality, one can imagine how such changes can lead to altered expression dynamics and/or novel biological functions subjected to natural selection.

Several human-specific gene duplications that are fixed and/or maintained in the human population have been mapped to genes that modify brain development. Increased *NOTCH2NL* copy number in humans has been linked to enhanced proliferation of neural progenitor cells and delayed differentiation, together increasing total neuronal output and cortical size (Fiddes et al. 2018b). Similarly, the human-specific paralog *ARGAP11B* is highly expressed in outer radial glia cells and promotes their proliferation and self-renewal (Antonacci et al. 2014; Florio et al. 2015). Other human-specific duplications associated with neural development include *BOLA2, SRGAP2C, HYDIN2,* and *NBPF* (Antonacci et al. 2010; Dougherty et al. 2017; Fossati et al. 2016; Keeney et al. 2014; Nuttle et al. 2016). Remarkably, *NBPF* genes represent the greatest copy number expansion of any protein-coding region in the human genome, making their expansion a potentially fascinating human evolutionary story. Despite this, little research has been carried out to determine their molecular function or to confirm their role in human evolution. This is largely the result of the difficulty working with highly identical paralogous sequence. Greater efforts to functionally study the suite of human-specific duplicate genes will be important for understanding the extent to which structural variation has shaped human adaptation.

Gene expression differences

Comparative *omics* approaches often lead to the discovery of species-specific molecular phenotypes with cell-type specificity (Florio et al. 2018). These include transcriptomics (RNA), proteomics (protein), and interactomics (molecular interactions) (Manzoni et al. 2018). Research studying human variation and disease is growing towards the integration of different omics techniques to concurrently evaluate different aspects of molecular biology (Karczewski and Snyder 2018). Currently, most efforts to identify human-specific molecular changes contributing to adaptive phenotypes have used comparative transcriptomics. Single-cell RNA sequencing analysis on cortical tissue from human and chimpanzee has identified 383 upregulated genes in human radial glia and 219 upregulated genes in human excitatory neurons (Kronenberg et al. 2018). Fifteen human-specific genes have preferential expression in cortical neural progenitor cells during development, and several have evolved distinct expression profiles from their ancestral paralogs (Florio et al. 2018). Upregulation of the plasma membrane-bound proteins INSR and ITGB8 in human radial glia compared to chimpanzees and macaques leads to increased activation of the PI3K-AKT-mTOR signaling that regulates the morphology and migration of human outer radial glia (Andrews et al. 2020; Pollen et al.

2019). Isoform-specific expression of membrane-bound PALMD protein in humans contributes to enhanced proliferation of outer radial glia (Kalebic et al. 2019). Differential expression of the *ZEB2* morphogenesis factor in humans versus non-human apes contributes to delayed neural epithelia to radial glia cell transition in humans with increased proliferation (Benito-Kwiecinski et al. 2021). Species differences in gene expression dynamics during neural development jointly with human-specific genetic changes highlighted above underscore the heterochrony of human adaptive neurobiology.

2.4 NECESSITY OF FUNCTIONAL VALIDATION OF CANDIDATE ADAPTIVE VARIATION

Functional investigation of candidate alleles under natural selection predicted to mediate adaptive phenotypic change is necessary to link putative adaptive biology to underlying molecular mechanisms. Except for forward genetics approaches noted above (e.g., malaria resistance alleles), most candidate adaptive alleles identified by evolutionary geneticists and molecular anthropologists have not been investigated functionally, or functional studies are limited. Myriad functional assays are currently available to investigate the contribution of high-priority candidate adaptive genetic variants to phenotypes and inform evolutionary hypotheses.

To devise an effective research strategy to test for putative adaptive roles, it is imperative to confirm a genetic mechanism of action for the candidate adaptive allele(s), identify a biological readout of the adaptive allele(s), and develop suitable model systems (Figure 2.2). Genetic mechanism refers to the allele’s functional effect on the DNA, RNA, and protein products (Marchetti et al. 2012). Genetic variants that alter the protein sequence can have consequences for the protein's function, including loss-of-function alleles that reduce (hypomorph) or completely abrogate (null) protein activity, or gain-of-function alleles that confer novel or augmented (hypermorph) protein activity (Marchetti et al. 2012). For example, functional studies suggest a gain-of-function for candidate adaptive alleles in *EGLN1* and *EPAS1* in Tibetan highlanders*,* acting to lower hemoglobin concentration (Tashi et al. 2017). Alternatively, candidate non-coding variants may impact gene expression; for instance, the fine-mapped, dark pigmentation-associated variants rs6510760-A and rs112332856-C, under selection in an upstream regulatory element of *MFSD12* in African populations, show ~4.9x decreased luciferase expression than light pigmentation-associated variants rs6510760-G and rs112332856-T (Crawford et al. 2017b).

In vitro and *in vivo* model systems offer the experimental opportunity to capture or enrich for relevant, affected biology of the candidate variant predicted to underlie physiological adaptations. Model systems include bacteria, yeast, animal models, and cell culture. Selection of appropriate model organisms (i.e., mouse, zebrafish, ferret, macaque, fruit fly, *C. elegans*) must consider the ability of that organism to recapitulate the human phenotype(s) of interest, feasibility of genetic manipulation, and generational time course. In recent years, there is a growing number of studies that have modified the genomes of animal models for functional characterization of candidate adaptive alleles to define their putative role in human evolution. For example, transgenic mouse models carrying the candidate derived regulatory allele (rs4911178-A) in the *Grow1* enhancer of *Gdf5* for human height adaptation have revealed significantly reduced long bone length compared to the ancestral allele (Capellini et al. 2017). Morpholino knockdown of *slc24a5* in zebrafish (Lamason et al. 2005) and CRISPR/Cas9 knockout of *Mfsd12* in mice (Crawford et al. 2017b)—two genes under positive selection in humans and harboring candidate pigmentation alleles—confirmed a functional role in melanogenesis. Additionally, the human-specific deletion of the conserved regulatory enhancer for penile spine development results in failure of penile spine formation when knocked out in mouse models (McLean et al. 2011). However, it is often necessary to utilize models that carry a human genetic background to investigate human biology.

Work using both monolayer human cell lines as well as human organoid model systems have offered compelling information about human adaptive biology. Human keratinocytes have been used to investigate the impact of a candidate adaptive SNP in *KITLG* under positive selection in Europeans associated with blond hair color phenotypes, revealing an altered LEF1 transcription factor binding site coupled with reduction in enhancer activity (Guenther et al. 2014). RNA-seq of cultured primary melanocytes obtained from African individuals carrying derived upstream regulatory alleles of *MFSD12* (rs6510760-A and rs112332856-C) confirmed a significantly decreased expression of *MFSD12* compared to cultured melanocytes derived from individuals carrying the ancestral alleles (Crawford et al. 2017b). Cerebral organoid models of human-specific *NOTCH2NL* gene duplicates identified a role in delayed differentiation in human neural development corresponding to increased neuronal number (Fiddes et al. 2018a; Suzuki et al. 2018). Likewise, the modification of modern human *NOVA1* to carry the ancestral allele found in Neanderthal and Denisovan genomes, creating an isoleucine-to-valine substitution, alters splicing activity and synaptic protein interactions in human organoid models (Trujillo et al. 2021). While the latter two studies highlight the utility of organoid modeling of macroevolutionary differences, one can also imagine the use of organoids for modeling candidate alleles underlying local adaptation within human populations (e.g., spleen organoids to model the candidate *PDE10A* SNP for deep-sea diving adaptation (Ilardo et al. 2018)).

2.5 CONCLUSION

Selection mapping can be informative for the broader scope of human biology, health, and disease when combined with association, fine-mapping, and functional analyses. Given the complexity and diversity of human genetic variation, environmental and cultural conditions, and gene-culture-environment interactions across geographically-defined human populations through time, it is crucial to increase the number of selection mapping studies in non-European and admixed populations in order to better understand how natural selection shapes human biology across varying ancestral genomic backgrounds (Bentley et al. 2017; Bentley et al. 2020; Popejoy and Fullerton 2016; Sirugo et al. 2019). Greater efforts must be made to increase representation within the community of scientific leaders who study human genetic variation, including beyond westernized scientists, to involve individuals from under-represented, vulnerable, and indigenous populations (Claw et al. 2018; Jackson et al. 2019). With advancements in whole genome sequencing technology, high-quality and accurate annotation of more and more primate genomes will continue to expand the landscape of mapped species-specific genetic variation. A more representative map of functional human-specific variants/genes together with functional variants under differential natural selection in human populations will have translational value for variant annotation in clinical disease cohorts which are currently barraged with hundreds of variants of unknown significance. In addition, given the limited sensitivity of current genotype-phenotype methods to polygenic traits or rare variants, selection mapping can enhance the power of these analyses to detect causal variants. Nonetheless, limitations of current selection detection methods need to be overcome to inform the genetics of complex traits, requiring improvement of both statistical inference methods and theory to better model complex soft selective sweeps and selection on structural variation. Beyond genotype-phenotype mapping, genome-wide selection mapping has the potential to enrich epigenetic, transcriptomic, proteomic, and interatomic research by pointing to functional factors that are evolutionarily conserved or adaptable.

ACKNOWLEDGEMENTS

We want to first and foremost thank all the study participants from diverse populations across the globe—and through time—who were included in the research papers reviewed here. We would like to thank Stephanie Bielas for providing helpful comments on the manuscript and Michael Bamshad for early ideas related to selection mapping.

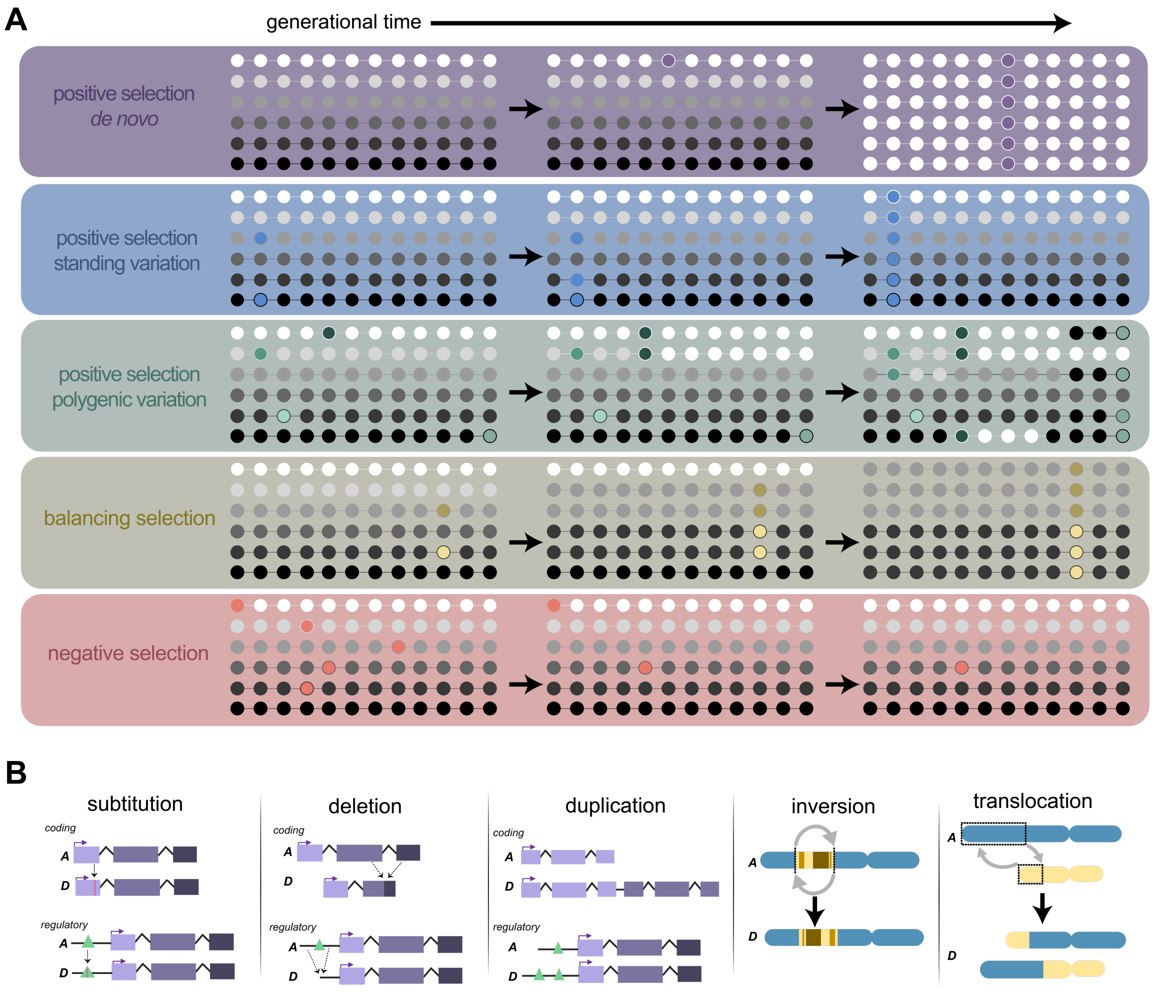


Figure 2.1. Natural selection influences genomic diversity.
(A) Theoretical basis for population genetics models of natural selection. Each dotted strand represents a haplotype. Alleles that increase or decrease in frequency over time due to selection are represented as dots in purple (positive selection on de novo variants), blue (positive selection on standing variation), green (polygenic positive selection), yellow (balancing selection), red (negative selection), and shades of grey (linked variation to causal allele). (B) Types of adaptive genetic variants that can be acted upon by natural selection. A, ancestral; D, derived.

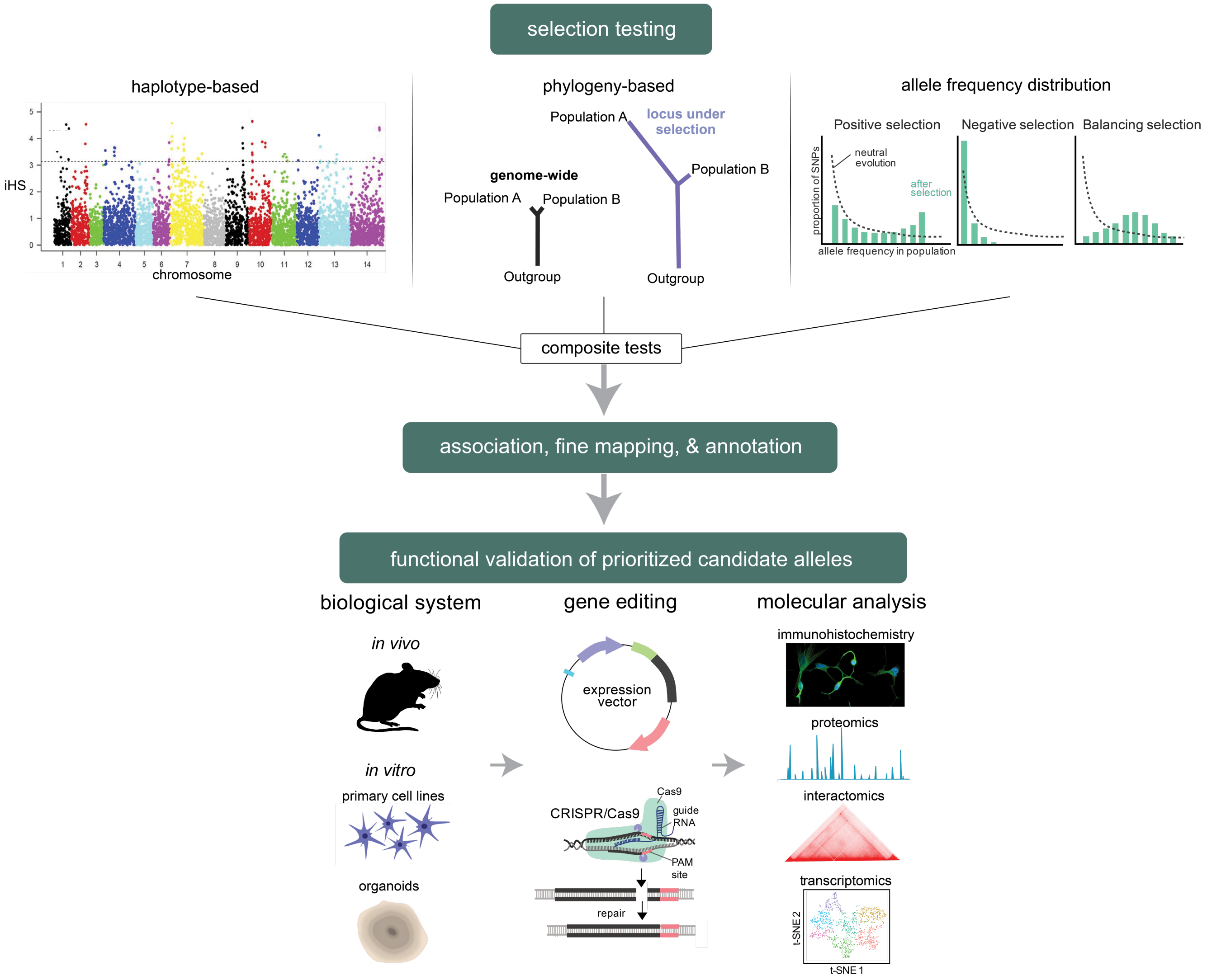


Figure 2.2. Reverse genetics selection mapping workflow.
First, perform statistical selection tests for a given dataset. Second, integrate with association mapping analyses and fine-mapping approaches to generate a list of high confidence candidate alleles for putative adaptive phenotypes. Third, identify the best research strategy to functionally investigate the candidate variation to model its biological role and inform hypotheses of adaptive human evolution.

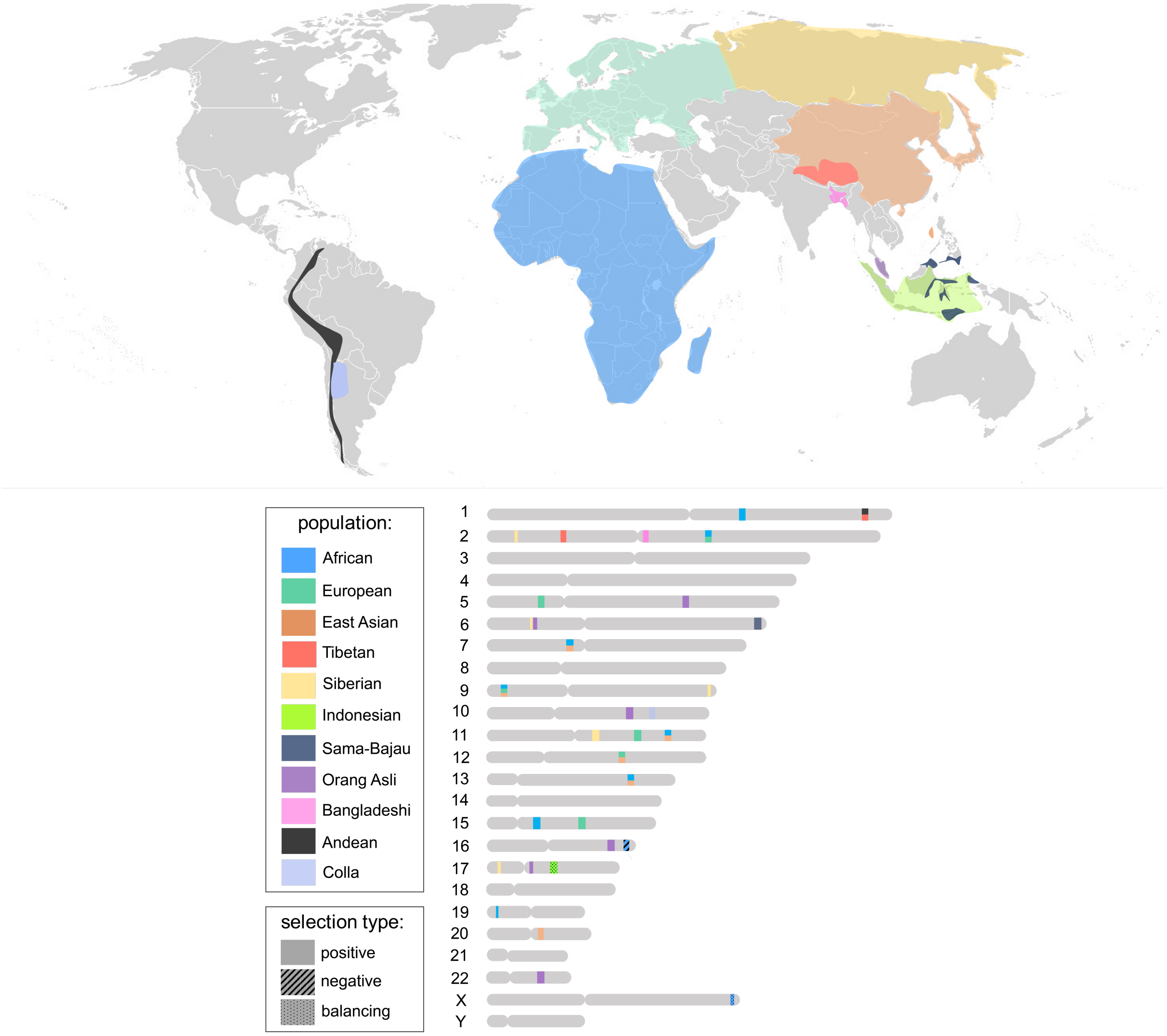
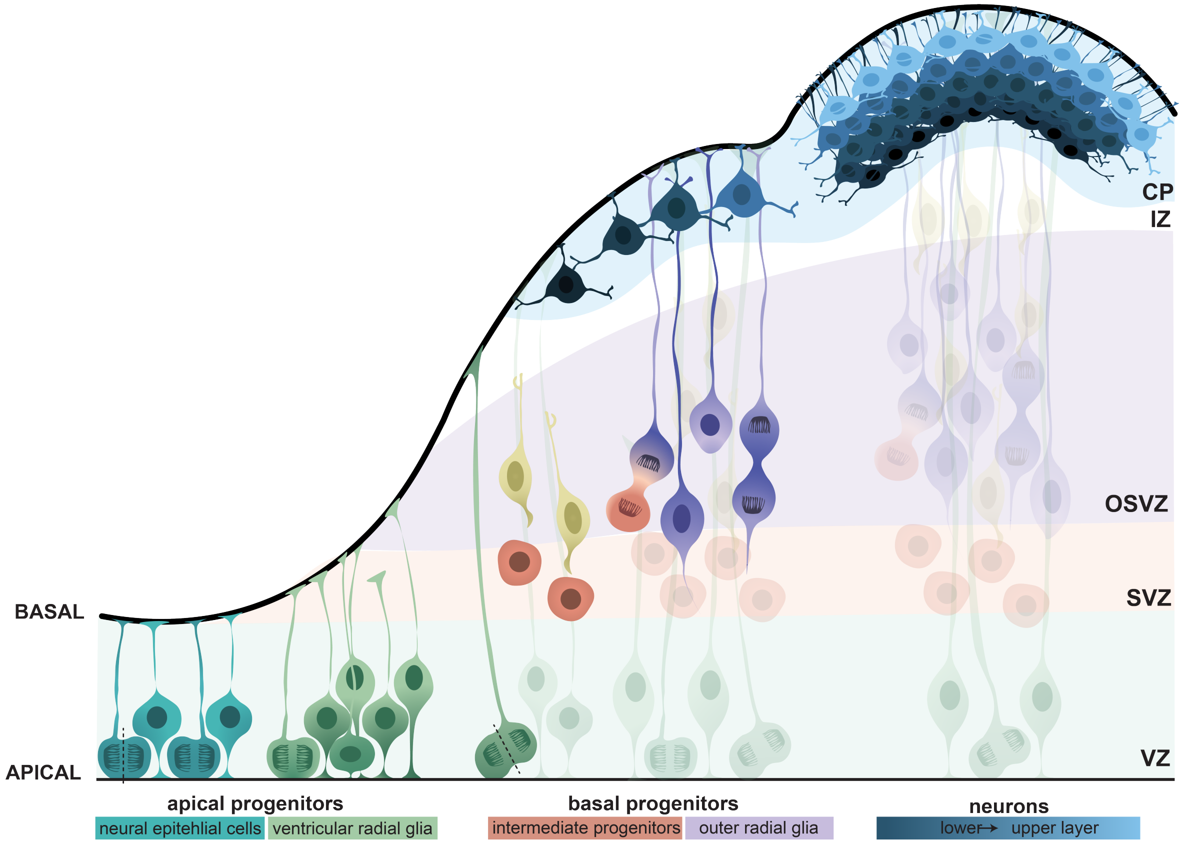


Figure 2.3. Map of discovered candidate adaptive alleles within diverse human populations. These alleles are represented by rectangular bands, color-coordinated by the general geographic region of the study population(s) and mapped to the respective broader chromosomal region. Selection type is indicated by pattern: solid (positive selection) and striped (balancing selection). Multi-colored bands indicate loci where adaptive alleles are found in multiple populations. Note: this is not a comprehensive map of all identified candidate adaptive alleles in humans.



	neural epitehlial cells	ventricular radial glia	intermediate progenitors	outer radial glia	neurons
HSS	***ASPM*** ***REST***				***NOVA1***
HAR		***FZD8***	***ΔGADD45G***		***AUTS2*** ***NPAS3***
HSD		***NOTCH2NL*** ***NBPF14/NBPF10***		***ARHGAP11B***	***SRGAP2C***
HSE	***ZEB2***			***PALMD*** ***INSR*** ***ITGB8***	

Figure 2.4. Graphic of human cortical development.
Corticogenesis through time from neural progenitor cell proliferation (left) to neurogenesis and cortical lamination (right). Table below indicates cell types corresponding to color in illustration and the human-specific genetic/molecular changes that are implicated in their cell biology. VZ, ventricular zone; SVZ, subventricular zone; OSVZ, outer subventricular zone; IZ, intermediate zone; CP, cortical plate; HSD, human-specific duplicate gene; HSE, human-specific expression; HSS, human-specific substitution; HAR, human-accelerated region.

Table 1. Summary of reviewed selection detection methods.

Type	Abbrev.	Test Name	Summary	Reference
Haplotype-based	EHH	Extended Haplotype Homozygosity	Probability and length of haplotype from a core haplotype by IBD.	Sabeti et al. 2002
	LRH	Long-Range Haplotype	The Long-Range Haplotype combines EHH with the haplotype frequency in the population to detect regions of recent positive selection, primarily hard sweeps	Sabeti et al. 2002
	WGLRH	Whole Genome Long-Range Haplotype Test	The Whole Genome Long-Range Haplotype test incorporates the LRH with patterns of LD to identify regions with long haplotypes, which may be indicative of positive selection	Zhang et al. 2006
	iHS	integrated Haplotype Score	The integrated Haplotype Score measures the ratio of EHH decay of haplotypes carrying derived alleles to those carrying ancestral alleles. This method is useful for detecting soft selective sweeps	Voight et al. 2006
	nSL	Number of Segregating Sites by Length Statistic	The Number of Segregating Sites by Length Statistic measures the ratio of EHH for haplotypes carrying derived alleles relative to those with ancestral alleles. However, unlike iHS, it incorporates segregating sites to measure distance rather than a genetic map. nSL is useful for detecting both hard and soft selective sweeps	Ferrer-Admetlla et al. 2014
	XP-EHH	Cross Population Extended Haplotype Homozygosity	The Cross Population Extended Haplotype Homozygosity incorporates the integral of EHH and compares haplotypes between two populations to detect those containing nearly fixed derived alleles. This is useful for detecting hard selective sweeps	Sabeti et al. 2007
Allele frequency	Tajima's *D*	Tajima's *D* Statistic	Tajima's *D* measures the differences between segregating sites (Θ_W) and the average nucleotide differences (π). It can be useful for detecting departure from neutrality, which may suggest balancing or positive selection. However, it is confounded by demographic conditions, such as population structure and growth	Tajima 1989
	Fu and Li's *D* and *F*	Fu and Li's *D* and *F* Statistics	Fu and Li's *D* and *F* are based on coalescence and compare the number of derived singleton variants to Θ_W and π, respectively. These statistics can suggest recent positive, purifying, or balancing selection. Fu and Li's *D* and *F* are more sensititve to detect selective sweeps acting on derived alleles compared to Tajima's *D*	Fu and Li 1993
	Fay and Wu's *H*	Fay and Wu's *H* Statistic	Fay and Wu's *H* looks for a high frequency of derived variants to determine whether the locus has undergone positive selection compared to neutral expectations	Fay and Wu 2000
	F_{ST}	Wright's Fixation Index	Wright's F_{ST} is a measure of genetic differentiation between two populations. This statistic can be used to identify loci under positive or balancing selection	Wright 1950
Phylogenetic-based	LSBL	Locus-Specific Branch Length	The Locus Specific Branch Length incorporates F_{ST} for a three-population comparison, where distances between populations are represented by branch lengths	Shriver et al. 2004
	PBS	Population Branch Statistic	The Population Branch Statistic, like LSBL, uses a three-population comparison based on F_{ST},	Yi et al. 2010

			except the values are log-transformed to incorporate phylogenetics	
	ω	d*N*/d*S*	d*N*/d*S* is the ratio of nonsynonymous to synonymous nucleotide changes which can provide evidence for positive or purifying selection across distantly-related lineages	Kimura 1977
	MKT	McDonald-Kreitman Test	The McDonald-Kreitman Test compares d*N*/d*S* ratio within a species to ratio of nonsynonymous to synonymous variants between species (p*N*/p*S*) to provide test for positive or purifying selection	McDonald and Kreitman 1991
Composite	CLR	Composite Likelihood Ratio	The Composite Likelihood Ratio measures an excess of derived alleles across several sites in a single population	Nielsen 2005
	XP-CLR	Cross Population Composite Likelihood Ratio	The Cross Population Composite Likelihood Ratio measures the differences in the excess of derived alleles following population divergence of two populations: the target population and an ancestral population	Chen et al. 2010
	CMS	Composite of Multiple Signals	The Composite of Multiple Signals Statistic combines results from $F_{ST,}$ iHS, and XP-EHH to increase power of detecting high frequency, derived alleles between populations	Grossman et al. 2010
	3P-CLR	3-Population Composite Likelihood Ratio	The 3-population Composite Likelihood Ratio is based on CLR and XP-CLR, except it uses three populations to allow for the detection of selection within each of two populations, or shared selection among the two populations before splitting from a third outgroup population	Racimo 2016
	SWIF(r)	Sweep Inference Framework (controlling for correlation)	The Sweep Inference Framework (controlling for correlation) combines a deep learning approach with a composite of the XP-EHH, iHS, and F_{ST} statistics. This method does not require a genome-wide comparison	Sugden et al. 2018

Table 2. Summary of reviewed human adaptive alleles.
illustrated in Figure 2.3. IDR, Infectious Disease Resistance; PT, Polycythemia Tolerance from High Altitude Adaptation; HA, High Altitude Adaptation; LP, Lactase Persistence; DD, Deep-Sea Diving Adaptation; HSD, High Starch Diet Adaptation; AW, Arsenic Water Adaptation; DR, Increased Dyslipidemia Risk (Gene-environment mismatch); SP, Skin Pigmentation Adaptation; CC, Cold Climate Adaptation; IDS, Infectious Disease Susceptibility; BSP, Body Size and Proportion.

Category	**Variant Type**	**Candidate Allele**	**Functional Annotation**	**Gene (within or nearby)**	**Population(s)**	**Selection Mode**	**Trait**	**References**
Immunity	Substitution	rs2814778-G	5' UTR	*DARC*	Sub-Saharan African, Malagasy	Positive	IDR	Hamblin and Di Rienzo 2000; Hamblin et al. 2002; Hodgson et al. 2014
Climatic	Substitution	rs11578671-A	Intergenic	*BRINP3*	High Altitude Andean	Positive	PT	Crawford et al. 2017a
Climatic	Substitution	rs12097901-C	Exonic: missense (p.Cys127Ser)	*EGLN1*	High Altitude Tibetan	Positive	HA	Lorenzo et al. 2014
Climatic	Substitution	rs186996510-G	Exonic: missense (p.Asp4Glu)	*EGLN1*	High Altitude Tibetan	Positive	HA	Lorenzo et al. 2014
Dietary	Substitution	rs41525747-G	Intronic	*MCM6*	African Pastoralist (Tanzania, Kenya, Sudan)	Positive	LP	Tishkoff et al. 2007
Dietary	Substitution	rs4988235-T	Intronic	*MCM6*	Northern European	Positive	LP	Enattah et al. 2002
Dietary	Substitution	rs41380347-G	Intronic	*MCM6*	African Pastoralist (Tanzania, Kenya, Sudan)	Positive	LP	Tishkoff et al. 2007
Dietary	Substitution	rs145946881-C	Intronic	*MCM6*	African Pastoralist (Tanzania, Kenya, Sudan)	Positive	LP	Tishkoff et al. 2007
Dietary	Substitution	rs182549-A	Intronic	*MCM6*	Northern European	Positive	LP	Enattah et al. 2002
Immunity	Deletion	*CCR5Δ32* (rs333)	Exonic: frameshift deletion	*CCR5*	European	Positive	IDR	Liu et al. 1996; Samson et al. 1996
Immunity	CNV	DUP4	Fusion protein: GYPB/GYPA	Affects glycoporin locus	East African	Positive	IDR	Algady et al. 2018
Immunity	Substitution	rs7687256-T	Exonic: missense (p.Glu24Gly)	*GYPA*	Sub-Saharan African, South Asian, European	Balancing	IDR	Bigham et al. 2018; Ko et al. 2011
Immunity	Substitution	rs7682260-A	Exonic: missense (p.Ser20Leu)	*GYPA*	Sub-Saharan African, South Asian, European	Balancing	IDR	Bigham et al. 2018; Ko et al. 2011
Immunity	Substitution	rs40401-T	Exonic: missense (p.Pro27Ser)	*IL3*	Orang Asli, Ghanaian, Nigerian, African American	Positive	IDR	Liu et al. 2015
Immunity	Substitution	rs2243250-T	Intergenic	*IL4*	Orang Asli	Positive	IDR	Liu et al. 2015
Immunity	Substitution	rs2070874-T	5' UTR	*IL4*	Orang Asli	Positive	IDR	Liu et al. 2015

Climatic	Substitution	rs16891982-G	Exonic: missense (p.Phe374Leu)	*MATP*	European	Positive	SP	Lao et al. 2007; Norton et al. 2007
Climatic	Substitution	rs3008052-T	Intronic	*PDE10A*	Sama-Bajau	Positive	DD	Ilardo et al. 2018
Dietary	Substitution	rs77768615-C	Intronic	*MGAM*	Andean	Positive	HSD	Lindo et al. 2018
Dietary	Substitution	rs7085104-G	Intronic	*AS3MT*	Colla	Positive	AW	Eichstaedt et al. 2015
Dietary	Substitution	rs1046778-C	Intronic	*AS3MT*	Colla	Positive	AW	Eichstaedt et al. 2015
Metabolic	Substitution	rs3135506-A	Exonic: missense (p.Ser19Trp)	*APOA5*	Mexican American	Positive	DR	Ko et al. 2014
Metabolic	Substitution	rs662799-C	Intergenic	*APOA5*	Mexican American	Positive	DR	Ko et al. 2014
Metabolic	Substitution	rs662799-C	Intergenic	nearby *LPL*	Mexican American	Positive	DR	Ko et al. 2014
Metabolic	Substitution	rs139961185-A	Intronic	*SIK3*	Mexican American	Positive	DR	Ko et al. 2014
Immunity	Substitution	rs334-T	Exonic: missense (p.Glu7Val)	*HBB*	Sub-Saharan African, Middle Eastern, South Asian	Balancing	IDR	Allison 1954; Piel et al. 2010
Climatic	Substitution	rs11230664-C	Intronic	*DDB1*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b
Climatic	Substitution	rs7948623-T	3' UTR	*TMEM138*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b
Climatic	Substitution	rs80356779-T	Exonic: missense (p.Pro479Leu)	*CPT1A*	Siberian	Positive	CC	Clemente et al. 2014; Cardona et al. 2014
Climatic	Substitution	rs80356779-T	Exonic: missense (p.Pro479Leu)	*CPT1A*	Alaskan Yup'ik	Positive	CC	Clemente et al. 2014
Climatic	Substitution	rs1042602-A	Exonic: missense (p.Ser192Tyr)	*TYR*	European	Positive	SP	Norton et al. 2007
Immunity	Substitution	rs34137742-T	Intronic	*OAS1*	European	Positive	IDR	Bigham et al. 2011
Climatic	Substitution	rs10744822-C	Intronic	*TBX5*	High Altitude Andean	Positive	PT	Crawford et al. 2017a
Climatic	Substitution	rs1800414-G	Exonic: missense (p.His615Arg)	*OCA2*	East Asian	Positive	SP	Lao et al. 2007
Climatic	Substitution	rs1800404-C	Exonic: synonymous	*OCA2*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b
Climatic	Substitution	rs1448484-T	Intronic	*OCA2*	European, East Asian	Positive	SP	Lao et al. 2007
Climatic	Substitution	rs6497271-A	Intronic	*HERC2*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b
Climatic	Substitution	rs4932620-T	Intronic	*HERC2*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b
Climatic	Substitution	rs1426654-A	Exonic: missense (p.Thr111Ala)	*SLC24A5*	European	Positive	SP	Lamason et al. 2005; Norton et al. 2007
Climatic	Substitution	rs34913965-T	Intronic	*NOS2*	High Altitude Andean	Positive	PT	Crawford et al. 2017a
Immunity	Deletion	*SLC4A1Δ27* (rs769664228)	Exonic: in-frame deletion	*SLC4A1*	Indonesian	Balancing	IDR	Jarolim et al. 1991; Paquette et al. 2015; Wilder et al. 2009
Climatic	Substitution	rs56203814-T	Exonic: synonymous	*MFSD12*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b

Climatic	Substitution	rs10424065-T	Intronic	*MFSD12*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b
Climatic	Substitution	rs6510760-A	Intergenic, regulatory	*MFSD12*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b
Climatic	Substitution	rs112332856-C	Intergenic, regulatory	*MFSD12*	African, South Asian, Australo-Melanesian	Positive	SP	Crawford et al. 2017b
Immunity	Substitution	rs2304207-C	Intronic	*IRF3*	European	Positive	IDS	Bigham et al. 2011
Climatic	Substitution	rs6058017-G	Intronic	*ASIP*	African	Positive	SP	Norton et al. 2007
Climatic	Substitution	rs4911178-A	Intronic, Regulatory	*GROW1* enhancer (of *GDF5*)	East Asian	Positive	BSP	Capellini et al. 2017
Immunity	Substitution	rs7280422-G	Intronic	*MX1*	European	Positive	IDS	Bigham et al. 2011
Immunity	Substitution	rs2071748-A	Intronic	*HMOX1*	Orang Asli	Positive	IDR	Liu et al. 2015
Immunity	Substitution	rs8139532-G	Intronic	*HMOX1*	Orang Asli	Positive	IDR	Liu et al. 2015
Immunity	Substitution	rs5030868-T	Exonic: missense (p.Ser188Phe)	*G6PD*	Southern European, Middle Eastern, Indian	Balancing	IDR	Tishkoff et al. 2001; Verrelli et al. 2002
Immunity	Substitution	rs1050828-A	Exonic: missense (p.Val98Met)	*G6PD*	Sub-Saharan African	Balancing	IDR	Saunders et al. 2002; Saunders et al. 2005; Tishkoff et al. 2001; Verrelli et al. 2002
Immunity	Substitution	rs137852331-G	Exonic: missense (p.Asn165Asp)	*G6PD*	Sub-Saharan African	Balancing	IDR	Saunders et al. 2002; Saunders et al. 2005; Tishkoff et al. 2001; Verrelli et al. 2002

REFERENCES

Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA, Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-73. doi: 10.1038/nature09534

Acheson DJ, Hagoort P (2013) Stimulating the brain's language network: syntactic ambiguity resolution after TMS to the inferior frontal gyrus and middle temporal gyrus. J Cogn Neurosci 25: 1664-77. doi: 10.1162/jocn_a_00430

Akbari A, Vitti JJ, Iranmehr A, Bakhtiari M, Sabeti PC, Mirarab S, Bafna V (2018) Identifying the favored mutation in a positive selective sweep. Nat Methods 15: 279-282. doi: 10.1038/nmeth.4606

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2: e286. doi: 10.1371/journal.pbio.0020286

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12: 1805-14. doi: 10.1101/gr.631202

Algady W, Louzada S, Carpenter D, Brajer P, Farnert A, Rooth I, Ngasala B, Yang F, Shaw MA, Hollox EJ (2018) The Malaria-Protective Human Glycophorin Structural Variant DUP4 Shows Somatic Mosaicism and Association with Hemoglobin Levels. Am J Hum Genet 103: 769-776. doi: 10.1016/j.ajhg.2018.10.008

Alkorta-Aranburu G, Beall CM, Witonsky DB, Gebremedhin A, Pritchard JK, Di Rienzo A (2012) The genetic architecture of adaptations to high altitude in Ethiopia. PLoS Genet 8: e1003110. doi: 10.1371/journal.pgen.1003110

Allison AC (1954) Notes on sickle-cell polymorphism. Ann Hum Genet 19: 39-51. doi: 10.1111/j.1469-1809.1954.tb01261.x

Alonso S, Izagirre N, Smith-Zubiaga I, Gardeazabal J, Diaz-Ramon JL, Diaz-Perez JL, Zelenika D, Boyano MD, Smit N, de la Rua C (2008) Complex signatures of selection for the melanogenic loci TYR, TYRP1 and DCT in humans. BMC Evol Biol 8: 74. doi: 10.1186/1471-2148-8-74

Amato R, Miele G, Monticelli A, Cocozza S (2011) Signs of selective pressure on genetic variants affecting human height. PLoS One 6: e27588. doi: 10.1371/journal.pone.0027588

Andolfatto P (2001) Adaptive hitchhiking effects on genome variability. Curr Opin Genet Dev 11: 635-41.

Andrews MG, Subramanian L, Kriegstein AR (2020) mTOR signaling regulates the morphology and migration of outer radial glia in developing human cortex. Elife 9. doi: 10.7554/eLife.58737

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164: 1229-36. doi: 10.1093/genetics/164.3.1229

Antonacci F, Dennis MY, Huddleston J, Sudmant PH, Steinberg KM, Rosenfeld JA, Miroballo M, Graves TA, Vives L, Malig M, Denman L, Raja A, Stuart A, Tang J, Munson B, Shaffer LG, Amemiya CT, Wilson RK, Eichler EE (2014) Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. Nat Genet 46: 1293-302. doi: 10.1038/ng.3120

Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, Girirajan S, Alkan C, Campbell CD, Vives L, Malig M, Rosenfeld JA, Ballif BC, Shaffer LG, Graves TA, Wilson RK, Schwartz DC, Eichler EE (2010) A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. Nat Genet 42: 745-50. doi: 10.1038/ng.643

Arribas JCD, Herrero AG, Martín-Lomas M, Cañada FJ, He S, Withers SG (2000) Differential mechanism-based labeling and unequivocal activityassignment of the two active sites of intestinal lactase/phlorizinhydrolase. European Journal of Biochemistry 267: 6996-7005.

Astling DP, Heft IE, Jones KL, Sikela JM (2017) High resolution measurement of DUF1220 domain copy number from whole genome sequence data. BMC Genomics 18: 614. doi: 10.1186/s12864-017-3976-z

Bamshad M, Wooding SP (2003) Signatures of natural selection in the human genome. Nat Rev Genet 4: 99-111. doi: 10.1038/nrg999

Bamshad MJ, Mummidi S, Gonzalez E, Ahuja SS, Dunn DM, Watkins WS, Wooding S, Stone AC, Jorde LB, Weiss RB, Ahuja SK (2002) A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A 99: 10539-44. doi: 10.1073/pnas.162046399

Barrett RD, Schluter D (2008) Adaptation from standing genetic variation. Trends Ecol Evol 23: 38-44. doi: 10.1016/j.tree.2007.09.008

Barton RA, Venditti C (2014) Rapid evolution of the cerebellum in humans and other great apes. Curr Biol 24: 2440-4. doi: 10.1016/j.cub.2014.08.056

Basu Mallick C, Iliescu FM, Möls M, Hill S, Tamang R, Chaubey G, Goto R, Ho SY, Gallego Romero I, Crivellaro F, Hudjashov G, Rai N, Metspalu M, Mascie-Taylor CG, Pitchappan R, Singh L, Mirazon-Lahr M, Thangaraj K, Villems R, Kivisild T (2013) The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. PLoS Genet 9: e1003912. doi: 10.1371/journal.pgen.1003912

Beall CM (2013) Human adaptability studies at high altitude: research designs and major concepts during fifty years of discovery. Am J Hum Biol 25: 141-7. doi: 10.1002/ajhb.22355

Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, Li C, Li JC, Liang Y, McCormack M, Montgomery HE, Pan H, Robbins PA, Shianna KV, Tam SC, Tsering N, Veeramah KR, Wang W, Wangdui P, Weale ME, Xu Y, Xu Z, Yang L, Zaman MJ, Zeng C, Zhang L, Zhang X, Zhaxi P, Zheng YT (2010) Natural selection on EPAS1

(HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders. Proc Natl Acad Sci U S A 107: 11459-64. doi: 10.1073/pnas.1002443107

Begun DJ, Aquadro CF (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 356: 519-20. doi: 10.1038/356519a0

Beniaminov A, Westhof E, Krol A (2008) Distinctive structures between chimpanzee and human in a brain noncoding RNA. RNA 14: 1270-5. doi: 10.1261/rna.1054608

Benito-Kwiecinski S, Giandomenico SL, Sutcliffe M, Riis ES, Freire-Pritchett P, Kelava I, Wunderlich S, Martin U, Wray GA, McDole K, Lancaster MA (2021) An early cell shape transition drives evolutionary expansion of the human forebrain. Cell 184: 2084-2102.e19. doi: 10.1016/j.cell.2021.02.050

Bentley AR, Callier S, Rotimi CN (2017) Diversity and inclusion in genomic research: why the uneven progress? J Community Genet. doi: 10.1007/s12687-017-0316-6

Bentley AR, Callier SL, Rotimi CN (2020) Evaluating the promise of inclusion of African ancestry populations in genomics. NPJ Genom Med 5: 5. doi: 10.1038/s41525-019-0111-x

Berg JJ, Coop G (2014) A population genetic signal of polygenic adaptation. PLoS Genet 10: e1004412. doi: 10.1371/journal.pgen.1004412

Bergthorsson U, Andersson DI, Roth JR (2007) Ohno's dilemma: evolution of new genes under continuous selection. Proc Natl Acad Sci U S A 104: 17004-9. doi: 10.1073/pnas.0707158104

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74: 1111-20. doi: 10.1086/421051

Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, Scherer SW, Julian CG, Wilson MJ, Lopez Herraez D, Brutsaert T, Parra EJ, Moore LG, Shriver MD (2010) Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet 6: e1001116. doi: 10.1371/journal.pgen.1001116

Bigham AW, Buckingham KJ, Husain S, Emond MJ, Bofferding KM, Gildersleeve H, Rutherford A, Astakhova NM, Perelygin AA, Busch MP, Murray KO, Sejvar JJ, Green S, Kriesel J, Brinton MA, Bamshad M (2011) Host genetic risk factors for West Nile virus infection and disease progression. PLoS One 6: e24745. doi: 10.1371/journal.pone.0024745

Bigham AW, Julian CG, Wilson MJ, Vargas E, Browne VA, Shriver MD, Moore LG (2014) Maternal PRKAA1 and EDNRA genotypes are associated with birth weight, and PRKAA1 with uterine artery diameter and metabolic homeostasis at high altitude. Physiological genomics 46: 687-697.

Bigham AW, Lee FS (2014) Human high-altitude adaptation: forward genetics meets the HIF pathway. Genes Dev 28: 2189-204. doi: 10.1101/gad.250167.114

Bigham AW, Magnaye K, Dunn DM, Weiss RB, Bamshad M (2018) Complex signatures of natural selection at GYPA. Hum Genet 137: 151-160. doi: 10.1007/s00439-018-1866-3

Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, Miller W, Hurles ME, Dermitzakis ET (2007) Fast-evolving noncoding sequences in the human genome. Genome Biol 8: R118. doi: 10.1186/gb-2007-8-6-r118

Biswas S, Akey JM (2006) Genomic insights into positive selection. TRENDS in Genetics 22: 437-446.

Bitar M, Kuiper S, O'Brien EA, Barry G (2019) Genes with human-specific features are primarily involved with brain, immune and metabolic evolution. BMC Bioinformatics 20: 406. doi: 10.1186/s12859-019-2886-2

Boyd JL, Skove SL, Rouanet JP, Pilaz LJ, Bepler T, Gordân R, Wray GA, Silver DL (2015) Human-chimpanzee differences in a FZD8 enhancer alter cell-cycle dynamics in the developing neocortex. Curr Biol 25: 772-779. doi: 10.1016/j.cub.2015.01.041

Brenner M, Hearing VJ (2008) The protective role of melanin against UV damage in human skin. Photochem Photobiol 84: 539-49. doi: 10.1111/j.1751-1097.2007.00226.x

Brutsaert TD, Kiyamu M, Revollendo GE, Isherwood JL, Lee FS, Rivera-Ch M, Leon-Velarde F, Ghosh S, Bigham AW (2019) Association of EGLN1 gene with high aerobic capacity of Peruvian Quechua at high altitude. Proceedings of the National Academy of Sciences 116: 24006-24011.

Bush EC, Lahn BT (2008) A genome-wide screen for noncoding elements important in primate evolution. BMC Evol Biol 8: 17. doi: 10.1186/1471-2148-8-17

Buxton P, Edwards C, Archer CW, Francis-West P (2001) Growth/differentiation factor-5 (GDF-5) and skeletal development. J Bone Joint Surg Am 83-A Suppl 1: S23-30.

Camus D, Hadley TJ (1985) A Plasmodium falciparum antigen that binds to host erythrocytes and merozoites. Science 230: 553-6. doi: 10.1126/science.3901257

Capellini TD, Chen H, Cao J, Doxey AC, Kiapour AM, Schoor M, Kingsley DM (2017) Ancient selection for derived alleles at a GDF5 enhancer influencing human growth and osteoarthritis risk. Nat Genet 49: 1202-1210. doi: 10.1038/ng.3911

Cardona A, Pagani L, Antao T, Lawson DJ, Eichstaedt CA, Yngvadottir B, Shwe MT, Wee J, Romero IG, Raj S, Metspalu M, Villems R, Willerslev E, Tyler-Smith C, Malyarchuk BA, Derenko MV, Kivisild T (2014) Genome-wide analysis of cold adaptation in indigenous Siberian populations. PLoS One 9: e98076. doi: 10.1371/journal.pone.0098076

Cavalli-Sforza LL (1966) Population structure and human evolution. Proc R Soc Lond B Biol Sci 164: 362-79.

Chaplin G (2004) Geographic distribution of environmental factors influencing human skin coloration. Am J Phys Anthropol 125: 292-302. doi: 10.1002/ajpa.10263

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-303.

Charlesworth J, Eyre-Walker A (2008) The McDonald-Kreitman test and slightly deleterious mutations. Mol Biol Evol 25: 1007-15. doi: 10.1093/molbev/msn005

Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20: 393-402. doi: 10.1101/gr.100545.109

Cheng YZ, Eley L, Hynes AM, Overman LM, Simms RJ, Barker A, Dawe HR, Lindsay S, Sayer JA (2012) Investigating embryonic expression patterns and evolution of AHI1 and CEP290 genes, implicated in Joubert syndrome. PLoS One 7: e44975. doi: 10.1371/journal.pone.0044975

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Pääbo S, Rocchi M, Eichler EE (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437: 88-93. doi: 10.1038/nature04000

Cholera R, Brittain NJ, Gillrie MR, Lopera-Mesa TM, Diakite SA, Arie T, Krause MA, Guindo A, Tubman A, Fujioka H, Diallo DA, Doumbo OK, Ho M, Wellems TE, Fairhurst RM (2008) Impaired cytoadherence of Plasmodium falciparum-infected erythrocytes containing sickle hemoglobin. Proc Natl Acad Sci U S A 105: 991-6. doi: 10.1073/pnas.0711401105

Chowdhury F, Rahman MA, Begum YA, Khan AI, Faruque AS, Saha NC, Baby NI, Malek MA, Kumar AR, Svennerholm AM, Pietroni M, Cravioto A, Qadri F (2011) Impact of rapid urbanization on the rates of infection by Vibrio cholerae O1 and enterotoxigenic Escherichia coli in Dhaka, Bangladesh. PLoS Negl Trop Dis 5: e999. doi: 10.1371/journal.pntd.0000999

Claw KG, Anderson MZ, Begay RL, Tsosie KS, Fox K, Garrison NA, Consortium SifIpiGS (2018) A framework for enhancing ethical genomic research with Indigenous communities. Nat Commun 9: 2957. doi: 10.1038/s41467-018-05188-3

Clemente FJ, Cardona A, Inchley CE, Peter BM, Jacobs G, Pagani L, Lawson DJ, Antao T, Vicente M, Mitt M, DeGiorgio M, Faltyskova Z, Xue Y, Ayub Q, Szpak M, Magi R, Eriksson A, Manica A, Raghavan M, Rasmussen M, Rasmussen S, Willerslev E, Vidal-Puig A, Tyler-Smith C, Villems R, Nielsen R, Metspalu M, Malyarchuk B, Derenko M, Kivisild T (2014) A Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. Am J Hum Genet 95: 584-589. doi: 10.1016/j.ajhg.2014.09.016

Collins SA, Sinclair G, McIntosh S, Bamforth F, Thompson R, Sobol I, Osborne G, Corriveau A, Santos M, Hanley B, Greenberg CR, Vallance H, Arbour L (2010) Carnitine palmitoyltransferase 1A (CPT1A) P479L prevalence in live newborns in Yukon, Northwest Territories, and Nunavut. Mol Genet Metab 101: 200-4. doi: 10.1016/j.ymgme.2010.07.013

Cooper DN, Kehrer-Sawatzki H (2011) Exploring the potential relevance of human-specific genes to complex disease. Hum Genomics 5: 99-107. doi: 10.1186/1479-7364-5-2-99

Corbett S, Courtiol A, Lummaa V, Moorad J, Stearns S (2018) The transition to modernity and chronic disease: mismatch and natural selection. Nat Rev Genet 19: 419-430. doi: 10.1038/s41576-018-0012-3

Crawford JE, Amaru R, Song J, Julian CG, Racimo F, Cheng JY, Guo X, Yao J, Ambale-Venkatesh B, Lima JA, Rotter JI, Stehlik J, Moore LG, Prchal JT, Nielsen R (2017a) Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. Am J Hum Genet 101: 752-767. doi: 10.1016/j.ajhg.2017.09.023

Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, Jewett E, Ranciaro A, Thompson S, Lo Y, Pfeifer SP, Jensen JD, Campbell MC, Beggs W, Hormozdiari F, Mpoloka SW, Mokone GG, Nyambo T, Meskel DW, Belay G, Haut J, Program NCS, Rothschild H, Zon L, Zhou Y, Kovacs MA, Xu M, Zhang T, Bishop K, Sinclair J, Rivas C, Elliot E, Choi J, Li SA, Hicks B, Burgess S, Abnet C, Watkins-Chow DE, Oceana E, Song YS, Eskin E, Brown KM, Marks MS, Loftus SK, Pavan WJ, Yeager M, Chanock S, Tishkoff SA (2017b) Loci associated with skin pigmentation identified in African populations. Science 358. doi: 10.1126/science.aan8433

Cvijovic I, Good BH, Desai MM (2018) The Effect of Strong Purifying Selection on Genetic Diversity. Genetics 209: 1235-1278. doi: 10.1534/genetics.118.301058

Daanen HA, Van Marken Lichtenbelt WD (2016) Human whole body cold adaptation. Temperature (Austin) 3: 104-18. doi: 10.1080/23328940.2015.1135688

Daub JT, Dupanloup I, Robinson-Rechavi M, Excoffier L (2015) Inference of Evolutionary Forces Acting on Human Biological Pathways. Genome Biol Evol 7: 1546-58. doi: 10.1093/gbe/evv083

de Gruijter JM, Lao O, Vermeulen M, Xue Y, Woodwark C, Gillson CJ, Coffey AJ, Ayub Q, Mehdi SQ, Kayser M, Tyler-Smith C (2011) Contrasting signals of positive selection in genes involved in human skin-color variation from tests based on SNP scans and resequencing. Investig Genet 2: 24. doi: 10.1186/2041-2223-2-24

Doan RN, Bae BI, Cubelos B, Chang C, Hossain AA, Al-Saad S, Mukaddes NM, Oner O, Al-Saffar M, Balkhy S, Gascon GG, Nieto M, Walsh CA, Autism HMCf (2016) Mutations in Human Accelerated Regions Disrupt Cognition and Social Behavior. Cell 167: 341-354.e12. doi: 10.1016/j.cell.2016.08.071

Doering JE, Kane K, Hsiao YC, Yao C, Shi B, Slowik AD, Dhagat B, Scott DD, Ault JG, Page-McCaw PS, Ferland RJ (2008) Species differences in the expression of Ahi1, a protein implicated in the neurodevelopmental disorder

Joubert syndrome, with preferential accumulation to stigmoid bodies. J Comp Neurol 511: 238-56. doi: 10.1002/cne.21824

Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, Harshman L, Duyzend MH, Ventura M, Antonacci F, Sandstrom R, Dennis MY, Eichler EE (2017) The birth of a human-specific neural gene by incomplete duplication and gene fusion. Genome Biol 18: 49. doi: 10.1186/s13059-017-1163-9

Dougherty ML, Underwood JG, Nelson BJ, Tseng E, Munson KM, Penn O, Nowakowski TJ, Pollen AA, Eichler EE (2018) Transcriptional fates of human-specific segmental duplications in brain. Genome Res 28: 1566-1576. doi: 10.1101/gr.237610.118

Eaaswarkhanth M, Dos Santos ALC, Gokcumen O, Al-Mulla F, Thanaraj TA (2020) Genome-Wide Selection Scan in an Arabian Peninsula Population Identifies a TNKS Haplotype Linked to Metabolic Traits and Hypertension. Genome Biol Evol 12: 77-87. doi: 10.1093/gbe/evaa033

Eichstaedt CA, Antao T, Cardona A, Pagani L, Kivisild T, Mormina M (2015) Positive selection of AS3MT to arsenic water in Andean populations. Mutat Res 780: 97-102. doi: 10.1016/j.mrfmmm.2015.07.007

Eichstaedt CA, Pagani L, Antao T, Inchley CE, Cardona A, Morseburg A, Clemente FJ, Sluckin TJ, Metspalu E, Mitt M, Magi R, Hudjashov G, Metspalu M, Mormina M, Jacobs GS, Kivisild T (2017) Evidence of Early-Stage Selection on EPAS1 and GPR126 Genes in Andean High Altitude Populations. Sci Rep 7: 13042. doi: 10.1038/s41598-017-13382-4

Elias PM, Williams ML (2013) Re-appraisal of current theories for the development and loss of epidermal pigmentation in hominins and modern humans. J Hum Evol 64: 687-92. doi: 10.1016/j.jhevol.2013.02.003

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Jarvela I (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30: 233-7. doi: 10.1038/ng826

Evans AG, Wellems TE (2002) Coevolutionary genetics of Plasmodium malaria parasites and their human hosts. Integr Comp Biol 42: 401-7. doi: 10.1093/icb/42.2.401

Evans PD, Anderson JR, Vallender EJ, Choi SS, Lahn BT (2004a) Reconstructing the evolutionary history of microcephalin, a gene controlling human brain size. Hum Mol Genet 13: 1139-45. doi: 10.1093/hmg/ddh126

Evans PD, Anderson JR, Vallender EJ, Gilbert SL, Malcom CM, Dorus S, Lahn BT (2004b) Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. Hum Mol Genet 13: 489-94. doi: 10.1093/hmg/ddh055

Evans PD, Vallender EJ, Lahn BT (2006) Molecular evolution of the brain size regulator genes CDK5RAP2 and CENPJ. Gene 375: 75-9. doi: 10.1016/j.gene.2006.02.019

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theor Popul Biol 3: 87-112. doi: 10.1016/0040-5809(72)90035-4

Faruque SM, Sack DA, Sack RB, Colwell RR, Takeda Y, Nair GB (2003) Emergence and evolution of Vibrio cholerae O139. Proc Natl Acad Sci U S A 100: 1304-9. doi: 10.1073/pnas.0337468100

Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-13.

Fay JC, Wyckoff GJ, Wu CI (2001) Positive and negative selection on the human genome. Genetics 158: 1227-34. doi: 10.1093/genetics/158.3.1227

Fehren-Schmitz L, Georges L (2016) Ancient DNA reveals selection acting on genes associated with hypoxia response in pre-Columbian Peruvian Highlanders in the last 8500 years. Sci Rep 6: 23485. doi: 10.1038/srep23485

Feldman MW, Cavalli-Sforza LL (1976) Cultural and biological evolutionary processes, selection for a trait under complex transmission. Theor Popul Biol 9: 238-59. doi: 10.1016/0040-5809(76)90047-2

Ferland RJ, Eyaid W, Collura RV, Tully LD, Hill RS, Al-Nouri D, Al-Rumayyan A, Topcu M, Gascon G, Bodell A, Shugart YY, Ruvolo M, Walsh CA (2004) Abnormal cerebellar development and axonal decussation due to mutations in AHI1 in Joubert syndrome. Nat Genet 36: 1008-13. doi: 10.1038/ng1419

Fernandez CI, Wiley AS (2017) Rethinking the starch digestion hypothesis for AMY1 copy number variation in humans. Am J Phys Anthropol 163: 645-657. doi: 10.1002/ajpa.23237

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. Mol Biol Evol 31: 1275-91. doi: 10.1093/molbev/msu077

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, Lorig-Roach R, Field AR, Haeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor MC, Zwolinski S, Katzman S, Kriegstein A, Eichler EE, Salama SR, Jacobs FMJ, Haussler D (2018a) Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell 173: 1356-1369 e22. doi: 10.1016/j.cell.2018.03.051

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, Lorig-Roach R, Field AR, Haeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor MC, Zwolinski S, Katzman S, Kriegstein A, Eichler EE, Salama SR, Jacobs FMJ, Haussler D (2018b) Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell 173: 1356-1369.e22. doi: 10.1016/j.cell.2018.03.051

Fischer A, Rausell A (2016) Primary immunodeficiencies suggest redundancy within the human immune system. Science Immunology 1: eaah5861. doi: 10.1126/sciimmunol.aah5861

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, Guhr E, Klemroth S, Prüfer K, Kelso J, Naumann R, Nüsslein I, Dahl A, Lachmann R, Pääbo S, Huttner WB (2015) Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science 347: 1465-70. doi: 10.1126/science.aaa1975

Florio M, Heide M, Pinson A, Brandl H, Albert M, Winkler S, Wimberger P, Huttner WB, Hiller M (2018) Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. Elife 7. doi: 10.7554/eLife.32332

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531-45. doi: 10.1093/genetics/151.4.1531

Fossati M, Pizzarelli R, Schmidt ER, Kupferman JV, Stroebel D, Polleux F, Charrier C (2016) SRGAP2 and Its Human-Specific Paralog Co-Regulate the Development of Excitatory and Inhibitory Synapses. Neuron 91: 356-69. doi: 10.1016/j.neuron.2016.06.013

Francis-West PH, Parish J, Lee K, Archer CW (1999) BMP/GDF-signalling interactions during synovial joint development. Cell Tissue Res 296: 111-9. doi: 10.1007/s004410051272

Fu YX (1995) Statistical properties of segregating sites. Theor Popul Biol 48: 172-97. doi: 10.1006/tpbi.1995.1025

Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Fukami M, Wada Y, Miyabayashi K, Nishino I, Hasegawa T, Nordenskjöld A, Camerino G, Kretz C, Buj-Bello A, Laporte J, Yamada G, Morohashi K, Ogata T (2006) CXorf6 is a causative gene for hypospadias. Nat Genet 38: 1369-71. doi: 10.1038/ng1900

Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Bresolin N, Clerici M, Sironi M (2010) Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. PLoS Genet 6: e1000849. doi: 10.1371/journal.pgen.1000849

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Ferrer-Admettla A, Pattini L, Nielsen R (2011) Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet 7: e1002355. doi: 10.1371/journal.pgen.1002355

Galvani AP, Slatkin M (2003) Evaluating plague and smallpox as historical selective pressures for the CCR5-Delta 32 HIV-resistance allele. Proceedings of the National Academy of Sciences of the United States of America 100: 15276-15279. doi: 10.1073/pnas.2435085100

Gillespie JH (1994) The causes of molecular evolution. Oxford University Press On Demand

Gittelman RM, Hun E, Ay F, Madeoy J, Pennacchio L, Noble WS, Hawkins RD, Akey JM (2015) Comprehensive identification and analysis of human accelerated regulatory DNA. Genome Res 25: 1245-55. doi: 10.1101/gr.192591.115

Gluckman PD, Beedle A, Hanson MA (2009) Principles of evolutionary medicine. Oxford University Press, Oxford ; New York

Goebel T (1999) Pleistocene human colonization of Siberia and peopling of the Americas: an ecological approach. Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews 8: 208-227.

Goebel T, Waters MR, O'Rourke DH (2008) The late Pleistocene dispersal of modern humans in the Americas. Science 319: 1497-502. doi: 10.1126/science.1153569

Gout JF, Duret L, Kahn D (2009) Differential retention of metabolic genes following whole-genome duplication. Mol Biol Evol 26: 1067-72. doi: 10.1093/molbev/msp026

Gout JF, Kahn D, Duret L, Consortium PP-G (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet 6: e1000944. doi: 10.1371/journal.pgen.1000944

Greenberg CR, Dilling LA, Thompson GR, Seargeant LE, Haworth JC, Phillips S, Chan A, Vallance HD, Waters PJ, Sinclair G, Lillquist Y, Wanders RJ, Olpin SE (2009) The paradox of the carnitine palmitoyltransferase type Ia P479L variant in Canadian Aboriginal populations. Mol Genet Metab 96: 201-7. doi: 10.1016/j.ymgme.2008.12.018

Gregg XT, Prchal JT (2018) Red blood cell enzymopathies. *Hematology*, 7th edn. Elsevier, pp 616-625

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, Cabili M, Adegbola RA, Bamezai RN, Hill AV, Vannberg FO, Rinn JL, Genomes P, Lander ES, Schaffner SF, Sabeti PC (2013) Identifying recent adaptations in large-scale genomic data. Cell 152: 703-13. doi: 10.1016/j.cell.2013.01.035

Grossman SR, Shlyakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O, Lander ES, Schaffner SF, Sabeti PC (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327: 883-6. doi: 10.1126/science.1183863

Guenther CA, Tasic B, Luo L, Bedell MA, Kingsley DM (2014) A molecular basis for classic blond hair color in Europeans. Nat Genet 46: 748-52. doi: 10.1038/ng.2991

Guernsey DL, Jiang H, Hussin J, Arnold M, Bouyakdan K, Perry S, Babineau-Sturk T, Beis J, Dumas N, Evans SC, Ferguson M, Matsuoka M, Macgillivray C, Nightingale M, Patry L, Rideout AL, Thomas A, Orr A, Hoffmann I, Michaud JL, Awadalla P, Meek DC, Ludman M, Samuels ME (2010) Mutations in centrosomal protein CEP152 in primary microcephaly families linked to MCPH4. Am J Hum Genet 87: 40-51. doi: 10.1016/j.ajhg.2010.06.003

Gunther T, Nettelblad C (2019) The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genet 15: e1008302. doi: 10.1371/journal.pgen.1008302

Guo J, Yang J, Visscher PM (2018) Leveraging GWAS for complex traits to detect signatures of natural selection in humans. Curr Opin Genet Dev 53: 9-14. doi: 10.1016/j.gde.2018.05.012

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, Fu Q, Mittnik A, Banffy E, Economou C, Francken M, Friederich S, Pena RG, Hallgren F, Khartanovich V, Khokhlov A, Kunst M, Kuznetsov P, Meller H, Mochalov O, Moiseyev V, Nicklisch N, Pichler SL, Risch R, Rojo Guerra MA, Roth C, Szecsenyi-Nagy A, Wahl J, Meyer M, Krause J, Brown D, Anthony D, Cooper A, Alt KW, Reich D (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. Nature 522: 207-11. doi: 10.1038/nature14317

Haas R, Stefanescu IC, Garcia-Putnam A, Aldenderfer MS, Clementz MT, Murphy MS, Llave CV, Watson JT (2017) Humans permanently occupied the Andean highlands by at least 7 ka. R Soc Open Sci 4: 170331. doi: 10.1098/rsos.170331

Haasl RJ, Payseur BA (2016) Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. Mol Ecol 25: 5-23. doi: 10.1111/mec.13339

Hadley TJ, Peiper SC (1997) From malaria to chemokine receptor: the emerging physiologic role of the Duffy blood group antigen. Blood 89: 3077-91.

Haldane JBS (1932) The causes of evolution. Longmans Green & Co, London

Haldane JBS (1949) Disease and evolution. Ricercha

Hamblin MT, Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66: 1669-79. doi: 10.1086/302879

Hamblin MT, Thompson EE, Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70: 369-83. doi: 10.1086/338628

Hardy S, Legagneux V, Audic Y, Paillard L (2010) Reverse genetics in eukaryotes. Biol Cell 102: 561-80. doi: 10.1042/BC20100038

Harris K (2018) The randomness that shapes our DNA. Elife 7. doi: 10.7554/eLife.41491

Harrison PW, Montgomery SH (2017) Genetics of Cerebellar and Neocortical Expansion in Anthropoid Primates: A Comparative Approach. Brain Behav Evol 89: 274-285. doi: 10.1159/000477432

Haygood R, Babbitt CC, Fedrigo O, Wray GA (2010) Contrasts between adaptive coding and noncoding changes during human evolution. Proc Natl Acad Sci U S A 107: 7853-7. doi: 10.1073/pnas.0911249107

Heft IE, Mostovoy Y, Levy-Sakin M, Ma W, Stevens AJ, Pastor S, McCaffrey J, Boffelli D, Martin DI, Xiao M, Kennedy MA, Kwok PY, Sikela JM (2020) The Driver of Extreme Human-Specific Olduvai Repeat Expansion Remains Highly Active in the Human Genome. Genetics 214: 179-191. doi: 10.1534/genetics.119.302782

Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics 169: 2335-52. doi: 10.1534/genetics.104.036947

Hodgson JA, Pickrell JK, Pearson LN, Quillen EE, Prista A, Rocha J, Soodyall H, Shriver MD, Perry GH (2014) Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. Proc Biol Sci 281: 20140930. doi: 10.1098/rspb.2014.0930

Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, Eskin E (2016) Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet 99: 1245-1260. doi: 10.1016/j.ajhg.2016.10.003

Hsieh P, Hallmark B, Watkins J, Karafet TM, Osipova LP, Gutenkunst RN, Hammer MF (2017) Exome Sequencing Provides Evidence of Polygenic Adaptation to a Fat-Rich Animal Diet in Indigenous Siberian Populations. Mol Biol Evol 34: 2913-2926. doi: 10.1093/molbev/msx226

Hubisz MJ, Pollard KS (2014) Exploring the genesis and functions of Human Accelerated Regions sheds light on their role in human evolution. Curr Opin Genet Dev 29: 15-21. doi: 10.1016/j.gde.2014.07.005

Hughes AL, Yeager M (1998) Natural selection and the evolutionary history of major histocompatibility complex loci. Front Biosci 3: d509-16. doi: 10.2741/a298

Ilardo MA, Moltke I, Korneliussen TS, Cheng J, Stern AJ, Racimo F, de Barros Damgaard P, Sikora M, Seguin-Orlando A, Rasmussen S, van den Munckhof ICL, Ter Horst R, Joosten LAB, Netea MG, Salingkat S, Nielsen R, Willerslev E (2018) Physiological and Genetic Adaptations to Diving in Sea Nomads. Cell 173: 569-580 e15. doi: 10.1016/j.cell.2018.03.054

Ip E, Chapman G, Winlaw D, Dunwoodie SL, Giannoulatou E (2019) VPOT: A Customizable Variant Prioritization Ordering Tool for Annotated Variants. Genomics Proteomics Bioinformatics 17: 540-545. doi: 10.1016/j.gpb.2019.11.001

Iskow RC, Gokcumen O, Lee C (2012) Exploring the role of copy number variants in human adaptation. Trends Genet 28: 245-57. doi: 10.1016/j.tig.2012.03.002

Izagirre N, Garcia I, Junquera C, de la Rua C, Alonso S (2006) A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. Mol Biol Evol 23: 1697-706. doi: 10.1093/molbev/msl030

Jablonski NG, Chaplin G (2000) The evolution of human skin coloration. J Hum Evol 39: 57-106. doi: 10.1006/jhev.2000.0403

Jablonski NG, Chaplin G (2013) Epidermal pigmentation in the human lineage is an adaptation to ultraviolet radiation. J Hum Evol 65: 671-5. doi: 10.1016/j.jhevol.2013.06.004

Jackson L, Kuhlman C, Jackson F, Fox KP (2019) Including Vulnerable Populations in the Assessment of Data From Vulnerable Populations. Frontiers in Big Data 2: 19. doi: 10.3389/fdata.2019.00019

Jarolim P, Palek J, Amato D, Hassan K, Sapak P, Nurse GT, Rubin HL, Zhai S, Sahr KE, Liu SC (1991) Deletion in erythrocyte band 3 gene in malaria-resistant Southeast Asian ovalocytosis. Proc Natl Acad Sci U S A 88: 11022-6.

Jarvis JP, Scheinfeldt LB, Soi S, Lambert C, Omberg L, Ferwerda B, Froment A, Bodo JM, Beggs W, Hoffman G, Mezey J, Tishkoff SA (2012) Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. PLoS Genet 8: e1002641. doi: 10.1371/journal.pgen.1002641

Jha P, Lu D, Xu S (2015) Natural Selection and Functional Potentials of Human Noncoding Elements Revealed by Analysis of Next Generation Sequencing Data. PLoS One 10: e0129023. doi: 10.1371/journal.pone.0129023

Johnsson M (2018) Integrating Selection Mapping With Genetic Mapping and Functional Genomics. Front Genet 9: 603. doi: 10.3389/fgene.2018.00603

Jovanovic VM, Sarfert M, Reyna-Blanco CS, Indrischek H, Valdivia DI, Shelest E, Nowick K (2021) Positive Selection in Gene Regulatory Factors Suggests Adaptive Pleiotropic Changes During Human Evolution. Front Genet 12: 662239. doi: 10.3389/fgene.2021.662239

Kaidbey KH, Agin PP, Sayre RM, Kligman AM (1979) Photoprotection by melanin--a comparison of black and Caucasian skin. J Am Acad Dermatol 1: 249-60.

Kalebic N, Gilardi C, Stepien B, Wilsch-Bräuninger M, Long KR, Namba T, Florio M, Langen B, Lombardot B, Shevchenko A, Kilimann MW, Kawasaki H, Wimberger P, Huttner WB (2019) Neocortical Expansion Due to Increased Proliferation of Basal Progenitors Is Linked to Changes in Their Morphology. Cell Stem Cell 24: 535-550.e9. doi: 10.1016/j.stem.2019.02.017

Kamm GB, Pisciottano F, Kliger R, Franchini LF (2013) The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. Mol Biol Evol 30: 1088-102. doi: 10.1093/molbev/mst023

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709-23. doi: 10.1534/genetics.107.080101

Karczewski KJ, Snyder MP (2018) Integrative omics for health and disease. Nat Rev Genet 19: 299-310. doi: 10.1038/nrg.2018.4

Karlsson EK, Harris JB, Tabrizi S, Rahman A, Shlyakhter I, Patterson N, O'Dushlaine C, Schaffner SF, Gupta S, Chowdhury F, Sheikh A, Shin OS, Ellis C, Becker CE, Stuart LM, Calderwood SB, Ryan ET, Qadri F, Sabeti PC, Larocque RC (2013) Natural selection in a bangladeshi population from the cholera-endemic ganges river delta. Sci Transl Med 5: 192ra86. doi: 10.1126/scitranslmed.3006338

Karlsson EK, Kwiatkowski DP, Sabeti PC (2014) Natural selection and infectious disease in human populations. Nat Rev Genet 15: 379-93. doi: 10.1038/nrg3734

Keeney JG, Dumas L, Sikela JM (2014) The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. Front Hum Neurosci 8: 427. doi: 10.3389/fnhum.2014.00427

Kim Y, Stephan W (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765-77.

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217: 624-6. doi: 10.1038/217624a0

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275-6. doi: 10.1038/267275a0

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge Cambridgeshire ; New York

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188: 107-16.

Knapp M, Hofreiter M (2010) Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. Genes (Basel) 1: 227-43. doi: 10.3390/genes1020227

Ko A, Cantor RM, Weissglas-Volkov D, Nikkola E, Reddy PM, Sinsheimer JS, Pasaniuc B, Brown R, Alvarez M, Rodriguez A, Rodriguez-Guillen R, Bautista IC, Arellano-Campos O, Munoz-Hernandez LL, Salomaa V, Kaprio J, Jula A, Jauhiainen M, Heliovaara M, Raitakari O, Lehtimaki T, Eriksson JG, Perola M, Lohmueller KE, Matikainen N, Taskinen MR, Rodriguez-Torres M, Riba L, Tusie-Luna T, Aguilar-Salinas CA, Pajukanta P (2014) Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. Nat Commun 5: 3983. doi: 10.1038/ncomms4983

Ko WY, Kaercher KA, Giombini E, Marcatili P, Froment A, Ibrahim M, Lema G, Nyambo TB, Omar SA, Wambebe C, Ranciaro A, Hirbo JB, Tishkoff SA (2011) Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. Am J Hum Genet 88: 741-754. doi: 10.1016/j.ajhg.2011.05.005

Kondrashov FA, Kondrashov AS (2006) Role of selection in fixation of gene duplications. J Theor Biol 239: 141-51. doi: 10.1016/j.jtbi.2005.08.033

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome Biol 3: RESEARCH0008. doi: 10.1186/gb-2002-3-2-research0008

Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208-22. doi: 10.1093/molbev/msi105

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23: 1891-901. doi: 10.1093/molbev/msl051

Kouprina N, Pavlicek A, Mochida GH, Solomon G, Gersch W, Yoon YH, Collura R, Ruvolo M, Barrett JC, Woods CG, Walsh CA, Jurka J, Larionov V (2004) Accelerated evolution of the ASPM gene controlling brain size begins prior to human brain expansion. PLoS Biol 2: E126. doi: 10.1371/journal.pbio.0020126

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, Munson KM, Hastie AR, Diekhans M, Hormozdiari F, Lorusso N, Hoekzema K, Qiu R, Clark K, Raja A, Welch AE, Sorensen M, Baker C, Fulton RS, Armstrong J, Graves-Lindsay TA, Denli AM, Hoppe ER, Hsieh P, Hill CM, Pang AWC, Lee J, Lam ET, Dutcher SK, Gage FH, Warren WC, Shendure J, Haussler D, Schneider VA, Cao H, Ventura M, Wilson RK, Paten B, Pollen A, Eichler EE (2018) High-resolution comparative analysis of great ape genomes. Science 360. doi: 10.1126/science.aar6343

Kwiatkowski DP (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77: 171-92. doi: 10.1086/432519

Laland KN, Odling-Smee J, Myles S (2010) How culture shaped the human genome: bringing genetics and the human sciences together. Nat Rev Genet 11: 137-48. doi: 10.1038/nrg2734

Lamason RL, Mohideen MA, Mest JR, Wong AC, Norton HL, Aros MC, Jurynec MJ, Mao X, Humphreville VR, Humbert JE, Sinha S, Moore JL, Jagadeeswaran P, Zhao W, Ning G, Makalowska I, McKeigue PM, O'Donnell D, Kittles R, Parra EJ, Mangini NJ, Grunwald DJ, Shriver MD, Canfield VA, Cheng KC (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science 310: 1782-6. doi: 10.1126/science.1116238

Lao O, de Gruijter JM, van Duijn K, Navarro A, Kayser M (2007) Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet 71: 354-69. doi: 10.1111/j.1469-1809.2006.00341.x

Latta RG (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. Am Nat 151: 283-92. doi: 10.1086/286119

Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait land trait in a subdivided population under selection. Genetics 164: 1205-19.

Lee S, Wang H, Xing EP (2017) Backward genotype-transcript-phenotype association mapping. Methods 129: 18-23. doi: 10.1016/j.ymeth.2017.09.004

Leffler EM, Band G, Busby GBJ, Kivinen K, Le QS, Clarke GM, Bojang KA, Conway DJ, Jallow M, Sisay-Joof F, Bougouma EC, Mangano VD, Modiano D, Sirima SB, Achidi E, Apinjoh TO, Marsh K, Ndila CM, Peshu N, Williams TN, Drakeley C, Manjurano A, Reyburn H, Riley E, Kachala D, Molyneux M, Nyirongo V, Taylor T, Thornton N, Tilley L, Grimsley S, Drury E, Stalker J, Cornelius V, Hubbart C, Jeffreys AE, Rowlands K, Rockett KA, Spencer CCA, Kwiatkowski DP, Network MGE (2017) Resistance to malaria through structural variation of red blood cell invasion receptors. Science 356. doi: 10.1126/science.aam6393

Lemas DJ, Wiener HW, O'Brien DM, Hopkins S, Stanhope KL, Havel PJ, Allison DB, Fernandez JR, Tiwari HK, Boyer BB (2012) Genetic polymorphisms in carnitine palmitoyltransferase 1A gene are associated with variation in body composition and fasting lipid traits in Yup'ik Eskimos. J Lipid Res 53: 175-84. doi: 10.1194/jlr.P018952

Levchenko A, Kanapin A, Samsonova A, Gainetdinov RR (2018) Human Accelerated Regions and Other Human-Specific Sequence Variations in the Context of Evolution and Their Relevance for Brain Development. Genome Biol Evol 10: 166-188. doi: 10.1093/gbe/evx240

Lewandowska M, Jedrychowska-Danska K, Ploszaj T, Witas P, Zamerska A, Mankowska-Pliszka H, Witas HW (2018) Searching for signals of recent natural selection in genes of the innate immune response - ancient DNA study. Infect Genet Evol 63: 62-72. doi: 10.1016/j.meegid.2018.05.008

León-Velarde F (2003) Pursuing international recognition of chronic mountain sickness. High Alt Med Biol 4: 256-9. doi: 10.1089/152702903322022857

Libert F CP, Beckman G, Samson M, Aksenova M, et al. (1998) The delta CCR5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe.

Lim JK, Lisco A, McDermott DH, Huynh L, Ward JM, Johnson B, Johnson H, Pape J, Foster GA, Krysztof D, Follmann D, Stramer SL, Margolis LB, Murphy PM (2009) Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. PLoS Pathog 5: e1000321. doi: 10.1371/journal.ppat.1000321

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, Martins AL, Massingham T, Moltke I, Raney BJ, Rasmussen MD, Robinson J, Stark A, Vilella AJ, Wen J, Xie X, Zody MC, Baldwin J, Bloom T, Chin CW, Heiman D, Nicol R, Nusbaum C, Young S, Wilkinson J, Worley KC, Kovar CL, Muzny DM, Gibbs RA, Cree A, Dihn HH, Fowler G, Jhangiani S, Joshi V, Lee S, Lewis LR, Nazareth LV, Okwuonu G, Santibanez J, Warren WC, Mardis ER, Weinstock GM, Wilson RK, Delehaunty K, Dooling D, Fronik C, Fulton L, Fulton B, Graves T, Minx P, Sodergren E, Birney E, Margulies EH, Herrero J, Green ED, Haussler D, Siepel A, Goldman N, Pollard KS, Pedersen JS, Lander ES, Kellis M, Team BISPaWGA, Team BCoMHGSCS, University GIaW (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476-82. doi: 10.1038/nature10530

Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, Watson JT, Viviano Llave C, Witonsky D, Beall C, Warinner C, Novembre J, Aldenderfer M, Di Rienzo A (2018) The genetic prehistory of the Andean highlands 7000 years BP though European contact. Sci Adv 4: eaau4921. doi: 10.1126/sciadv.aau4921

Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, MacDonald ME, Stuhlmann H, Koup RA, Landau NR (1996) Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. Cell 86: 367-377.

Liu X, Yunus Y, Lu D, Aghakhanian F, Saw WY, Deng L, Ali M, Wang X, Nor FM, Rahman TA, Shaari SA, Salleh MZ, Phipps ME, Ong RT, Xu S, Teo YY, Hoh BP (2015) Differential positive selection of malaria resistance genes in three indigenous populations of Peninsular Malaysia. Hum Genet 134: 375-92. doi: 10.1007/s00439-014-1525-2

Livingstone FB (1984) The Duffy blood groups, vivax malaria, and malaria selection in human populations: a review. Hum Biol 56: 413-25.

Loomis WF (1967) Skin-pigment regulation of vitamin-D biosynthesis in man. Science 157: 501-6.

Lorenzo FR, Huff C, Myllymäki M, Olenchock B, Swierczek S, Tashi T, Gordeuk V, Wuren T, Ri-Li G, McClain DA, Khan TM, Koul PA, Guchhait P, Salama ME, Xing J, Semenza GL, Liberzon E, Wilson A, Simonson TS, Jorde LB, Kaelin WG, Koivunen P, Prchal JT (2014) A genetic mechanism for Tibetan high-altitude adaptation. Nat Genet 46: 951-6. doi: 10.1038/ng.3067

Lucas ER, Miles A, Harding NJ, Clarkson CS, Lawniczak MKN, Kwiatkowski DP, Weetman D, Donnelly MJ, Consortium AgG (2019) Whole-genome sequencing reveals high complexity of copy number variation at insecticide resistance loci in malaria mosquitoes. Genome Res 29: 1250-1261. doi: 10.1101/gr.245795.118

Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz JM, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas SA, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161: 1012-1025. doi: 10.1016/j.cell.2015.04.004

Luzzatto L (2012) Sickle cell anaemia and malaria. Mediterr J Hematol Infect Dis 4: e2012065. doi: 10.4084/MJHID.2012.065

Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, Richardson MF, Anholt RR, Barrón M, Bess C, Blankenburg KP, Carbone MA, Castellano D, Chaboub L, Duncan L, Harris Z, Javaid M, Jayaseelan JC, Jhangiani SN, Jordan KW, Lara F, Lawrence F, Lee SL, Librado P, Linheiro RS, Lyman RF, Mackey AJ, Munidasa M, Muzny DM, Nazareth L, Newsham I, Perales L, Pu LL, Qu C, Ràmia M, Reid JG, Rollmann SM, Rozas J, Saada N, Turlapati L, Worley KC, Wu YQ, Yamamoto A, Zhu Y, Bergman CM, Thornton KR, Mittelman D, Gibbs RA (2012) The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173-8. doi: 10.1038/nature10811

Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, Ferrari R (2018) Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform 19: 286-302. doi: 10.1093/bib/bbw114

Marchetti G, Pinotti M, Lunghi B, Casari C, Bernardi F (2012) Functional genetics. Thromb Res 129: 336-40. doi: 10.1016/j.thromres.2011.10.028

Marques-Bonet T, Kidd JM, Ventura M, Graves TA, Cheng Z, Hillier LW, Jiang Z, Baker C, Malfavon-Borja R, Fulton LA, Alkan C, Aksay G, Girirajan S, Siswara P, Chen L, Cardone MF, Navarro A, Mardis ER, Wilson RK, Eichler EE (2009) A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457: 877-81. doi: 10.1038/nature07744

Marth GT, Czabarka E, Murvai J, Sherry ST (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics 166: 351-372.

Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M, Sirak K, Gamba C, Jones ER, Llamas B, Dryomov S, Pickrell J, Arsuaga JL, de Castro JM, Carbonell E, Gerritsen F, Khokhlov A, Kuznetsov P, Lozano M, Meller H, Mochalov O, Moiseyev V, Guerra MA, Roodenberg J, Verges JM, Krause J, Cooper A, Alt KW, Brown D, Anthony D, Lalueza-Fox C, Haak W, Pinhasi R, Reich D (2015) Genome-wide patterns of selection in 230 ancient Eurasians. Nature 528: 499-503. doi: 10.1038/nature16152

Matsumoto Y, Goto T, Nishino J, Nakaoka H, Tanave A, Takano-Shimizu T, Mott RF, Koide T (2017) Selective breeding and selection mapping using a novel wild-derived heterogeneous stock of mice revealed two closely-linked loci for tameness. Sci Rep 7: 4607. doi: 10.1038/s41598-017-04869-1

Maynard-Smith J, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23-35.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-4. doi: 10.1038/351652a0

McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, Guenther C, Indjeian VB, Lim X, Menke DB, Schaar BT, Wenger AM, Bejerano G, Kingsley DM (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471: 216-9. doi: 10.1038/nature09774

McVicker G, Gordon D, Davis C, Green P (2009) Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5: e1000471. doi: 10.1371/journal.pgen.1000471

Messer PW, Petrov DA (2013) Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci U S A 110: 8615-20. doi: 10.1073/pnas.1220835110

Meyer CG, Calixto Fernandes MH, Intemann CD, Kreuels B, Kobbe R, Kreuzberg C, Ayim M, Ruether A, Loag W, Ehmen C, Adjei S, Adjei O, Horstmann RD, May J (2011) IL3 variant on chromosomal region 5q31-33 and protection from recurrent malaria attacks. Hum Mol Genet 20: 1173-81. doi: 10.1093/hmg/ddq562

Miller LH, Mason SJ, Clyde DF, McGinniss MH (1976) The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. N Engl J Med 295: 302-4. doi: 10.1056/NEJM197608052950602

Miyata T, Yasunaga T (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol 16: 23-36. doi: 10.1007/BF01732067

Monge C (1942) LIFE IN THE ANDES AND CHRONIC MOUNTAIN SICKNESS. Science 95: 79-84. doi: 10.1126/science.95.2456.79

Monroy Kuhn JM, Jakobsson M, Gunther T (2018) Estimating genetic kin relationships in prehistoric populations. PLoS One 13: e0195491. doi: 10.1371/journal.pone.0195491

Montgomery SH, Capellini I, Venditti C, Barton RA, Mundy NI (2011) Adaptive evolution of four microcephaly genes and the evolution of brain size in anthropoid primates. Mol Biol Evol 28: 625-38. doi: 10.1093/molbev/msq237

Mozzi A, Guerini FR, Forni D, Costa AS, Nemni R, Baglio F, Cabinio M, Riva S, Pontremoli C, Clerici M, Sironi M, Cagliani R (2017) REST, a master regulator of neurogenesis, evolved under strong positive selection in humans and in non human primates. Sci Rep 7: 9530. doi: 10.1038/s41598-017-10245-w

Mugal CF, Wolf JB, Kaj I (2014) Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol Biol Evol 31: 212-31. doi: 10.1093/molbev/mst192

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K (2013) FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol Evol 30: 1196-205. doi: 10.1093/molbev/mst030

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, Scheffler K, Kosakovsky Pond SL (2015) Gene-wide identification of episodic selection. Mol Biol Evol 32: 1365-71. doi: 10.1093/molbev/msv035

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL (2012) Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8: e1002764. doi: 10.1371/journal.pgen.1002764

Myles S, Somel M, Tang K, Kelso J, Stoneking M (2007) Identifying genes underlying skin pigmentation differences among human populations. Hum Genet 120: 613-21. doi: 10.1007/s00439-006-0256-4

Nesse RM, Bergstrom CT, Ellison PT, Flier JS, Gluckman P, Govindaraju DR, Niethammer D, Omenn GS, Perlman RL, Schwartz MD, Thomas MG, Stearns SC, Valle D (2010) Evolution in health and medicine Sackler colloquium: Making evolutionary biology a basic science for medicine. Proc Natl Acad Sci U S A 107 Suppl 1: 1800-7. doi: 10.1073/pnas.0906224106

Nesse RM, Stearns SC (2008) The great opportunity: Evolutionary applications to medicine and public health. Evol Appl 1: 28-48. doi: 10.1111/j.1752-4571.2007.00006.x

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857-68. doi: 10.1038/nrg2187

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15: 1566-75. doi: 10.1101/gr.4252305

Niermeyer S, Zamdio S, Moore LG (2001) *High altitude: an exploration of human adaptation*. The People: pp. 42–100.

Nish S, Medzhitov R (2011) Host defense pathways: role of redundancy and compensation in infectious disease phenotypes. Immunity 34: 629-36. doi: 10.1016/j.immuni.2011.05.009

Nordborg M, Tavare S (2002) Linkage disequilibrium: what history has to tell us. Trends Genet 18: 83-90. doi: 10.1016/s0168-9525(02)02557-x

Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy B, Shriver MD (2007) Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. Mol Biol Evol 24: 710-22. doi: 10.1093/molbev/msl203

Novembre J, Galvani AP, Slatkin M (2005) The geographic spread of the CCR5 Delta32 HIV-resistance allele. PLoS Biol 3: e339. doi: 10.1371/journal.pbio.0030339

Nuttle X, Giannuzzi G, Duyzend MH, Schraiber JG, Narvaiza I, Sudmant PH, Penn O, Chiatante G, Malig M, Huddleston J, Benner C, Camponeschi F, Ciofi-Baffoni S, Stessman HA, Marchetto MC, Denman L, Harshman L, Baker C, Raja A, Penewit K, Janke N, Tang WJ, Ventura M, Banci L, Antonacci F, Akey JM, Amemiya CT, Gage FH, Reymond A, Eichler EE (2016) Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. Nature 536: 205-9. doi: 10.1038/nature19075

O'Bleness M, Searles VB, Varki A, Gagneux P, Sikela JM (2012) Evolution of genetic and genomic features unique to the human lineage. Nat Rev Genet 13: 853-66. doi: 10.1038/nrg3336

Ohno S (1970) Evolution by gene duplication. Springer, New York

Oksenberg N, Stevison L, Wall JD, Ahituv N (2013) Function and regulation of AUTS2, a gene implicated in autism and human evolution. PLoS Genet 9: e1003221. doi: 10.1371/journal.pgen.1003221

Oliveira TY, Harris EE, Meyer D, Jue CK, Silva WA (2012) Molecular evolution of a malaria resistance gene (DARC) in primates. Immunogenetics 64: 497-505. doi: 10.1007/s00251-012-0608-2

Otto SP, Yong P (2002) The evolution of gene duplicates. Adv Genet 46: 451-83. doi: 10.1016/s0065-2660(02)46017-8

Pan HX, Bai HS, Guo Y, Cheng ZY (2019) Bioinformatic analysis of the prognostic value of ZNF860 in recurrence-free survival and its potential regulative network in gastric cancer. Eur Rev Med Pharmacol Sci 23: 162-170. doi: 10.26355/eurrev_201901_16760

Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424: 194-7. doi: 10.1038/nature01771

Paquette AM, Harahap A, Laosombat V, Patnode JM, Satyagraha A, Sudoyo H, Thompson MK, Yusoff NM, Wilder JA (2015) The evolutionary origins of Southeast Asian Ovalocytosis. Infect Genet Evol 34: 153-9. doi: 10.1016/j.meegid.2015.06.002

Parcerisas A, Rubio SE, Muhaisen A, Gómez-Ramos A, Pujadas L, Puiggros M, Rossi D, Ureña J, Burgaya F, Pascual M, Torrents D, Rábano A, Avila J, Soriano E (2014) Somatic signature of brain-specific single nucleotide variations in sporadic Alzheimer's disease. J Alzheimers Dis 42: 1357-82. doi: 10.3233/JAD-140891

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256-60. doi: 10.1038/ng2123

Pervaiz N, Abbasi AA (2016) Molecular evolution of WDR62, a gene that regulates neocorticogenesis. Meta Gene 9: 1-9. doi: 10.1016/j.mgene.2016.02.005

Peter BM, Huerta-Sanchez E, Nielsen R (2012) Distinguishing between selective sweeps from standing variation and from a de novo mutation. PLoS Genet 8: e1003011. doi: 10.1371/journal.pgen.1003011

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, Srinivasan BS, Barsh GS, Myers RM, Feldman MW, Pritchard JK (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826-37. doi: 10.1101/gr.087577.108

Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Williams TN, Weatherall DJ, Hay SI (2010) Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. Nat Commun 1: 104. doi: 10.1038/ncomms1104

Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares M, Vanderhaeghen P, Haussler D (2006) An RNA gene expressed during cortical development evolved rapidly in humans. Nature 443: 167-72. doi: 10.1038/nature05113

Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS (2018) Long reads: their purpose and place. Hum Mol Genet 27: R234-R241. doi: 10.1093/hmg/ddy177

Pollen AA, Bhaduri A, Andrews MG, Nowakowski TJ, Meyerson OS, Mostajo-Radji MA, Di Lullo E, Alvarado B, Bedolli M, Dougherty ML, Fiddes IT, Kronenberg ZN, Shuga J, Leyrat AA, West JA, Bershteyn M, Lowe CB, Pavlovic BJ, Salama SR, Haussler D, Eichler EE, Kriegstein AR (2019) Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. Cell 176: 743-756.e17. doi: 10.1016/j.cell.2019.01.017

Popejoy AB, Fullerton SM (2016) Genomics is failing on diversity. Nature 538: 161-164. doi: 10.1038/538161a

Prabhakar S, Noonan JP, Pääbo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314: 786. doi: 10.1126/science.1130738

Prabhakar S, Visel A, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Morrison H, Fitzpatrick DR, Afzal V, Pennacchio LA, Rubin EM, Noonan JP (2008) Human-specific gain of function in a developmental enhancer. Science 321: 1346-50. doi: 10.1126/science.1159974

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20: R208-15. doi: 10.1016/j.cub.2009.11.055

Przeworski M (2002) The signature of positive selection at randomly chosen loci. Genetics 160: 1179-89.

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. Evolution 59: 2312-23.

Przeworski M, Hudson RR, Di Rienzo A (2000) Adjusting the focus on human variation. Trends Genet 16: 296-302. doi: 10.1016/s0168-9525(00)02030-8

Qian W, Liao BY, Chang AY, Zhang J (2010) Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet 26: 425-30. doi: 10.1016/j.tig.2010.07.002

Quillen EE, Bauchet M, Bigham AW, Delgado-Burbano ME, Faust FX, Klimentidis YC, Mao X, Stoneking M, Shriver MD (2012) OPRM1 and EGFR contribute to skin pigmentation differences between Indigenous Americans and Europeans. Hum Genet 131: 1073-80. doi: 10.1007/s00439-011-1135-1

Racimo F (2016) Testing for Ancient Selection Using Cross-population Allele Frequency Differentiation. Genetics 202: 733. doi: 10.1534/genetics.115.178095

Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5: 28. doi: 10.1186/1471-2148-5-28

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-7. doi: 10.1126/science.273.5281.1516

Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA (2005) Ancient and recent positive selection transformed opioid cis-regulation in humans. PLoS Biol 3: e387. doi: 10.1371/journal.pbio.0030387

Rodriguez-Murillo L, Greenberg DA (2008) Genetic association analysis: a primer on how it works, its strengths and its weaknesses. Int J Androl 31: 546-56. doi: 10.1111/j.1365-2605.2008.00896.x

Rumold CU, Aldenderfer MS (2016) Late Archaic-Early Formative period microbotanical evidence for potato at Jiskairumoko in the Titicaca Basin of southern Peru. Proc Natl Acad Sci U S A 113: 13672-13677. doi: 10.1073/pnas.1604265113

Ruwende C, Khoo SC, Snow RW, Yates SN, Kwiatkowski D, Gupta S, Warn P, Allsopp CE, Gilbert SC, Peschu N (1995) Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature 376: 246-9. doi: 10.1038/376246a0

S O (1970) Evolution by gene duplication. Springer, New York

Sabeti P, Usen S, Farhadian S, Jallow M, Doherty T, Newport M, Pinder M, Ward R, Kwiatkowski D (2002a) CD40L association with protection from severe malaria. Genes Immun 3: 286-91. doi: 10.1038/sj.gene.6363877

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES (2002b) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-7. doi: 10.1038/nature01140

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, International HapMap C, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-8. doi: 10.1038/nature06250

Sambo MR, Trovoada MJ, Benchimol C, Quinhentos V, Gonçalves L, Velosa R, Marques MI, Sepúlveda N, Clark TG, Mustafa S, Wagner O, Coutinho A, Penha-Gonçalves C (2010) Transforming growth factor beta 2 and heme oxygenase 1 genes are risk factors for the cerebral malaria syndrome in Angolan children. PLoS One 5: e11141. doi: 10.1371/journal.pone.0011141

Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber CM, Saragosti S, Lapoumeroulie C, Cognaux J, Forceille C, Muyldermans G, Verhofstede C, Burtonboy G, Georges M, Imai T, Rana S, Yi Y, Smyth RJ, Collman RG, Doms RW, Vassart G, Parmentier M (1996) Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 382: 722-5. doi: 10.1038/382722a0

Saunders MA, Hammer MF, Nachman MW (2002) Nucleotide variability at G6pd and the signature of malarial selection in humans. Genetics 162: 1849-61.

Saunders MA, Slatkin M, Garner C, Hammer MF, Nachman MW (2005) The extent of linkage disequilibrium caused by selection on G6PD in humans. Genetics 171: 1219-29. doi: 10.1534/genetics.105.048140

Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet 19: 491-504. doi: 10.1038/s41576-018-0016-z

Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Woldemeskel D, Beggs W, Lambert C, Jarvis JP, Abate D, Belay G, Tishkoff SA (2012) Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol 13: R1. doi: 10.1186/gb-2012-13-1-r1

Scheinfeldt LB, Tishkoff SA (2013) Recent human adaptation: genomic approaches, interpretation and insights. Nat Rev Genet 14: 692-702. doi: 10.1038/nrg3604

Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156: 879-91. doi: 10.1093/genetics/156.2.879

Schlebusch CM, Lewis CM, Jr., Vahter M, Engstrom K, Tito RY, Obregon-Tito AJ, Huerta D, Polo SI, Medina AC, Brutsaert TD, Concha G, Jakobsson M, Broberg K (2013) Possible positive selection for an arsenic-protective haplotype in humans. Environ Health Perspect 121: 53-8. doi: 10.1289/ehp.1205504

Schmitz S, Thomas PD, Allen TM, Poznansky MJ, Jimbow K (1995) Dual role of melanins and melanin precursors as photoprotective and phototoxic agents: inhibition of ultraviolet radiation-induced lipid peroxidation. Photochem Photobiol 61: 650-5.

Schwartz JJ, Roach DJ, Thomas JH, Shendure J (2014) Primate evolution of the recombination regulator PRDM9. Nat Commun 5: 4370. doi: 10.1038/ncomms5370

Sharp PM, Bailes E, Chaudhuri RR, Rodenburg CM, Santiago MO, Hahn BH (2001) The origins of acquired immune deficiency syndrome viruses: where and when? Philos Trans R Soc Lond B Biol Sci 356: 867-76. doi: 10.1098/rstb.2001.0863

Shriner D, Nickle DC, Jensen MA, Mullins JI (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genet Res 81: 115-21. doi: 10.1017/s0016672303006128

Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, Huang J, Akey JM, Jones KW (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics 1: 274-86.

Sikela JM (2006) The jewels of our genome: the search for the genomic changes underlying the evolutionarily unique capacities of the human brain. PLoS Genet 2: e80. doi: 10.1371/journal.pgen.0020080

Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH (1994) Receptor and ligand domains for invasion of erythrocytes by Plasmodium falciparum. Science 264: 1941-4. doi: 10.1126/science.8009226

Simonson TS, McClain DA, Jorde LB, Prchal JT (2012) Genetic determinants of Tibetan high-altitude adaptation. Hum Genet 131: 527-33. doi: 10.1007/s00439-011-1109-3

Simonson TS, Yang Y, Huff CD, Yun H, Qin G, Witherspoon DJ, Bai Z, Lorenzo FR, Xing J, Jorde LB, Prchal JT, Ge R (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329: 72-5. doi: 10.1126/science.1189406

Singh BD, Singh AK (2015) Association mapping. *Marker-Assisted Plant Breeding: Principles and Practices*. Springer, New Delhi, pp pp. 217-256

Sirugo G, Williams SM, Tishkoff SA (2019) The Missing Diversity in Human Genetic Studies. Cell 177: 1080. doi: 10.1016/j.cell.2019.04.032

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL (2015) Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol 32: 1342-53. doi: 10.1093/molbev/msv022

Stein MB, Chen CY, Ursano RJ, Cai T, Gelernter J, Heeringa SG, Jain S, Jensen KP, Maihofer AX, Mitchell C, Nievergelt CM, Nock MK, Neale BM, Polimanti R, Ripke S, Sun X, Thomas ML, Wang Q, Ware EB, Borja S, Kessler RC, Smoller JW, Collaborators AStARaRiSS (2016) Genome-wide Association Studies of Posttraumatic Stress Disorder in 2 Cohorts of US Army Soldiers. JAMA Psychiatry 73: 695-704. doi: 10.1001/jamapsychiatry.2016.0350

Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L, Gerrard B, Malasky M, Ramos MD, Morlot S, Tzetis M, Oddoux C, di Giovine FS, Nasioulas G, Chandler D, Aseev M, Hanson M, Kalaydjieva L, Glavac D, Gasparini P, Kanavakis E, Claustres M, Kambouris M, Ostrer H, Duff G, Baranov V, Sibul H, Metspalu A, Goldman D, Martin N, Duffy D, Schmidtke J, Estivill X, O'Brien SJ, Dean M (1998) Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes. Am J Hum Genet 62: 1507-15. doi: 10.1086/301867

Sudmant PH, Huddleston J, Catacchio CR, Malig M, Hillier LW, Baker C, Mohajeri K, Kondova I, Bontrop RE, Persengiev S, Antonacci F, Ventura M, Prado-Martinez J, Marques-Bonet T, Eichler EE, Project GAG (2013) Evolution and diversity of copy number variation in the great ape lineage. Genome Res 23: 1373-82. doi: 10.1101/gr.158543.113

Sugden LA, Atkinson EG, Fischer AP, Rong S, Henn BM, Ramachandran S (2018) Localization of adaptive variants in human genomes using averaged one-dependence estimation. Nat Commun 9: 703. doi: 10.1038/s41467-018-03100-7

Suktitipat B, Naktang C, Mhuantong W, Tularak T, Artiwet P, Pasomsap E, Jongjaroenprasert W, Fuchareon S, Mahasirimongkol S, Chantratita W, Yimwadsana B, Charoensawan V, Jinawath N (2014) Copy number variation in Thai population. PLoS One 9: e104355. doi: 10.1371/journal.pone.0104355

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F, Detours V, Vanderhaeghen P (2018) Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. Cell 173: 1370-1384 e16. doi: 10.1016/j.cell.2018.03.067

Szpak M, Mezzavilla M, Ayub Q, Chen Y, Xue Y, Tyler-Smith C (2018) FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. Genome Biol 19: 5. doi: 10.1186/s13059-017-1380-2

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-95.

Tak YG, Farnham PJ (2015) Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin 8: 57. doi: 10.1186/s13072-015-0050-4

Tashi T, Scott Reading N, Wuren T, Zhang X, Moore LG, Hu H, Tang F, Shestakova A, Lorenzo F, Burjanivova T, Koul P, Guchhait P, Wittwer CT, Julian CG, Shah B, Huff CD, Gordeuk VR, Prchal JT, Ge R (2017) Gain-of-function EGLN1 prolyl hydroxylase (PHD2 D4E:C127S) in combination with EPAS1 (HIF-2α) polymorphism lowers hemoglobin concentration in Tibetan highlanders. J Mol Med (Berl) 95: 665-670. doi: 10.1007/s00109-017-1519-3

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31-40. doi: 10.1038/ng1946

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J, Piro A, Stoneking M, Tagarelli A, Tagarelli G, Touma EH, Williams SM, Clark AG (2001)

Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science 293: 455-62. doi: 10.1126/science.1061573

Tolia NH, Enemark EJ, Sim BK, Joshua-Tor L (2005) Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite Plasmodium falciparum. Cell 122: 183-93. doi: 10.1016/j.cell.2005.05.033

Trujillo CA, Rice ES, Schaefer NK, Chaim IA, Wheeler EC, Madrigal AA, Buchanan J, Preissl S, Wang A, Negraes PD, Szeto RA, Herai RH, Huseynov A, Ferraz MSA, Borges FS, Kihara AH, Byrne A, Marin M, Vollmers C, Brooks AN, Lautz JD, Semendeferi K, Shapiro B, Yeo GW, Smith SEP, Green RE, Muotri AR (2021) Reintroduction of the archaic variant of. Science 371. doi: 10.1126/science.aax2537

van de Bunt M, Cortes A, Brown MA, Morris AP, McCarthy MI, Consortium I (2015) Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. PLoS Genet 11: e1005535. doi: 10.1371/journal.pgen.1005535

Verrelli BC, McDonald JH, Argyropoulos G, Destro-Bisol G, Froment A, Drousiotou A, Lefranc G, Helal AN, Loiselet J, Tishkoff SA (2002) Evidence for balancing selection from nucleotide sequence analyses of human G6PD. Am J Hum Genet 71: 1112-28. doi: 10.1086/344345

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. Annu Rev Genet 47: 97-120. doi: 10.1146/annurev-genet-111212-133526

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72. doi: 10.1371/journal.pbio.0040072

Wang J, Huang D, Zhou Y, Yao H, Liu H, Zhai S, Wu C, Zheng Z, Zhao K, Wang Z, Yi X, Zhang S, Liu X, Liu Z, Chen K, Yu Y, Sham PC, Li MJ (2020) CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. Nucleic Acids Res 48: D807-D816. doi: 10.1093/nar/gkz1026

Wang M, Huang X, Li R, Xu H, Jin L, He Y (2014) Detecting recent positive selection with high accuracy and reliability by conditional coalescent tree. Mol Biol Evol 31: 3068-80. doi: 10.1093/molbev/msu244

Wang YQ, Su B (2004) Molecular evolution of microcephalin, a gene determining human brain size. Hum Mol Genet 13: 1131-7. doi: 10.1093/hmg/ddh127

Wiener P, Pong-Wong R (2011) A regression-based approach to selection mapping. J Hered 102: 294-305. doi: 10.1093/jhered/esr014

Wilder JA, Stone JA, Preston EG, Finn LE, Ratcliffe HL, Sudoyo H (2009) Molecular population genetics of SLC4A1 and Southeast Asian ovalocytosis. J Hum Genet 54: 182-7. doi: 10.1038/jhg.2009.12

Williams GC, Nesse RM (1991) The dawn of Darwinian medicine. Q Rev Biol 66: 1-22. doi: 10.1086/417048

Wisser RJ, Murray SC, Kolkman JM, Ceballos H, Nelson RJ (2008) Selection mapping of loci for quantitative disease resistance in a diverse maize population. Genetics 180: 583-99. doi: 10.1534/genetics.108.090118

Wooding SP, Watkins WS, Bamshad MJ, Dunn DM, Weiss RB, Jorde LB (2002) DNA sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 gene: implications for human population history and natural selection. Am J Hum Genet 71: 528-42. doi: 10.1086/342260

Wright S (1950) Genetical structure of populations. Nature 166: 247-9.

Wu DD, Li GM, Jin W, Li Y, Zhang YP (2012) Positive selection on the osteoarthritis-risk and decreased-height associated variants at the GDF5 gene in East Asians. PLoS One 7: e42553. doi: 10.1371/journal.pone.0042553

Xu K, Schadt EE, Pollard KS, Roussos P, Dudley JT (2015) Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions. Mol Biol Evol 32: 1148-60. doi: 10.1093/molbev/msv031

Xu Z, Wu C, Wei P, Pan W (2017) A Powerful Framework for Integrating eQTL and GWAS Summary Data. Genetics 207: 893-902. doi: 10.1534/genetics.117.300270

Yakub I, Lillibridge KM, Moran A, Gonzalez OY, Belmont J, Gibbs RA, Tweardy DJ (2005) Single nucleotide polymorphisms in genes for 2'-5'-oligoadenylate synthetase and RNase L inpatients hospitalized with West Nile virus infection. J Infect Dis 192: 1741-8. doi: 10.1086/497340

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, Zheng H, Liu T, He W, Li K, Luo R, Nie X, Wu H, Zhao M, Cao H, Zou J, Shan Y, Li S, Yang Q, Asan, Ni P, Tian G, Xu J, Liu X, Jiang T, Wu R, Zhou G, Tang M, Qin J, Wang T, Feng S, Li G, Huasang, Luosang J, Wang W, Chen F, Wang Y, Zheng X, Li Z, Bianba Z, Yang G, Wang X, Tang S, Gao G, Chen Y, Luo Z, Gusang L, Cao Z, Zhang Q, Ouyang W, Ren X, Liang H, Zheng H, Huang Y, Li J, Bolund L, Kristiansen K, Li Y, Zhang Y, Zhang X, Li R, Li S, Yang H, Nielsen R, Wang J, Wang J (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75-8. doi: 10.1126/science.1190371

Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, Valmeekam V, Retief J, Matsuzaki H, Taub M, Seielstad M, Kennedy GC (2006) A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics 22: 2122-8. doi: 10.1093/bioinformatics/btl365

Zhang J (2003) Evolution of the human ASPM gene, a major determinant of brain size. Genetics 165: 2063-70. doi: 10.1093/genetics/165.4.2063

Zhao Z, Jin L, Fu YX, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB, Ramos-Onsins S, Yu N, Li WH (2000) Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc Natl Acad Sci U S A 97: 11354-8. doi: 10.1073/pnas.200348197

Zhou L, Zhao F (2018) Prioritization and functional assessment of noncoding variants associated with complex diseases. Genome Med 10: 53. doi: 10.1186/s13073-018-0565-y

Chapter 3
Functional Fates of *NBPF* Duplicate Genes in Human and Non-Human Primate Corticogenesis

3.1 INTRODUCTION

A major feature of human brain evolution is the exceptional enlargement of the neocortex, particularly of the prefrontal cortex, relative to size scaling predictions for a primate brain (Smaers et al. 2017). The biological tradeoff of this cortical expansion, which provides the anatomical substrate for further adaptations in neuronal circuitry underlying human-specific cognitive traits, is an increased burden of neurodevelopmental and neuropsychiatric disorders (Crow 2000; Dennis and Eichler 2016; Dennis et al. 2017; Sikela and Searles Quick 2018). Segmental duplication is a rich source of highly plastic genomic variation important for human-specific adaptive evolution as well as a major cause of developmental disorders (Crow 2000; Dennis and Eichler 2016; Dennis et al. 2017; Sikela and Searles Quick 2018). Gene duplicates that arise from segmental duplication are a common basis of novel biological functions that underlie phenotypic variation (Dennis and Eichler 2016; Kronenberg et al. 2018). Functional duplicate genes expressed in the developing brain, specific to the human lineage and disrupted in neurodevelopmental disorders, are great candidates for the molecular basis of human-specific features of brain development. One such candidate, the neuroblastoma breakpoint gene family (*NBPF*), has undergone significant gene duplication in primate genomes, coupled with intragenic expansion in the human genome (Figure 3.1A, 3.1B, and 3.1C) (Astling et al. 2017; Dumas et al. 2007; O'Bleness et al. 2014), yet the molecular functions of NBPF proteins remain unknown.

Phylogenetic evidence indicates that a single ancestral copy of *NBPF* first evolved in placental mammals and was lost in the rodent lineage prior to duplication by segmental duplication in the primate order. In the human genome, there are 24 *NBPF* paralogs, all of which are located on chromosome 1 and enriched in the 1q21 region (Figure 3.1B) (Astling et al. 2017). Among these, four human-specific (hs) duplicates (*NBPF10, NBPF14, NBPF19, NBPF20*) arose from human-specific segmental duplications (hsSDs), and three paralogs have human-specific changes (*NBPF9, NBPF12*, *NBPF26*). Three duplicative transposition events involving three hs*NBPF* genes (*NBPF10, NBPF14, NBPF19*) have led to the placement of each of these genes adjacent to a hs*NOTCH2NL* gene (Figure 3.1B) (Fiddes et al. 2019). HsNO*TCH2NL* genes, which have been implicated in delayed differentiation during human corticogenesis (Fiddes et al. 2018; Suzuki et al. 2018), show evidence of coregulation with hs*NBPF* genes in human neural progenitor cells (hNPCs), potentially suggesting paired, dosage-related involvement of these genes in human brain evolution (Fiddes et al. 2019). This expression profile of hs*NBPF* genes is consistent with previous RNA *in situ* hybridizations in fetal brain showing *NBPF* enrichment at the ventricular zone (Keeney et al. 2015a). *De novo* heterozygous distal microdeletions and microduplications of 1q21.1 in humans harboring the hs*NBPF* genes have been linked to several neurodevelopmental disorders, most often including microcephaly and macrocephaly, respectively (Bernier et al. 2016; Dennis et al. 2017; Linden et al. 2021). Although a few candidate genes within this interval have been proposed to contribute to the clinical brain phenotypes in individuals with 1q21.1 deletion/duplication syndrome (1q21DDS), (e.g., hs*NOTCH2NL* genes and *HYDIN2*)*,* there are nevertheless cases where the dosage of these candidates are not affected due to differential breakpoints (Dougherty et al. 2017; Fiddes et al. 2018). While *NBPF* genes have yet to be implicated in the 1q21 disease etiology, *NBPF* copy number appears to be altered in all cases, and *NBPF* sequence has been associated with brain size variation across primates, as well as autism and schizophrenia severity (Davis et al. 2014; Keeney et al. 2015a; Keeney et al. 2014). This suggests that *NBPF* genes are important for neural development.

NBPF genes predominantly encode DUF1220 (Olduvai) protein domains of uncharacterized function that are roughly 65 amino acids in length, as well as coiled coil domains at their N-terminus (O'Bleness et al. 2012; Popesco et al. 2006). Phylogenetic analysis of NBPF-DUF1220 domains has identified six distinct clades based on sequence similarity, which have been termed CON1, CON2, HLS1, HLS2, HLS3, and CON3 (O'Bleness et al. 2012; Popesco et al. 2006). While the copy number of DUF1220 domains is variable across paralogs, interestingly, their order is highly conserved (Figure 3.1B) (O'Bleness et al. 2012; Popesco et al. 2006). NBPF-DUF1220 sequences have undergone amplification exclusively in primates, with a trend towards increased copy number with increasing phylogenetic relatedness to humans, ranging from 1-8 copies in non-primate mammals to 48-75 copies in monkeys to 97-138 copies in non-human great apes to over 300 copies in humans (O'Bleness et al. 2012; Popesco et al. 2006; Zimmer and Montgomery 2015). Tandem intragenic duplication of a three-domain block (HLS1-HLS2-HLS3) contributes to the exceptional expansion of DUF1220 domains in hs*NBPF* paralogs relative to ancestral *NBPF* paralogs, making DUF1220 expansion the greatest human-specific copy number increase of coding sequence in the human genome (O'Bleness et al. 2012; Popesco et al. 2006; Zimmer and Montgomery 2015). The evolutionary hyper-amplification of the triplet DUF1220 sequence in hs*NBPF* paralogs was likely driven by nonallelic homologous recombination via a G-quadraplex-based mechanism that promotes genomic instability at *NBPF* loci and underlies extant copy number variation of NBPF-DUF1220 triplets in the human genome (Heft et al. 2020). While there is a distinct human-specific range in copy number of NBPF-DUF1220 domains compared to other species, the domains remain polymorphic in the general human population as well as in clinical cohorts (Davis et al. 2014). Clinically, increased NBPF-DUF1220 copy number has been linearly associated with autism severity as well as severity of ‘negative’ symptoms of schizophrenia (e.g., social deficits), whereas reduced NBPF-DUF1220 copy number has been linearly associated with severity of ‘positive’ symptoms of schizophrenia (e.g., hallucinations, delusions) (Davis et al. 2019; Davis et al. 2015; Davis et al. 2014; Quick et al. 2016).

Despite extensive progress in the genomic annotation of *NBPF* genes in the human genome and improved *NBPF* paralog expression profiling, very little functional data on NBPF proteins exists. This is largely due to the extreme technical difficulty of investigating evolutionarily young duplicates of high sequence identity comprised of copious repetitive elements. Nevertheless, research efforts have implicated NBPF1 overexpression in apoptosis via regulation of PI3K/mTOR signaling in HeLa cells (Qin et al. 2016), while overexpression of NBPF15 has been shown to increase proliferation of human neural stem cells (Keeney et al. 2015a). Here, we expand on the latter finding to test the hypothesis that NBPF-DUF1220 dosage coupled with changes in expression dynamics underlie variations in cortical development that distinguish humans from other primates.

Increase in the dosage of DUF1220 protein domains has been proposed as a molecular contributor to cortical expansion in primates, with the most exaggerated effect in humans (Keeney et al. 2014). However, this has not been functionally tested. To gain insight into the function of NBPF proteins in corticogenesis, we leveraged cortical organoid models and cultured NPCs. Our goals were to characterize molecular biology of endogenous NBPF proteins and identify interspecific differences in cortical growth related to DUF1220 dosage using overexpression and deletion approaches. While we initially pursued hs*NBPF14* for overexpression modeling given its high endogenous expression in hNPCs and altered dosage in 1q21DDS, we had to switch experiments to human NBPF15 due to significant technical hurdles with the molecular cloning of *NBPF14*. *NBPF15*, which is exclusive to primates, does not have the same expression profile as *NBPF14* and encodes fewer DUF1220 protein domains. However, it still allows us to utilize *NBPF15* ectopic overexpression as a proxy for DUF1220-expanded hs*NBPF* expression and characterize interactions with ancestral NBPF protein in NPCs. For investigation of depletion of NBPF, hs*NBPF14* was targeted. Our findings highlight species differences important for interpreting NBPF function as well as further implicate *NBPF* genes in the etiology of 1q21 neurodevelopmental disorders. This represents the first study to utilize human and chimpanzee cortical organoids to investigate NBPF function in brain development.

3.2 RESULTS

3.2.1 *NBPF* paralog expression differences in corticogenesis support *cis*-regulatory divergence during human and non-human primate brain evolution

A common phenomenon underlying gene retention following duplication in evolution is alteration of the expression profile, or transcriptional fate, of duplicate genes. *Cis*-regulatory divergence is pervasive across duplicate genes that arose from hsSDs (Shew et al. 2021). To investigate this in the context of *NBPF* genes within humans, we first analyzed recently published transcriptomics data in lymphoblastoid cell lines (LCL) from 463 individuals and assessed *NBPF* paralog expression differences with isoform variation across individuals (Shew et al. 2021). We observed differences in expression based on paralog, with *NBPF4* and *NBPF6* (representing the most ancestral protein-coding *NBPF* genes) showing low to no expression compared to the rest, which may reflect changes in *cis* regulatory elements between paralogs (Figure 3.2). Second, based on short read RNA sequencing data deposited in the GTEX portal, *NBPF4* and *NBPF6* show enrichment of expression in the testes with little to no expression in other tissues (data not shown). Conversely, *NBPFs* that underwent subsequent gene duplication during primate evolution, which resulted in the acquisition of a distinct promoter obtained from the gene *EVI5* (*pEVI5*), show low, ubiquitous expression across all adult tissues (data not shown). Expression profiles of *pEVI5-NBPF* and *EVI5* were comparable (data not shown). Together, this data suggests changes to the spatial and temporal specificity of *NBPF* paralog expression that underlie expression profiles, supporting *cis*-regulatory divergence.

During human *NBPF* evolution, the loss of the *EVI5* promoter with the acquisition of new sequence occurs in a subset of human *NBPF* genes (Figure 3.1B and 3.1C) (Fiddes et al. 2019). Shared expression profiles have been previously observed between these hs*NBPF* and hs*NOTCH2NL* in hNPCs suggestive of shared regulatory elements as results of their duplicative evolution (Fiddes et al. 2019). Given this finding, we compared scRNAseq data across human and chimpanzee cerebral organoid models deposited in the UCSC Cell Browser and scAPEX to identify differences in *NBPF* gene expression between species (data not shown). Consistent with above findings, *NBPF4* and *NBPF6* showed no evidence of expression in cortical organoids of both species. In human organoids, *NBPFs* with orthologs in chimpanzees showed overall low expression but slightly higher than in chimpanzee organoids. Lastly, chimpanzees showed no expression of hs*NBPF*s. hsNBPF14 showed the most marked increase in expression in organoids with an enrichment in cortical NPCs. To assess paralog expression patterns across cell types in primary tissue, we obtained gene-level quantifications from primary fetal brain (Nowakowski et al. 2017) (Figure 3.1D). hs*NBPF14* and hs*NBPF10* are highly expressed in human mitotic NPCs and outer radial glia (oRG), with notable enrichment in ventricular radial glia (vRG) (Figure 3.1D). Additionally, a general trend towards increased expression in vRGs and oRGs was noted for other *NBPF*s except for *NBPF4* and *NBPF6.* However, it is important to note that while the older NBPF paralogs have unique sequence that is captured in short read scRNAseq data allowing for more accurate annotation of expression, the human-specific *NBPF* genes do not and the data may reflect multi-mapping issues. Nevertheless, this data does indicate that hs*NBPF*s are expressed in NPCs.

Together, *cis*-regulatory differences across *NBPF* genes indicate that younger duplicates in primates obtain ubiquitous, low expression across tissues, including the brain. Subsequently, hs*NBPF* genes underwent additional *cis*-regulatory evolution that enabled higher expression in NPCs in the developing human cerebral cortex.

3.2.2 NBPF proteins localize to the cytoplasm with enriched expression in mitotic hNPCs

While RNA expression of *NBPF* genes has shown enrichment in cortical NPCs, expression on a protein level has not been assessed. To characterize endogenous NBPF protein expression, we differentiated human and chimpanzee iPSCs to 3D cerebral organoids and monolayer cortical NPC cultures (Figure 3.3A). Protein expression and localization was performed using the Invitrogen pan-NBPF-DUF1220 antibody (Figure S3.1A). We confirmed specificity of the pan-NBPF-DUF1220 antibody to NBPF proteins by overexpression of NBPF15-mCherry in hNPCs followed by co-immunostaining (Figure S3.1B). Protein expression was assessed across developmental time at days 28, 56, and 84 in cortical organoids differentiated from three human induced pluripotent stem cell (iPSC) lines derived from two female individuals (H20682^{WT}, H28834^{WT}) and one male individual (H23555^{WT}) together with three chimpanzee iPSC lines derived from two female individuals (C40280^{WT}, C3647^{WT}) and one male individual (C3649^{WT}) (Figure 3.3A). All six iPSC lines were reprogrammed and passaged together to reduce batch effects and improve interspecific comparisons (Gallego Romero et al. 2015). By immunostaining, protein expression profiling corroborates the transcriptomics findings (Figure 3.3B) and implicates NBPF function at several stages of corticogenesis.

Subcellular localization of endogenous NBPF proteins was characterized in monolayer cortical NPCs differentiated from the six human and chimpanzee iPSC lines using the pan-NBPF-DUF1220 antibody. We observed cytoplasmic localization of NBPF proteins at the Golgi apparatus, microtubules, the mitotic spindle apparatus, and the midbody of mitotic hNPCs, which may suggest a role in abscission (Figure 3.3C and S3.1C). There was a notable enrichment in NBPF signal in mitotic hNPCs compared to interphase hNPCs. Conversely, we see little to no NBPF signal in interphase chimpanzee NPCs (cNPCs), consistent with transcriptomic findings (Figure 3.3D). However, a slight signal was observed at the metaphase plate as well as at the spindle poles during anaphase in mitotic cNPCs, suggesting lowly expressed NBPF proteins that are not human-specific may function during cell division in primate cortical NPCs. This potentially supports the hypothesis that hs*NBPF* DUF1220-expanded proteins have retained a shared, ancestral function with other primate *NBPF* paralogs.

Since NBPF proteins are almost exclusively comprised of DUF1220 protein domains of uncharacterized function, we searched the literature for other DUF1220-containing proteins. The only other gene in the human genome containing a DUF1220 protein domain is *PDE4DIP*, or Myomegalin, which encodes a single copy of DUF1220 that is distinct in sequence identity (by ~70%) from NBPF-DUF1220 sequences (O'Bleness et al. 2014; O'Bleness et al. 2012). The PDE4DIP-DUF1220 domain may be the ancestral sequence of NBPF-DUF1220 domains based on a minimum evolution-based consensus phylogenetic tree on all DUF1220 sequences in the human genome (Keeney et al. 2015b). Shared subcellular localization pattern of NBPF proteins with

PDE4DIP raises the possibility that NBPF proteins may retain putative ancestral functions of PDE4DIP via DUF1220 domains in NPCs.

3.2.3 NBPF proteins putatively interact with proteins important for NPC proliferation

To explore molecular functions of NBPF proteins in cortical NPCs, we tagged human NBPF15 for controlled immunostaining and pull-down experiments. NBPF15 represents a primate-specific ortholog, and while we sought to investigate hsNBPF14 given its copy number expansion, enrichment in NPCs, and disruption in 1q21DDS, we ultimately had to settle on NBPF15 overexpression after significant technical hurdles with isolating and cloning hs*NBPF14* sequence into an expression plasmid. However, using NBPF15 overexpression as a proxy for increased DUF1220 expression still allows us to test for a DUF1220 dosage effect. We used a piggybac tetracycline-on system ($H9^{iNBPF15\text{-}FLAG}$) to generate inducible hESC lines, followed by differentiated to hNPCs for induction together with uninduced, empty vector control hESCs ($H9^{EV}$) (Figure 3.4A). Following 48 hours of doxycycline-induction, replicate $H9^{iNBPF15\text{-}FLAG}$ and $H9^{EV}$ NPCs were collected for immunostaining as well as FLAG-immunoprecipitation followed by mass-spectrometry (IP-MS) to identify putative protein interactors. FLAG-immunostaining confirmed cytoplasm localizations observed with pan-NBPF-DUF1220 antibody, with an enrichment of signal at the spindle apparatus and near the midbody of dividing cells (Figure 3.4B). There is also a similar localization of NBPF15-FLAG in surrounding interphase cells as described for DUF1220-containing PDE4DIP, consistent with Golgi association and microtubules (Figure 3.4B). Given microtubule localizations throughout the cell cycle, NBPF proteins could help facilitate the protracted window of symmetric divisions of vRG by ensuring proper spindle orientation during mitosis that distinguish humans from chimpanzees, coupled with establishment of proper apicobasal polarity upon which both interkinetic nuclear migration of vRG and mitotic somal translocation of oRG depend for proliferative divisions.

To identify proteins that interact with NBPFs at the Golgi, microtubules, and the midbody, we prioritized IP-MS hits with greater than two-fold enrichment in bait samples based on either peptide sequence matches or abundance measures (Figure 3.4C). To exclude potential false positives, top hits were filtered through an online repository of common protein contaminants in AP-MS experiments (CRAPome), resulting in a list of 114 proteins. Gene Ontology Pathway Analysis was performed on the final list, and enriched categories of interest included cell-cell adhesion, regulation of apoptosis in brain development, proliferation, and mitotic spindle organization (Figure 3.4D). Using a more stringent filtering of less than 10% in the CRAPome database, STRING Network analysis was performed to identify enriched protein-protein networks (Figure 3.4E). From this analysis, we identified enriched interaction networks of proteins associated with the extracellular matrix (TNC, FN1, and LGALS1) and microtubules/filaments (NES, DCX, and TUBB3). Both the extracellular environment as well as cytoskeletal organization is important for proper development and maintenance of the NPC architecture upon which neuronal migration relies, and disruptions can cause several cortical malformations including premature differentiation leading to microcephaly (Ferent et al. 2020). Further, the delayed shift from proliferative to neurogenic cell divisions in the developing human cortex relative to chimpanzees is partly regulated by temporal differences in establishment of extracellular matrix and cytoskeletal organizations—biology which may be afforded by increased NBPF proteins in hNPCs. Interestingly, other NBPF proteins (NBPF1 and NBPF9) interact with NBPF15, validated by replicate co-immunoprecipitation of NBPF15 at ~70 kD with NBPF proteins at ~130 kD (Figure 3.4F). While NBPF proteins remain unassociated with any interactors in current public databases, highlighting the paucity of data on NBPF proteins, this finding of NBPF-NBPF interactions has important implications for novel biology in hNPCs.

For species comparisons, we generated inducible chimpanzee iPSC lines (C40280$^{iNBPF15\text{-}FLAG}$, C3647$^{iNBPF15\text{-}FLAG}$) to express hNBPF15-FLAG in differentiated cNPCs together with empty vector controls. In cNPCs, NBPF15-FLAG IP-MS identified far fewer enriched proteins and no specific proteins overlapped between human and chimpanzee datasets (Figure 3.4G). However, enriched functional categories of hits included regulation of cell growth, intra-Golgi and retrograde Golgi-to-ER traffic, and remodeling of extracellular matrix—findings consistent with hNPC observations (Figure 3.4H). While not identified in hNPCs, top candidate interactors in cNPCs highlight potential shared functions with hNPCs at mitotic biology and microtubule dynamics: KLHL9 (regulates AURKB dynamics on mitotic chromosomes and influences mitotic progression and cytokinesis completion (Maerki et al. 2009; Sumara et al. 2007)), ARFGAP2 (involved in protein transport between the Golgi and endoplasmic reticulum (Frigerio et al. 2007)), PAFAH1B3 (involved in neuronal migration and implicated in lissencephaly (Sweeney et al. 2000)), and CEP128 (centriole and spindle pole associated protein that functions in neural development (Mönnich et al. 2018)). Nevertheless, the fewer detected interacting proteins may reflect differences between NBPF-mediated biology in a human versus chimpanzee cellular background due to a greater impact of NBPF dosage in humans, which enhances the proliferative capacity of vRG and oRG hNPCs.

3.2.4 Ectopic expression of *NBPF15* in human and chimpanzee cortical organoids suggests critical dosage threshold necessary for elevated proliferation

To test the hypothesis that *NBPF* dosage impacts human neural development, we evaluated the effect of ectopic human *NBPF* overexpression in chimpanzee cortical organoids and overexpression in human cortical organoids. Again, although *NBPF15* lacks the extreme triplet amplification of hs*NBPF* that makes up the majority of interspecific DUF1220 copy number differences, we sought to utilize NBPF15 overexpression throughout development as a proxy for increased expression of DUF1220-expanded sequence. We generated three inducible chimpanzee iPSC lines (C40280$^{iNBPF15\text{-}3xFLAG}$, C3647$^{iNBPF15\text{-}3xFLAG}$, C3649$^{iNBPF15\text{-}3xFLAG}$) to ectopically express human *NBPF15* together with their respective unedited control cell lines (C40280^{WT}, C3647^{WT}, C3649^{WT}). We also generated three inducible human iPSC lines (H20682$^{iNBPF15\text{-}3xFLAG}$, H28834$^{iNBPF15\text{-}3xFLAG}$, H23555$^{iNBPF15\text{-}3xFLAG}$) for *NBPF15* overexpression along with respective unedited controls (H20682^{WT}, H28834^{WT}, H23555 ^{WT}).

We differentiated all control and inducible NBPF15-overexpression iPSC lines to cortical organoids across two differentiation replicates per condition. Organoids were induced for overexpression with doxycycline at day 14 in neural development with fresh doxycycline added to media every 48 hours (Figure 3.5A). Cross section organoid area was measured weekly starting at day 28 neural differentiation (ND) through day 56ND. Given that the human genome contains ~115 additional DUF1220 haploid copies relative to the chimpanzee genome (Fiddes et al. 2019), coupled with enriched expression in the developing brain, we predicted that *NBPF* overexpression in a DUF1220-rich genomic background (human) would have a marginal effect on cortical growth compared to overexpression in a DUF1220-low genetic background (chimpanzee). Unexpectedly, based on overall size, we observed a stronger effect in human cortical organoids starting at day 42ND (*p*=<0.0001, unpaired two-tailed *t* test) (Figure 3.5B). There is a subtle trend towards increased size captured in chimpanzee organoids at days 49ND and 56ND, however this finding is not significant (unpaired two-tailed *t* test) (Figure 3.5C). This finding may suggest that human cells meet a critical DUF1220 dosage threshold wherein excess has a more profound phenotypic impact that is not recapitulated in chimpanzee cells.

3.2.5 Depletion of DUF1220 diminishes proliferation and causes premature differentiation in human cortical organoids

To test the hypothesis that depletion of endogenously expressed *NBPF* in hNPCs (i.e., *NBPF14* and *NBPF10*) results in a decrease in NPC proliferation and/or premature differentiation, biology that is known to cause a reduction in cortical neuron number and lead to smaller brain size, we used CRISPR/Cas9 editing to introduce an early termination codon in hs*NBPF14* in H9 ESCs. Briefly, we used an sgRNA designed to target an early exon downstream of any predicted transcription start sites with 100% match to NBPF14/10 in conjunction with a ssODN retaining complete homology to the target locus except for 17 bp changes that served to 1) abolish the PAM site and 2) introduce unique primer sequence for screening (M13R) upon homologous recombination during double-strand break repair (Figures 3.6A and 3.6B). To screen colonies and identify edited clones, PCR and Sanger sequencing were used with an upstream *NBPF* forward primer and M13R (Figure 3.6C). Using the pan-NBPF-DUF1220 antibody, western blot analysis confirmed NBPF protein reduction in NPCs differentiated from three independently CRISPR-edited clonal ESC lines compared to unedited control NPCs (Figures 3.6D and 3.6E). Given the inability to distinguish between paralogs and isoforms of NBPF proteins with western analysis, further genotyping was performed using PacBio long-read HiFi sequencing, achieving high quality reads with ~5x genome-wide coverage (Figure S3.2A). We confirmed the M13R-disrupting allele in one of the three edited lines that may suggest heterozygous editing of *NBPF14*, which would be consistent with *NBPF14* genotypes of 1q21DDS (Figure S3.2B). In addition to confirmation of M13R allele, mapping of HiFi reads to hs*NBPF* sequence indicates several putative indels surrounding the target locus (Figure S3.2B) as well as at other *NBPF* loci (data not shown) that may have been generated during editing and contribute to the observed reductions in protein levels by western analysis. However, this may also reflect alignment issues using long-read HiFi reads for highly repetitive *NBPF* genes. Although sequencing and western blot were inconclusive, they nevertheless support depletion of NBPF-DUF1220 sequence on a protein level. Given this, we further explored phenotypes in these models to evaluate a reduced dosage effect on cortical growth.

The remaining analyses focus on organoid findings from the one edited line with HiFi reads confirming the presence of the M13R allele and will be referred to as H9^{NBPF-}

$^{DUF1220\Delta}$. H9$^{NBPF\text{-}NBPF\text{-}DUF1220\Delta}$ ESCs were differentiated to cerebral organoids and collected at day 32ND together with H9^{WT} controls. Immunofluorescent staining was performed on 12 organoids per genotype from two replicate differentiations (6 per replicate). Grossly, H9$^{NBPF\text{-}DUF1220\Delta}$ organoids have a thinner and less organized neuroepithelium with fewer cells per neural rosette (NR) compared to controls (Figures 3.6F and 3.6G). H9$^{NBPF\text{-}DUF1220\Delta}$ NRs showed a significant reduction in KI67-positive proliferating cells ($p = 0.0352$, two-tailed *t* test) and a significant increase in NEUN-positive immature neurons ($p = 0.0223$, two-tailed *t* test) suggesting a potential trend towards premature differentiation (Figure 3.6H and 3.6I). Together, depletion of NBPF-DUF1220 protein in human cerebral organoids impairs proliferation and may result in slight premature differentiation during early corticogenesis—biology consistent with pathogenic mechanisms of microcephaly. These findings support the hypothesis that NBPF proteins are important for establishing cortical neuron number via regulating NPC dynamics.

To further investigate cellular composition and subtle biological defects not detected by immunohistochemistry, scRNAseq was performed on four organoids at day 32ND per genotype. The total cell yield was lower in the control lines compared to H9$^{NBPF\text{-}DUF1220\Delta}$, and relatively low for both genotypes compared to other scRNAseq datasets. Thus, to fully complement the immunohistochemistry, additional replicates and higher cell numbers are required to evaluate cellular ratios important for dissecting proliferation versus differentiation in an unbiased way. Nevertheless, we were able to identify predicted cell clusters for early-stage organoids (Figure 3.7A). Differences in specific cell cluster proportions were observed. A larger proportion of H9$^{NBPF\text{-}DUF1220\Delta}$ organoid cells made up the deep-layer (DL) cortical neuron and immature neuron clusters compared to

H9$^{+/+}$ cells (Figure 3.7B). Conversely, a smaller proportion of H9$^{NBPF\text{-}DUF1220\Delta}$ organoid cells was observed for mitotic, cortical neuron, and interneuron clusters (Figure 3.7B). In terms of common markers, fewer H9$^{NBPF\text{-}DUF1220\Delta}$ progenitor cells expressed proliferation markers *PCNA* and *MKI67* than control cells whereas more H9$^{NBPF\text{-}DUF1220\Delta}$ cells expressed *DCX* in immature neuron cluster, findings consistent with immunostaining (Figure 3.7D). In the cortical DL neuron cluster, more H9$^{NBPF\text{-}DUF1220\Delta}$ cells expressed early born neuron marker, *TBR1*, yet fewer H9$^{NBPF\text{-}DUF1220\Delta}$ cells expressed layer 5 cortical neuron marker, *CTIP2*, in the general cortical neuron cluster. The latter finding is consistent with the subtle trend towards decreased CTIP2 per neural rosette observed from immunostaining (Figure 3.6J). These cortical neuron cluster differences could indicate premature differentiation and/or cortical lamination defects.

GO-terms enriched biological categories of the differential genes identified across cell clusters in H9$^{NBPF\text{-}DUF1220\Delta}$ organoids included: G2/M mitotic cell cycle, actin filament organization, neuron migration, positive regulation of cell size, negative regulation of microtubule polymerization, axon guidance, protein processing at the endoplasmic reticulum, microtubule binding, cognitive trait, Schizophrenia and Autism, and neuropsychiatric disorders. These categories suggest *NBPF* genes are important for brain development and proliferation dynamics, potentially through regulation of cytoskeleton components and cell cycle kinetics. These findings are consistent with subcellular localization patterns and enriched pathways/networks in the NBPF15-FLAG IP-MS results, together implicating NBPF proteins in proliferative biology during corticogenesis.

3.3 DISCUSSION

Phylogenetic copy number expansion of *NBPF-DUF1220* sequence has been proposed as a contributor to cortical size expansion in primates (Keeney et al. 2014). Humans, who have triple the size of the chimpanzee brain with double the number of neurons and the most profound prefrontal cortex expansion across apes, acquire double the copies of *NBPF-DUF1220* as chimpanzees from hsSDs that arose following the divergence of human and chimpanzee lineages (Kronenberg et al. 2018; Mora-Bermúdez et al. 2016; O'Bleness et al. 2014; O'Bleness et al. 2012; Smaers et al. 2017). Further, hs*NBPF* duplicate genes that contribute to this doubling of *NBPF-DUF1220* sequence show evidence of a remarkable enrichment in expression in the developing brain (Fiddes et al. 2019; Kronenberg et al. 2018; Pollen et al. 2019). While there are multiple contenders for cortical expansion that arose with the sequential SDs throughout ape evolution (Dougherty et al. 2017; Fiddes et al. 2018; Florio et al. 2015; Suzuki et al. 2018), we provide compelling evidence that *NBPF* genes are players in cortical expansion.

The functional fate of duplicate genes in the *NBPF* gene family in humans certainly involves gene structure divergence from ancestral copies in the form of copy number expansion as well as transcriptional divergence due to differential promoter acquisition. There are two general models of functionalization following gene duplication for genes that are maintained and escape pseudogenization that we will consider for *NBPF* duplicates. One is retention of ancestral function, often in the form of subfunctionalization whereby duplicates acquire changes that allow them to divvy up ancestral functions (Force et al. 1999; Hughes 1994; Piatigorsky and Wistow 1991). The other is neofunctionalization, where a derived function(s) arises from sequence divergence of either or both copies (Lynch and Katju 2004; Ohno 1970). Our findings support a model of *cis*-regulatory divergence during *NBPF* evolution in primates that resulted in enrichment of NBPF protein expression in the developing human brain, potentially reflecting subfunctionalization, coupled with sparse, low expression of other *NBPF* genes across cell types. Further, we provide evidence to suggest a NBPF dosage model with retained/redundant ancestral functions afforded by DUF1220 protein domains during neural development wherein increased dosage elevates NPC proliferative capacity and reduced dosage impairs proliferation.

Our findings of shared subcellular localizations of NBPF proteins in human and chimpanzee mitotic NPCs (albeit low signal in mitotic cNPCs), as well as between NBPF and DUF1220-carrying PDE4DIP proteins in human interphase and mitotic NPCs, suggest shared functions via DUF1220 domains in human corticogenesis. *PDE4DIP* expression is largely confined to heart and skeletal muscle where it localizes to both the Golgi apparatus and the centrosome to activate nucleation and help anchor gamma-tubulin nucleating ring complexes—processes crucial for directed cell migration and establishing mitotic spindle orientation (Roubin et al. 2013). While PDE4DIP has not been directly linked to microcephaly, its functions at affected biology in microcephaly and interacts with primary microcephaly gene, *CDK5RAP2*. From IP-MS, we identified putative evidence for NBPF interaction with LGALS1 which interacts with the PDE4DIP-interacting partner, LGALS3BP, that is in complex with PDE4DIP at the Golgi apparatus and centrosomes (Roubin et al. 2013), further suggesting shared functions of NBPF15 with PDE4DIP via DUF1220. In support of NBPF proteins functioning at cell biology important for NPC proliferation in cortical growth, we identified putative interactions of

NBPF15 with syndromic microcephaly genes, cytoskeletal proteins, and extracellular matrix proteins. Additionally, our comparative IP-MS analysis of NBPF15 in human and chimpanzee backgrounds indicates interaction with different proteins but at similar biology, supporting a model of shared NBPF functions in both species. However, there may be a critical threshold for which increased or decreased DUF1220 dosage has a substantial effect on cortical growth that is achieved in human cells, as evidenced by our interspecies NBPF15-overexpression cortical organoid analysis. A genomic trade-off of NBPF-DUF1220 dosage has been previously pondered wherein a minimal and maximum threshold in copy number has been proposed for unaffected corticogenesis, and dosage decrease or increase beyond this threshold causes neurodevelopmental/neuropsychiatric disorders (Sikela and Searles Quick 2018).

The lack of shared protein partners between human and chimpanzee IP-MS experiments may reflect structural plasticity of NBPF15 afforded by its intrinsically disordered regions (IDRs). NBPF proteins are characterized as intrinsically disordered proteins (IDP) (Van Bibber et al. 2020), which are predicted to have conformational flexibility typically in absence of interaction with specific protein partners (Wright and Dyson 2015). At least one molecular recognition feature (MoRF)—features that undergo disorder-to-order transitions upon protein binding—per DUF1220 region has been observed, suggesting that intrinsic-disorder-based interactions are important for NBPF protein function (Van Bibber et al. 2020). Proteins with IDRs and low complexity regions tend to undergo liquid-liquid phase separation (LLPS) (Banani et al. 2017; Ditlev et al. 2018; Lin et al. 2018), especially for proteins that have several copies of domains that interact with each other at low affinity (Banani et al. 2017; Ditlev et al. 2018; Lin et al.

2018). Indeed, from IP-MS, we found evidence of NBPF15 interaction with NBPF1 and NBPF9 in hNPCs. Consistent with this, previous evidence has shown that NBPF15 N-terminal region is able to form aggregates and undergo phase transition (Wu et al. 2020a). Considering our findings strongly implicate NBPF proteins at the mitotic spindle apparatus, formation of which is known to be driven in part by LLPS (Jiang et al. 2015; Tiwary and Zheng 2019), one avenue for future investigation is to test if NBPF-DUF1220 domains interact with each other to drive biomolecular condensation important for microtubule dynamics in NPC activity and polarity, and that an increase in NBPF-DUF1220 dosage corresponds to elevated capacity for LLPS.

Given the repetitive nature of young *NBPF* duplicates, multimapping reads cannot be ruled out in short read sequencing data which likely obscures the characterization of paralog-specific expression profiles as well as the contribution of isoform diversity in the presented data. This renders our transcriptomic data only suggestive and warrants follow-up investigation using long-read RNA sequencing approaches to dissect *NBPF* paralog expression across development. In addition, whole organoid area measurements from overexpression models do not tell us what is happening on a NR level, so further investigation of correlated NR phenotypes are necessary to assess proliferative mechanisms. It also remains to be tested if DUF1220 domain clades are functionally distinct in addition to their sequence diversity. If they are, this could afford human-specific triplet DUF1220-expanded NBPFs novel functional divergence in the developing brain that will be necessary to assess for implications in 1q21DDS pathogenicity and human cortical evolution. Lastly, while gene duplication is a prevalent mechanism across eukaryotes, estimated to have created >30% of genes (Zhang 2003), it often leads to pseudogenization (Ohno 1970). Thus, we acknowledge that most duplicate genes are expected to undergo non-functionalization with an estimated half-life of 4 million years, which is older than most hsSDs including those that created hs*NBPF* and hs*NOTCH2NL* duplicates (Dennis et al. 2017; Lynch and Conery 2000, 2003). It is likely that most recent *NBPF* genes are in a transitory phase either towards loss or towards fixation by natural selection. Nevertheless, this research provides evidence that NBPF proteins are important for primate evolutionary biology and contribute to extant human biology.

In conclusion, our study provides novel data into NBPF protein expression, localization, and function in human and chimpanzee models of neural development. Utilizing human and chimpanzee NPC and cortical organoid models, we confirm enriched NBPF expression profiles in the human brain on a protein level and provide novel characterization of NBPF cytoplasmic subcellular localizations consistent with Golgi apparatus, microtubules, mitotic spindle, and the midbody. Organoid dosage models provide evidence that NBPF-overexpression enhances cortical growth and NBPF-depletion reduces NPC proliferative capacity, implicating NBPF proteins in establishing cortical size. Our findings further implicate *NBPF* in the etiology of neurodevelopmental disorders.

3.4 METHODS

Human and chimpanzee ESC/iPSC culture

H9 human embryonic stem cells (female, WA09, WiCell) and all human and chimpanzee iPSC lines were cultured using feeder free conditions on Matrigel (Corning with mTeSR-1 (STEMCELL Technologies). Passaging was performed using mTeSR-1 supplemented

1 nM ROCK inhibitor (BD Biosciences 562822) to prevent differentiation. Both manual and chemical dissociation with Versene (Gibco 15040066) were performed for splitting. Karyotype and CNV microarray analyses were performed on H9 lines prior to CRISPR/Cas9 editing, and again before long read genomic and RNA sequencing, to ensure no pathogenic changes were acquired during culturing. iPSC lines H20682 (human female), H28834 (human female), H23555 (human male), C40280 (chimpanzee female), C3647 (chimpanzee female), and C3649 (chimpanzee female) were a gift from Yoav Gilad and are described in (Gallego Romero et al. 2015). No further validation of the iPSCs lines was performed in our lab.

Generation of overexpression cell lines

For NBPF15-mCherry expression, 3 μg phCMV1-NBPF15-mCherry (gifted to our lab from Aaron Issaian and Kirk Hansen and cloned from isolated NBPF15 cDNA) was co-electroporated into H9-derived hNPCs (H9$^{NBPF15\text{-}mCherry}$). To establish the doxycycline-inducible NBPF15 lines (*H9*$^{iNBPF15\text{-}FLAG}$ and C40280$^{iNBPF15\text{-}FLAG}$), NBPF15 sequence was cloned from phCMV1-NBPF15-mCherry into the piggybac plasmid pPBhCMV1-NBPF15-FLAG-pA, followed by co-electroporation of pPBhCMV1-NBPF15-FLAG-pA (2 μg) and pPBCAG-rtTAM2-IRES-NEO-pA (1 μg) into ~1 x 10^6 ESCs or iPSCs with the piggybac transposase helper plasmid pPBase (1 μg). In addition, ~1 x 10^6 ESCs or iPSCs were electroporated at the same time with pPBhCMV1-NBPF15-FLAG-pA (2 μg) and pPBase (1 μg) only as a control for cell survival during drug selection. NBPF15-3xFLAG inducible lines with an induction fluorescence marker (H20682$^{iNBPF15\text{-}3xFLAG}$, H28834$^{iNBPF15\text{-}3xFLAG}$

H23555$^{iNBPF15\text{-}3xFLAG}$, C40280$^{iNBPF15\text{-}3xFLAG}$, C3647$^{iNBPF15\text{-}3xFLAG}$, C3649$^{iNBPF15\text{-}3xFLAG}$) were created by modifying pPBhCMV1-NBPF15-FLAG-pA plasmid to pPBhCMV1-NBPF15-3XFLAG-pA-eGFP-P2a and electroporated as described above. CON1 and HLS1-HLS2-HLS3 domain sequence was isolated from NBPF15 cDNA and cloned into pPBhCMV1-pA-mCherry-P2a followed by electroporation, as indicated above, into human and chimpanzee NPCs to generate inducible lines H20682$^{CON1\text{-}3xFLAG}$, C40280$^{CON1\text{-}3xFLAG}$ H20682$^{HLS1\text{-}HLS2\text{-}HLS3\text{-}3xFLAG}$, C40280 $^{HLS1\text{-}HLS2\text{-}HLS3\text{-}3xFLAG}$.

All electroporations were performed with addition of pmaxGFP for visualization of electroporation efficiency (1 μl) using the Human Stem Cell Nucleofector Kit 1 (Lonza, VPH-5012) and Amaxa Nucleofector following manufacturer's instructions (Lonza, A-23). After 48 hours post-electroporation, cells with integrated pPBCAG-rtTAM2-IRES-NEO-pA were selected for by supplementing mTeSR-1 media with G418 (300 μg/ml, Corning 61234RF). After three days of selection, complete loss of cells is observed in the control well. After seven days of drug selection, cells are split using serial dilution for single colony screening for integration of pPBhCMV1-pA using PCR amplification primers targeting a unique region of the plasmid upstream of NBPF15. For NBPF15-FLAG and NBPF15-3xFLAG transgene induction, hNPCs and cNPCs were treated with 2 μg/mL Doxycycline (Thermo Scientific 446060050) for 48 hours for immunostaining and immunoprecipitation.

CRISPR/Cas9 editing of cell lines

To generate H9$^{NBPF\text{-}DUF1220\Delta}$ ESCs, CRISPR/Cas9 editing was performed. The sgRNA was designed to target *NBPF14,* with as few matches as possible to other *NBPF* paralogs, using the online tool provided by the Zhang lab (http://crispr.mit.edu). The selected sgRNA had 100% match to both *NBPF14* and *NBPF10*, with 3 or more mismatches to other *NBPF* loci. Cloning of the sgRNA was performed by first annealing and phosphorylating sense and antisense oligos of the sgRNA and then cloning the duplex into pSpCas9(BB) (PX330) (Addgene plasmid # 42230) via digestion with BbsI (NEB R3539S) and ligation with T4 ligase. The U6 forward primer was used to confirm plasmid presence in selected bacterial colonies. Selected cultures were expanded, and plasmids were isolated and purified by maxiprep (Qiagen A29739). Confluent H9 ESCs were electroporated as described above using 3 μg of pX330 plasmid together with 3 μg of a 200 bp single-stranded oligodeoxynucleotide (ssODN) to promote homologous recombination and designed with homology to the target locus except for 17 bp differences in the middle of the ssODN to introduce M13R sequence at the target site, serving to 1) ablate the PAM site, 2) introduce an early termination codon, and 3) provide unique primer sequencing for colony screening. Following electroporation, cells were serially diluted and passaged into 96-well plates and screened by PCR and Sanger sequencing using an NBPF14/10 forward primer and M13R. Serial dilution of positive colonies was performed until single colonies were obtained. Based on PCR screening, three H9-edited clonal lines were maintained for further genotyping and analysis. Primers, sgRNA, and ssODN sequences are provided in Table 3.

Human and chimpanzee cortical organoid and NPC differentiation

Telencephalic cerebral organoids were generated based on previously published protocols (Lancaster et al. 2013), with few modifications to start with low cell density in order to generate smaller and more consistent embryoid bodies (EBs). Briefly, human and chimpanzee iPSCs were passaged into 96-well V-shaped bottom ultra-low attachment cell culture plates (PrimeSurface® 3D culture, MS-9096VZ) to achieve a starting cell density of 600-1,000 cells per well in 30 µl of mTesR^{TM}1 with 1 nM ROCK inhibitor. After 36 hours, 150 µl of N-2/SMAD inhibition media (cocktail of 1X N-2 supplement (Invitrogen 17502048), 2 μM A-83-01 inhibitor (Tocris Bioscience 2939), and 1 mM dorsomorphin (Tocris Bioscience 309350) in DMEM-F12 (Gibco 11330032)) was added for neural induction. On day 7, EBs were transferred to Matrigel-coated plates to enrich for neural rosettes at a density of 20-30 EBs per well of a 6-well plate, and media was changed to neural differentiation media (0.5X N-2 supplement, 0.5X B-27 supplement (Invitrogen 17504044) with 20 pg/μl bFGF and 1mM dorsomorphin inhibitor in DMEM/F-12). For organoid differentiation EBs were outlined on day 14 using a pipet tip and uplifted carefully with a cell scraper to minimize organoid fusion and tissue ripping. Media was changed once more to N-2/B-27 with bFGF only and plates with uplifted organoids were placed on a shaker in the incubator set at a rotation speed of 90. On day 14, media was changed once more to N-2/B-27 with bFGF only. Prior to day 14, media changes were performed every 48 hours. After day 14, daily media changes were performed until collection. For monolayer NPC differentiation, neural rosettes were scored and uplifted on day 14, dissociated in Accutase (Gibco A1110501), and re-plated on poly-L-ornithine (PLO)/Laminin-coated plates for NPC expansion, selection, and passaging. 15 µg/mL PLO (Sigma-Aldrich P4957) diluted in DPBS (Gibco 14040-133); 10 µg/mL laminin (Sigma-Aldrich L2020) diluted in DMEM/F-12.

H9^{WT} and H9$^{NBPF\text{-}DUF1220\Delta}$ organoids were differentiated based on an older version of this protocol that followed the steps as described above except for differentiation began from a confluent dish of PSCs and tissue was manually scored and uplifted for EB formation by shaking at day 3.

HMW genomic DNA isolation

Phenol-chloroform extractions were performed on the screened *NBPF14*-targeted H9 ESC lines to obtain high molecular weight (HMW), high quality gDNA for long read sequencing. Briefly, cells were transferred to 1.5 mL Eppendorf tubes and pelleted at 3400 rpm. Media was removed and replaced with ~550 μl lysis buffer per ~1 x 10^6 cells. Cells in lysis buffer are incubated for digestion overnight at 55°C. 2x volumes of phenol is added to remove protein, mixed by vertexing, and spun at 3000 rpm for 5 minutes. Upper aqueous layer is transferred to a fresh tube and 1x volume of chloroform is added, mixed, and spun. Supernatant is transferred for DNA precipitation by addition of 0.1 volume of 3 M sodium acetate and 2.5 volumes of cold 100% ethanol (-20°C). Sample is mixed by inversion and incubated in the -80°C for 1 hr. Sample is spun at 10,000 rpm and supernatant is removed. Precipitate is washed with cold 70% ethanol (4°C) and then air dried for 10 minutes at room temperature. gDNA pellet is resuspended in low TE for 48 hours at 4°C without mixing. Concentration was assessed using Nanodrop and Qubit.

PacBio HiFi sequencing and analysis

For long-read genotyping, 70-100 μg of HMW gDNA per sample was shipped on dry ice to the University of Oregon Genomics and Cell Characterization Core Facility (GC3F) for QC, single-plex genome assembly library preparation, and sequencing. For HiFi read preparation, gDNA was sheared using BluePippin selection at 10-20kb fragments and libraries were sequenced on 2-3 SMRT cells per sample. Average read lengths from sequencing were 10-20 kb with average genome coverage of ~5x. NanoPlot (De Coster et al. 2018) was used to visualize read lengths by average read quality using kernel density estimates and histogram plots in Python3 (v3.10) (Van Rossum and Drake Jr 1995). HiFi reads were then aligned to the hg38 reference genome using Winnowmap (Jain et al. 2022; Jain et al. 2020) and Minimap2 (Li 2018) as described and indexed bam files were loaded into Integrated Genome Viewer (IGV) for alignment visualization (igv.org).

Organoid dissociation and single-cell RNA sample preparation

Organoid tissue was dissociated for single-cell RNA preparation as follows. First, 2-4 medium-sized organoids were transferred to a glass slide and cut with a sterile razor into several, tiny pieces. Pieces were then transferred to a fresh tube containing acidified EBSS using an unfiltered and cut pipette tip and suspended. EBSS was then aspirated and a 1:10 ratio of DNase solution (17.5 µL Papain, 0.5 µL 1 M EDTA, 11 µL 0.5 M Cysteine HCL and 971 µL acidified EBSS) to papain solution (0.8 µg DNAse in 800µL of acidified EBSS) was added to the tissue, mixed by flicking, and incubated in 37˚C bead bath for 8 minutes. Following incubation, tissue was triturated carefully to avoid bubbles using a glass pipette tip (~13 times) and then centrifuged for 5 minutes at 2,000 rpm. Supernatant was removed then pellet was resuspended in 1:9 DNase solution with albumin-ovomucoid inhibitor (AOI) to acidified EBSS and filtered through a 70 µm cell strainer. Fresh AOI was added to create a continuous density gradient and centrifuged at

1000 rpm for 5 minutes. Supernatant was removed, pellet was resuspended in N-2/B-27 + bFGF media by flicking, and cell number and viability were assessed.

Seq-Well single-cell RNA preparation was performed as described in (Gierahn et al. 2017). Briefly, 90,000-picowell Seq-Well arrays were functionalized and loaded with barcoded beads (ChemeGenes). ~20,000 dissociated cells were loaded onto the arrays and incubated for 15 minutes. PBS washes were performed to remove residual BSA and excess cells. Functionalized membranes were sealed in an Agilent clamp to the top of the arrays and incubated at 37° for 45 minutes. Sealed arrays were incubated in lysis buffer (5 M guanidine thiocynate, 1 mM EDTA, 0.5% sarkosyl, 1% BME) for 20 minutes followed by 45-minute incubation in hybridization buffer (2 M NaCl, 1X PBS, 8% PEG8000). Beads were removed from arrays by centrifugation at 2,000 rcf for 5 minutes in wash buffer (2 M NaCl, 3 mM $MgCl_2$, 20 mM Tris-HCl pH 8.0, 8% PEG8000). For reverse transcription, beads were incubated with the Maxima Reverse Transcriptase (Thermo Scientific EP0742) for 30 minutes at room temperature then overnight incubation at 52°C, followed by Exonuclease 1 treatment (New England Biolabs) for 45 minutes at 37°C. Second strand synthesis was then performed with Maxima Reverse Transcriptase (Thermo Scientific EP0742) for 1 hour at 37°C as described (Hughes et al. 2020). Whole transcriptome amplification was performed using the 2x KAPA Hifi Hotstart Readymix (KAPA Biosystems). Beads were split to 1,500-2,000 per reaction and run under the following conditions: 4 cycles (98°C, 20s; 65°C, 45s; 72°C, 3m), 12 Cycles (98°C, 20s; 67°C, 20s; 72°C, 3m), final extension (72°C, 3m, 4°C, hold). Products were purified with Ampure SPRI beads (Beckman Coulter CNGS005) at 0.6X volumetric ratio then 1.0X volumetric ratio. Libraries were prepared using the Illumina Nextera XT kit and sequenced on Illumina NextSeq (75-cycle).

Single-cell RNA sequencing and bioinformatics analysis

Sequencing reads were processed using a Drop-seq pipeline as described (Macosko et al. 2015). First, FASTQ files were converted to bam files, tagged with cell and molecular barcodes, and trimmed. Reads were mapped to hg38 using STAR alignment. Bam files were subsequently sorted, merged, and tagged with gene exons. Gene expression count matrices were generated. Cells with <300 detectable genes, >5000 genes, and/or >10% mitochondrial genes were removed from downstream analysis. Genes detected in less than 5 cells were excluded. Alter quality control filtering, a total of 2,890 cells were analyzed from H9$^{NBPF\text{-}DUF1220\Delta}$ (4 organoid replicates) whereas only 970 total cells were sequenced from H9 controls (4 organoid replicates).

We used the Seurat package (v4.0.6) in R for scRNAseq analysis and visualizations (Hao et al. 2021). We followed the Fast integration using reciprocal PCA (RPCA) vignette for integrative genotype analysis, which uses an optimized algorithm to identify anchors for any two datasets. This allowed us to identify cell clusters combining genotypes and visualize using UMAP dimensionality reduction plots. Cluster annotation was performed using the top defining genes in each cluster using Wilcoxon rank-sum test followed by cross-referencing with published scRNAseq from primary cortex and cortical organoid data available via the UCSC Cell Browser and scAPEx. Differential expression testing between genotypes for each cluster was performed with the FindMarkers() function with logfc.threshold=log(2), which perfoms a Wilcoxon rank-sum test. Genes were considered differentially expressed if adjusted p-value was <0.05. DAVID analysis (david.ncifcrf.gov/tools) and Metascape analysis (metascape.org) (Zhou et al., 2019) were performed to identify enriched biological pathways based on Benjamini-Hochberg multiple hypothesis corrections of the *p*-values.

Immunofluorescence staining and image quantifications

Human and chimpanzee NPCs were fixed in 4% paraformaldehyde (PFA) for 20 minutes. Human and chimpanzee cortical organoids were fixed in 4% PFA for 24 hours at 4°C, cryoprotected in 15% and 30% sucrose in 1x DPBS for 24 hours at 4°C, then embedded in OCT with quick freezing in -50°C 2-methylbutane, followed by cryosectioning for immunostaining. Primary antibodies used: rabbit anti-NBPF (1:250, Invitrogen PA5-83644), mouse anti-SOX2 (1:250, R&D Systems MAB2018), rabbit anti-KI67 (1:200, Abcam ab16667), rabbit anti-cleaved caspase3 (1:400, Cell Signaling 9661), goat anti-DCX (1:400, Santa Cruz Biotechnology A1313), rat anti-CTIP2 (1:500, Abcam ab18465), mouse anti-NEUN (1:250, EMD Millipore MAB377); mouse anti-FLAG (10 μg/ml, Sigma-Aldrich F3165). Samples for immunostaining were incubated for 1 hour with blocking buffer (5% NDS (Jackson ImmunoResearch) 0.1% Triton X-100, 5% BSA) at room temperature, then overnight with primary antibodies diluted in blocking buffer at 4°C, and for 2 hours in secondary dilution (1:400) at room temperature. Washes performed in PBS. For nuclear staining, samples were incubated at room temperature for 10 minutes in Hoescht (1:1000 dilution in PBS) prior to final washes.

Glass covers were mounted onto all slides with Prolong Gold (Molecular Probes S36972) and incubated for 24 hours at room temperature prior to imaging. Imaging was performed with a Nikon A1ss inverted confocal microscope using NIS-Elements Advanced Research software. Image analysis was performed using Fiji (ImageJ) software. Statistical significance of image quantifications was tested using a student's two-tailed *t*-test, and data was plotted as mean ±SEM using GraphPad Prism (v9.3.1).

Western blot analysis

ESCs, iPSCs, and NPCs used for western blot analysis were pelleted and lysed in RIPA buffer supplemented with 1:50 protease inhibitor cocktail (Sigma-Aldrich P8340) and 1:100 phosphatase inhibitor cocktail 3 (Sigma-Aldrich P0044) using mortar and pestle coupled with end-over-end rotation for 30 minutes to 1 hr at 4°C. Protein concentration was quantified by BCA (Thermo Scientific Pierce A53227). Lysis samples were then incubated at a 1:3 ratio with 4x Laemali sample buffer (Bio-Rad) supplemented with 10% BME and incubated at 95°C on a heat block for 5 minutes for denaturation. Samples were loaded into 4-20% SDS-polyacrylamide gels (Bio-Rad) and proteins were separated by electrophoresis at 30V for ~4 hours room temperature. Separated proteins were then transferred to PVDF membranes (Millipore) overnight using a wet transfer system (Bio-Rad) at 4°C. For immunoblotting, membranes were incubated in 5% milk blocking buffer (1x TBS-T) followed by primary antibody incubation overnight at 4°C with rotation. Membranes were washed 3 times for 5 minutes in 1x TBS-T and then incubated with secondary antibodies for 1-2 hours at room temperature. Membranes underwent final washes before developing using West Femto Substrate (ThermoFisher 34095) with film exposure.

Primary antibodies used: anti-NBPF (1:1000, Invitrogen PA5-83644), anti-FLAG (10 μg/ml), and anti-β-actin (1:2500, Abcam ab6276). Secondary antibodies used: donkey anti-rabbit HRP-conjugated (1:5000, Cytiva NA9340V) and goat anti-mouse HRP-conjugated (1:1000, Invitrogen 32430).

Immunoprecipitation and mass spectrometry

FLAG-immunoprecipitations were performed using anti-FLAG M2 Magnetic Beads (Sigma-Aldrich M8823) according to manufacturer's instructions. Briefly, whole cell lysates were prepared as described above. 40 μl M2 beads were used for ~5 x 10^6 cells and washed with TBS as recommended. 1 mL lysate was added to M2 beads and agitated overnight at 4C to increase binding efficiency. Supernatant was removed from beads using a magnetic separator. Resin was washed gently three times with TBS, with the suspension being split into two fresh tubes during the third wash: one with 80% volume and the other with the remaining 20%. The 20% bound fraction was used for protein elution in 4x Laemali sample buffer (Bio-Rad) with heat activation for analysis by western blot as described above to confirm sufficient NBPF15-FLAG pull-down (Figure S3.1D). The 80% fraction, containing ~3-5 mg protein, was submitted to the UM Proteomics Core for Liquid Chromatography with tandem mass-spectrometry (LC-MS/MS). For H9 hNPCs, two control and two NBPF15-FLAG-bait replicates were submitted. For C40280 cNPCs, three controls and three NBPF15-FLAG-bait replicates were submitted for TMT Mass Tagging 6-plex (Thermo Scientific, 90061).

LC-MS/MS data from hNPCs was searched against the Human Uniprot protein database (20,286 entries) whereas cNPC data was searched against *Pan troglodytes*

(Chimpanzee) Uniprot protein database (48,769 entries) appended with human *NPPF* proteins (18 entries). Both datasets were filtered for high confidence proteins based on number of peptides identified. Student's two-tailed *t*-test was used to generate *p*-values of abundance differences. The ggplot2 package () in R (v4.1.2) was used to create volcano plots. Top hits with $\log_2$(Abundance ratio) >1 and *p*-value <0.001 were filtered through the Contaminant Repository for Affinity Purification (CRAPome, reprint-apms.org) to remove commonly identified proteins. Data provided in Table 4 for hNPCs and Table 5 for cNPCs. Pathway enrichment analysis was performed using the DAVID database (david.ncifcrf.gov/tools) and Metascape (metascape.org) (Zhou et al., 2019) with Benjamini-Hochberg multiple hypothesis corrections of the *p*-values*.* To identify enriched protein networks, a STRING Network (v11.5, https://string-db.org) was generated after filtering for hits with less than 10% representation in the CRAPome database.

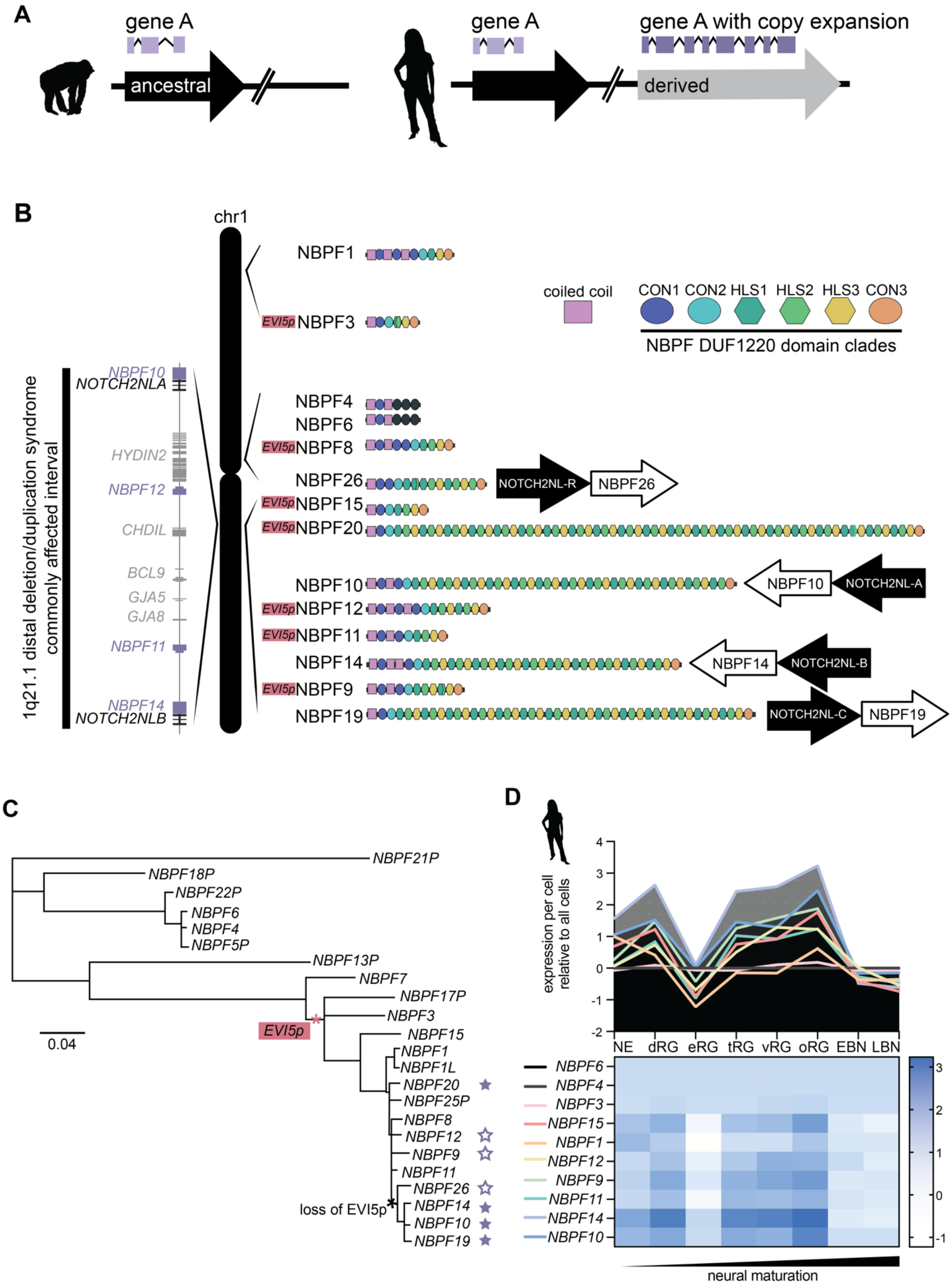

A
gene A
ancestral
gene A
gene A with copy expansion
derived
B
chr1
1q21.1 distal deletion/duplication syndrome
commonly affected interval
NBPF10
NOTCH2NLA
HYDIN2
NBPF12
CHDIL
BCL9
GJA5
GJA8
NBPF11
NBPF14
NOTCH2NLB
NBPF1
EVI5p NBPF3
coiled coil
CON1 CON2 HLS1 HLS2 HLS3 CON3
NBPF DUF1220 domain clades
NBPF4
NBPF6
EVI5p NBPF8
NBPF26
NOTCH2NL-R
NBPF26
EVI5p NBPF15
EVI5p NBPF20
NBPF10
NBPF10
NOTCH2NL-A
EVI5p NBPF12
EVI5p NBPF11
NBPF14
NBPF14
NOTCH2NL-B
EVI5p NBPF9
NBPF19
NOTCH2NL-C
NBPF19
C
NBPF21P
NBPF18P
NBPF22P
NBPF6
NBPF4
NBPF5P
NBPF13P
NBPF7
NBPF17P
NBPF3
EVI5p
0.04
NBPF15
NBPF1
NBPF1L
NBPF20 ★
NBPF25P
NBPF8
NBPF12 ☆
NBPF9 ☆
NBPF11
NBPF26 ☆
loss of EVI5p
NBPF14 ★
NBPF10 ★
NBPF19 ★
D
expression per cell
relative to all cells
4
3
2
1
0
-1
-2
NE dRG eRG tRG vRG oRG EBN LBN
NBPF6
NBPF4
NBPF3
NBPF15
NBPF1
NBPF12
NBPF9
NBPF11
NBPF14
NBPF10
3
2
1
0
-1
neural maturation

Figure 3.1. *Cis*-regulatory evolution of *NBPF* duplicate genes.
(A) Illustration of human-specific gene duplication by segmental duplication followed by sequence expansion. (B) Distribution of *NBPF* protein-coding genes in the human genome on chromosome 1, showing enrichment in 1q21 region. The six DUF1220 domain clades are represented by different colors. Proteins with *pEVI5* are labeled. Duplicated hs*NBPF*-hs*NOTCH2NL* gene pairs are indicated with directionality. Far left shows the most affected 1q21DDS interval. This illustration is re-created from Fiddes et al., 2019. (C) Phylogenetic NBPF tree depicting relatedness of hs*NBPF* paralogs. Re-created with same sequence data and software parameters as Fiddes et al., 2019. *NBPF* 5' sequences (1 kb sequence flanking the predicted start codon) were aligned using the Geneious Tree Builder function in Geneious Prime software (v2022.2) with a genetic distance model of Tamura-Nei (Tamura and Nei 1993) and the neighbor-joining build method. (D) NBPF paralog-specific expression levels across cortical cell types represented with line graph (top) and heat map (bottom). Expression data was obtained from Kowakoski et al., 2017. NE, neuroepithelium. dRG, diving radial glia. eRG, early radial glia. tRG, truncated radial glia. vRG, ventricular radial glia. oRG, outer radial glia. EBN, early-born neurons. LBN, late-born neurons.

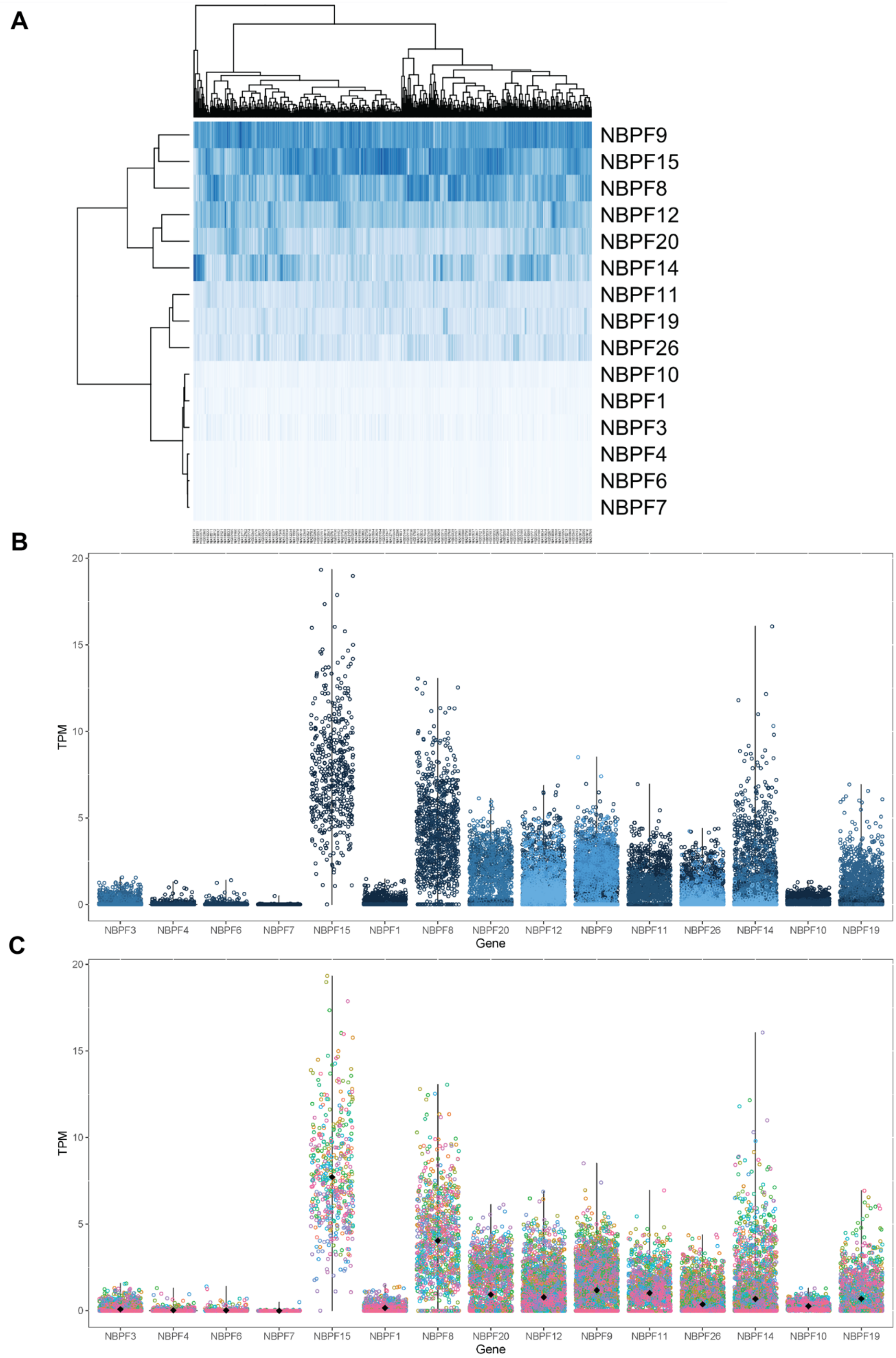

A
NBPF9
NBPF15
NBPF8
NBPF12
NBPF20
NBPF14
NBPF11
NBPF19
NBPF26
NBPF10
NBPF1
NBPF3
NBPF4
NBPF6
NBPF7
B
TPM
20
15
10
5
0
NBPF3
NBPF4
NBPF6
NBPF7
NBPF15
NBPF1
NBPF8
NBPF20
NBPF12
NBPF9
NBPF11
NBPF26
NBPF14
NBPF10
NBPF19
Gene
C
TPM
20
15
10
5
0
NBPF3
NBPF4
NBPF6
NBPF7
NBPF15
NBPF1
NBPF8
NBPF20
NBPF12
NBPF9
NBPF11
NBPF26
NBPF14
NBPF10
NBPF19
Gene

Figure 3.2. LCL RNAseq data from 463 individuals.
Data obtained from (Shew et al. 2021)**.** (A) Heatmap of expression level of *NBPF* genes (y-axis) per individual (x-axis) with phylogenetic trees based on expression values. Colors indicate increasing values from white to dark blue. (B) Jitter plot of TPMs per *NBPF* gene, colored by *NBPF* isoform. (C) Jitter plot of TPMs per *NBPF* gene colored by individual. Heatmap and jitter plots made in R (v4.1.2).

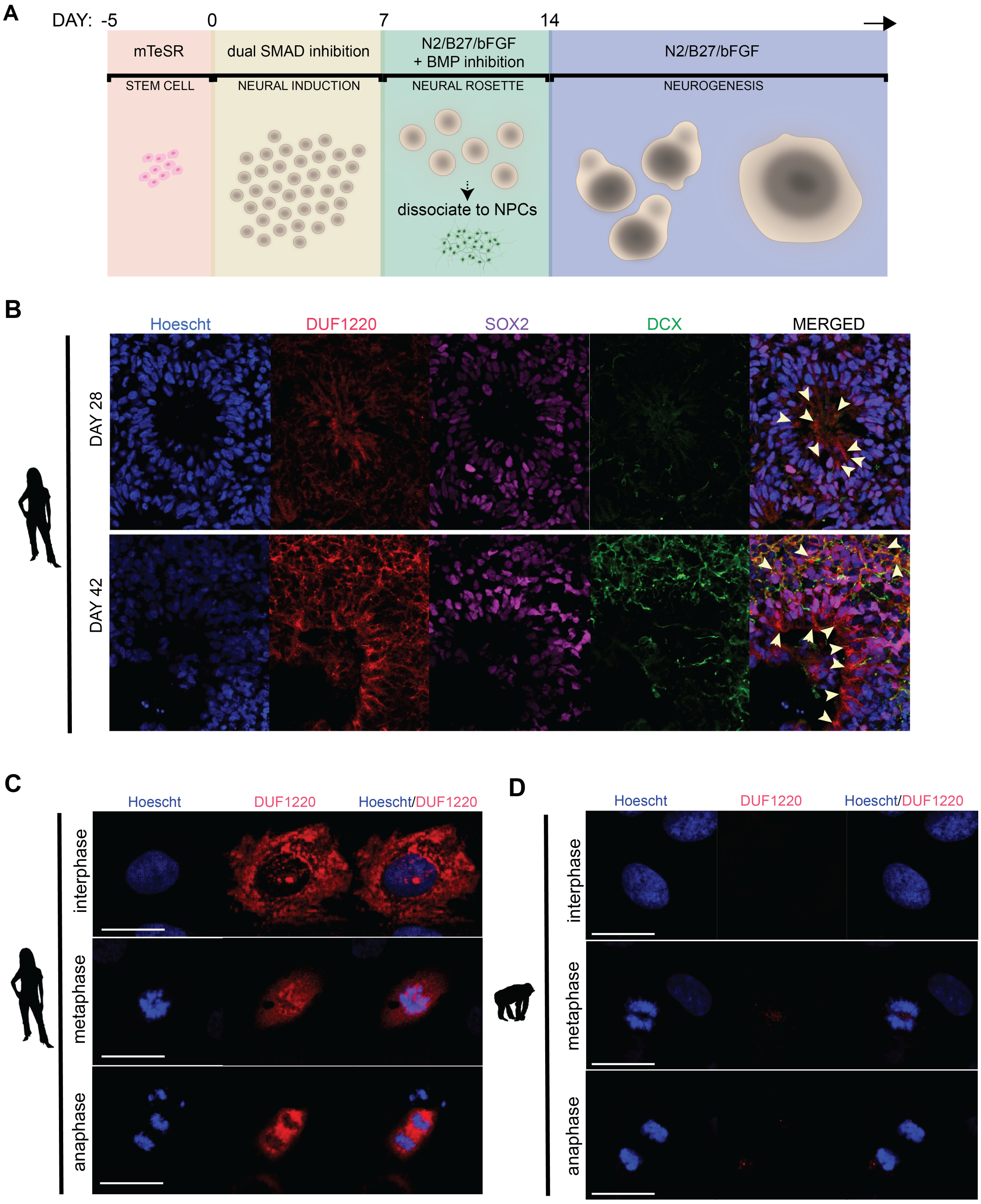

A
DAY: -5
0
7
14
mTeSR
dual SMAD inhibition
N2/B27/bFGF
+ BMP inhibition
N2/B27/bFGF
STEM CELL
NEURAL INDUCTION
NEURAL ROSETTE
NEUROGENESIS
dissociate to NPCs
B
Hoescht
DUF1220
SOX2
DCX
MERGED
DAY 28
DAY 42
C
Hoescht
DUF1220
Hoescht/DUF1220
interphase
metaphase
anaphase
D
Hoescht
DUF1220
Hoescht/DUF1220
interphase
metaphase
anaphase

Figure 3.3. NBPF protein expression and localization in developing cortical organoids.
(A) Schematic of differentiation timeline for human and chimpanzee cortical organoids and NPC cultures. (B) Immunofluorescence staining of human cortical organoids (H20682^{WT}) with markers for radial glia progenitors (SOX2), immature neurons (DCX), and NBPF (DUF1220). Scale bar, 10 μm. Subcellular localization by immunofluorescence staining of NBPF-DUF1220 protein in cycling NPCs derived from (C) human (H20682^{WT}) and (D) chimpanzee (C3647^{WT}). Scale bar, 25 μm.

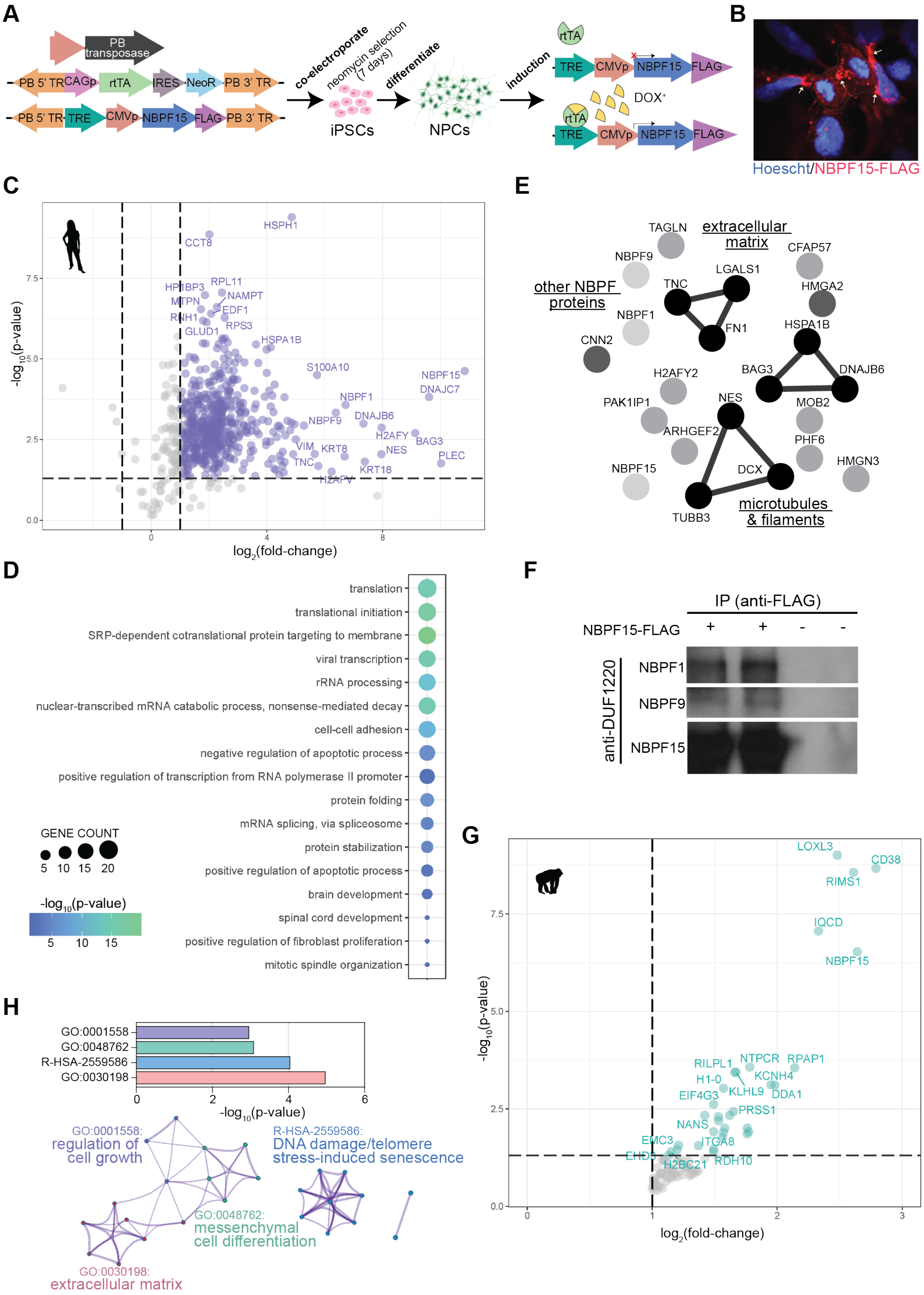

A
PB transposase
PB 5' TR
CAGp
rtTA
IRES
NeoR
PB 3' TR
PB 5' TR
TRE
CMVp
NBPF15
FLAG
PB 3' TR
co-electroporate
neomycin selection (7 days)
iPSCs
differentiate
NPCs
induction
rtTA
TRE
CMVp
NBPF15
FLAG
DOX+
rtTA
TRE
CMVp
NBPF15
FLAG
B
Hoescht/NBPF15-FLAG
C
-log10(p-value)
log2(fold-change)
0.0
2.5
5.0
7.5
0
4
8
HSPH1
CCT8
HP1BP3
RPL11
MTPN
NAMPT
EDF1
RNH1
RPS3
GLUD1
HSPA1B
S100A10
NBPF15
DNAJC7
NBPF1
NBPF9
DNAJB6
H2AFY
BAG3
VIM
KRT8
NES
PLEC
TNC
KRT18
H2AFV
D
translation
translational initiation
SRP-dependent cotranslational protein targeting to membrane
viral transcription
rRNA processing
nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
cell-cell adhesion
negative regulation of apoptotic process
positive regulation of transcription from RNA polymerase II promoter
protein folding
mRNA splicing, via spliceosome
protein stabilization
positive regulation of apoptotic process
brain development
spinal cord development
positive regulation of fibroblast proliferation
mitotic spindle organization
GENE COUNT
5 10 15 20
-log10(p-value)
5 10 15
E
TAGLN
extracellular matrix
NBPF9
CFAP57
other NBPF proteins
TNC
LGALS1
HMGA2
NBPF1
FN1
HSPA1B
CNN2
H2AFY2
BAG3
DNAJB6
PAK1IP1
NES
MOB2
ARHGEF2
PHF6
NBPF15
DCX
HMGN3
TUBB3
microtubules & filaments
F
IP (anti-FLAG)
NBPF15-FLAG
+
+
-
-
anti-DUF1220
NBPF1
NBPF9
NBPF15
G
-log10(p-value)
log2(fold-change)
0.0
2.5
5.0
7.5
0
1
2
3
LOXL3
CD38
RIMS1
IQCD
NBPF15
RILPL1
NTPCR
RPAP1
H1-0
KCNH4
KLHL9
DDA1
EIF4G3
PRSS1
NANS
EMC3
ITGA8
EHD3
H2BC21
RDH10
H
GO:0001558
GO:0048762
R-HSA-2559586
GO:0030198
0
2
4
6
-log10(p-value)
GO:0001558: regulation of cell growth
R-HSA-2559586: DNA damage/telomere stress-induced senescence
GO:0048762: messenchymal cell differentiation
GO:0030198: extracellular matrix

Figure 3.4. NBPF15-FLAG IP-MS in human versus chimpanzee NPCs.
(A) Schematic of piggybac three-vector system to generate hNBPF15-FLAG, hNBPF15-3xFLAG, hNBPF15-CON1-3xFLAG, and hNBPF15-HLS1-HLS2-HLS3-3xFLAG inducible lines followed by transgene induction in NPCs. (B) Immunostaining of Doxycycline-treated H9$^{iNBPF15\text{-}FLAG}$ hNPCs confirming transgene expression and subcellular localization. (C) Volcano plot of enriched proteins in NBPF15-FLAG bait hNPCs (H9$^{iNBPF15\text{-}FLAG}$) compared to controls (H9^{WT}), highlighted in purple. Two replicates per genotype. (D) Dotplot of enriched GO terms for top candidate NBPF15-FLAG IP-MS hits in hNPCs. (E) Graphic of STRING Network analysis on Crapome-filtered top IP-MS hits. Grayscale represents increasing evidence for functional link. (F) FLAG co-immunoprecipitation and immunoblotting with anti-NBPF-DUF1220 in bait and control hNPC replicates confirms interaction of NBPF15 (~70 kD) with other NBPF proteins (~130 kD and ~100 kD). (G) Volcano plot of enriched proteins in NBPF15-FLAG bait cNPCs (C40280$^{iNBPF15\text{-}FLAG}$) compared to controls (C40280^{WT}), highlighted in green. Three replicates per genotype. (H) Metascape network visualization of top hits within significantly enriched biological categories and corresponding *p*-values.

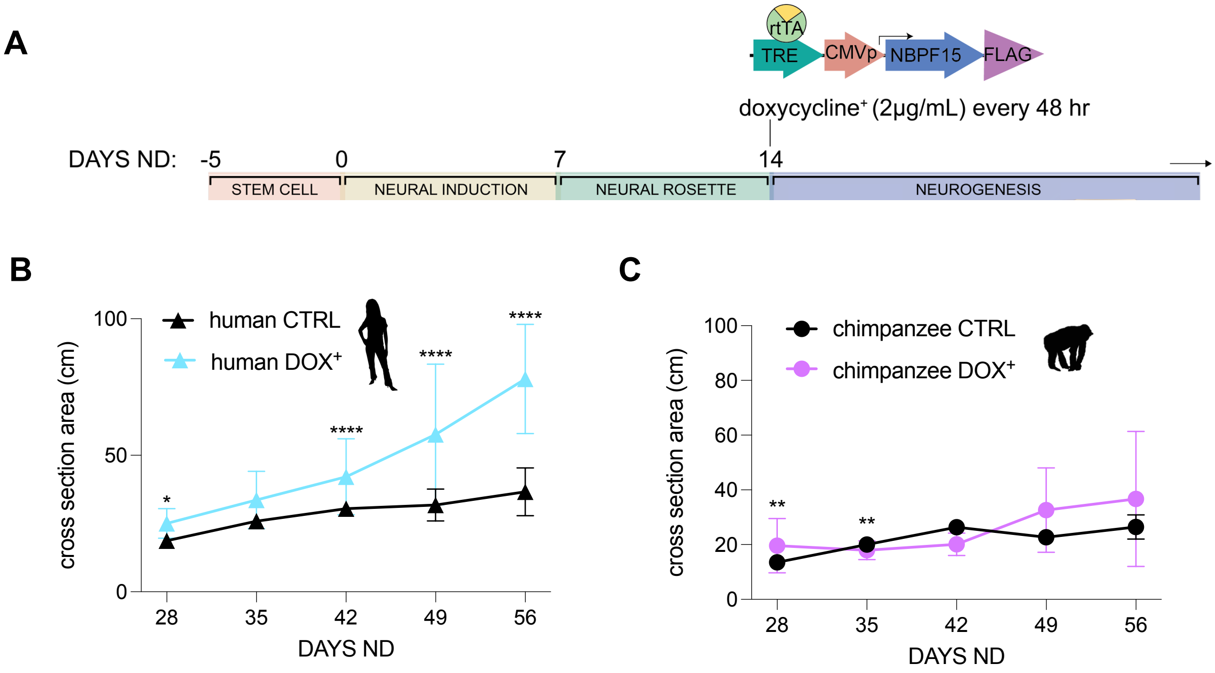
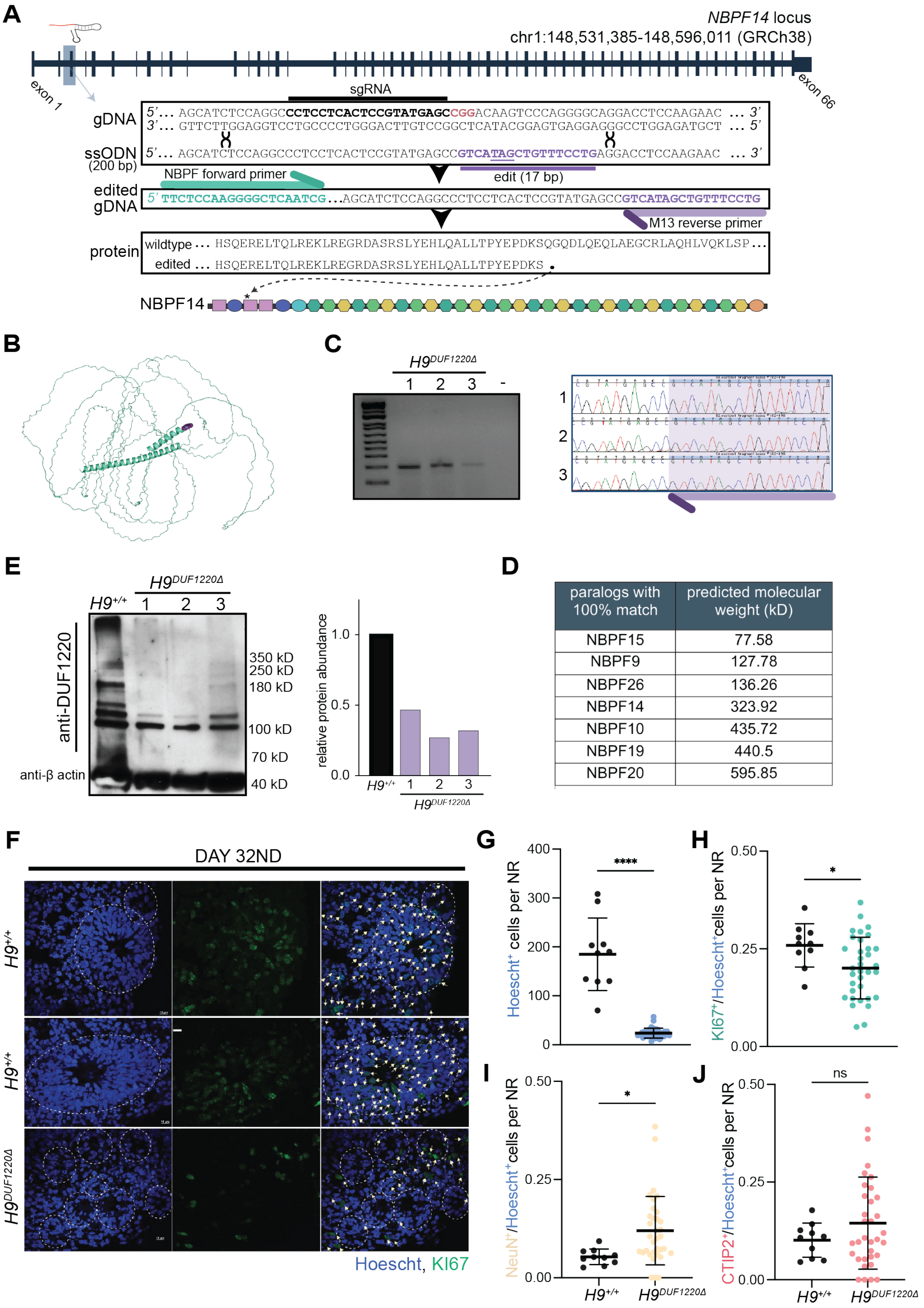


Figure 3.5. NBPF15 overexpression in human and chimpanzee cerebral organoids.
(A) Neural differentiation timeline of doxycycline induction for NBPF15 overexpression in organoids. Weekly cross section area measurements of organoids at day 28ND to day 56ND for human (B) chimpanzee (C). Data shown as mean±SEM. Measurements collected from organoids differentiated across two independent differentiations per condition, and from iPSCs derived from three individuals, per species.



paralogs with 100% match	predicted molecular weight (kD)
NBPF15	77.58
NBPF9	127.78
NBPF26	136.26
NBPF14	323.92
NBPF10	435.72
NBPF19	440.5
NBPF20	595.85

Figure 3.6. DUF1220-depleted cortical organoids exhibit proliferation defects.
(A) Schematic of CRISPR/Cas9 editing strategy to introduce M13R truncating sequence by homologous recombination at NBPF14 target locus. (B) NBPF14 predicted 3D protein structure based on AlphaFold predictions. NBPF14 (Q5TI25) PDB structure file was downloaded from Uniprot and edited in PyMOL (v2.5.2) to include predicted edit location. Coiled coil domain is towards center and DUF1220 domains are peripheral and less structured. Edit site is colored in purple. (C) PCR and Sanger validation of edited clonal H9 ESC lines. (D) Predicted molecular weights for each NBPF paralog predicted to be recognized by the anti-NBPF-DUF1220 antibody based on the canonical isoform in UCSC genome browser. (E) Western blot analysis on NPC whole cell lysates from H9 edited lines and controls. (F) Immunofluorescence staining of neural rosettes (circled) in day 32 cortical organoids with marker of proliferation (KI67) for wildtype H9 controls (top), unedited H9 electroporation-controls (middle), and H9$^{NBPF\text{-}DUF1220\Delta}$. Scale bar, 10 μm. Marker quantifications per neural rosette: (G) Hoescht+ nuclei (*p*=<0.001), (H) proliferating KI67+ cells/Hoescht+ cells (*p*=0.0352), (I) NEUN+ immature neurons/Hoescht+ cells (*p*=0.0223), (J) layer V CTIP2+ cortical neurons/Hoescht+ cells (*p*=0.2605). Significance determined by two-tailed unpaired *t* test. Two differentiation replicates of 6 organoids per replicate were analyzed per genotype (H9$^{+/+}$, H9$^{NBPF\text{-}DUF1220\Delta}$). Neural rosette count: H9$^{NBPF\text{-}DUF1220\Delta}$, *n* = 34; H9$^{+/+}$, *n* = 10. Cryosection thickness, 13 μm.

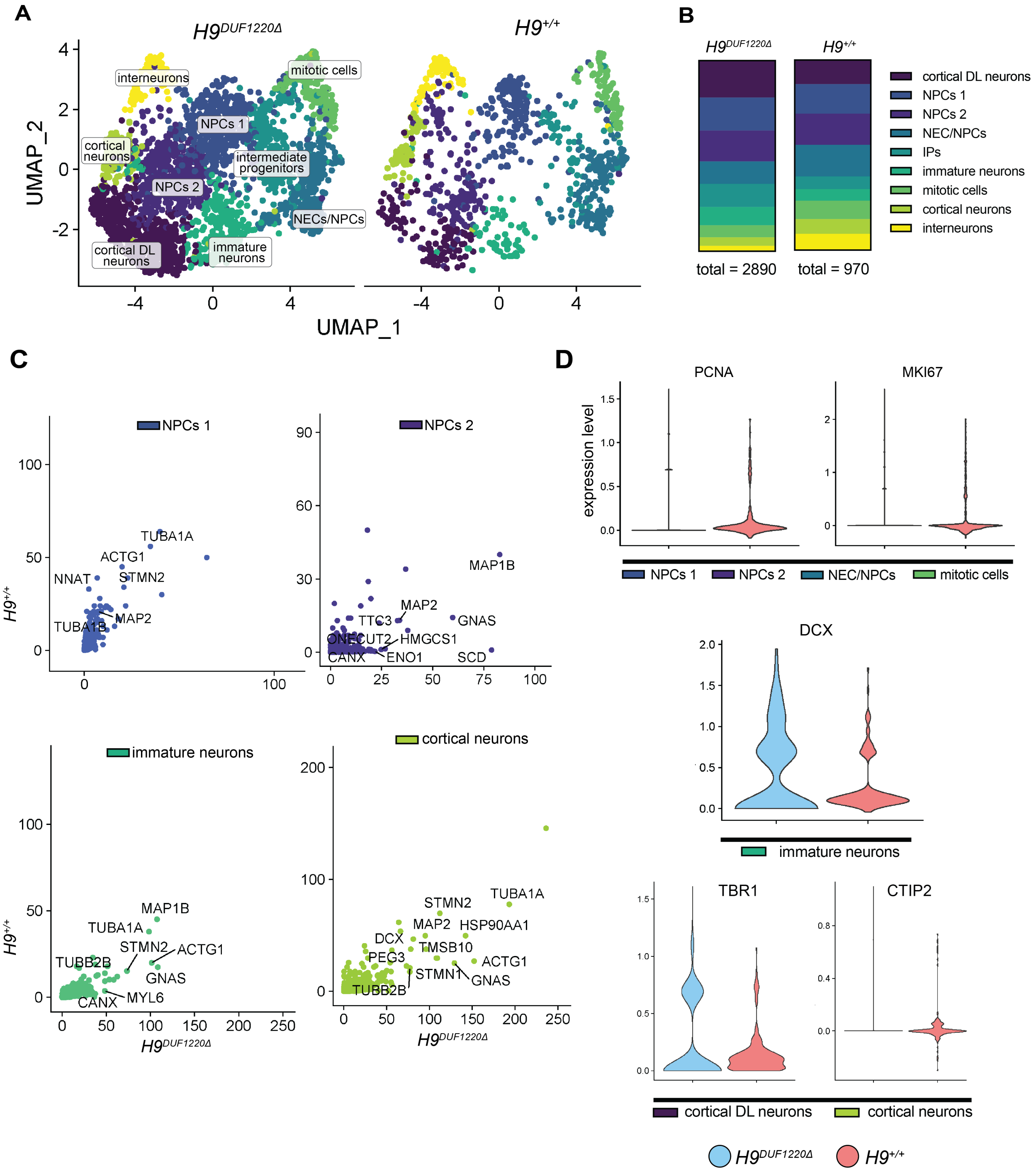
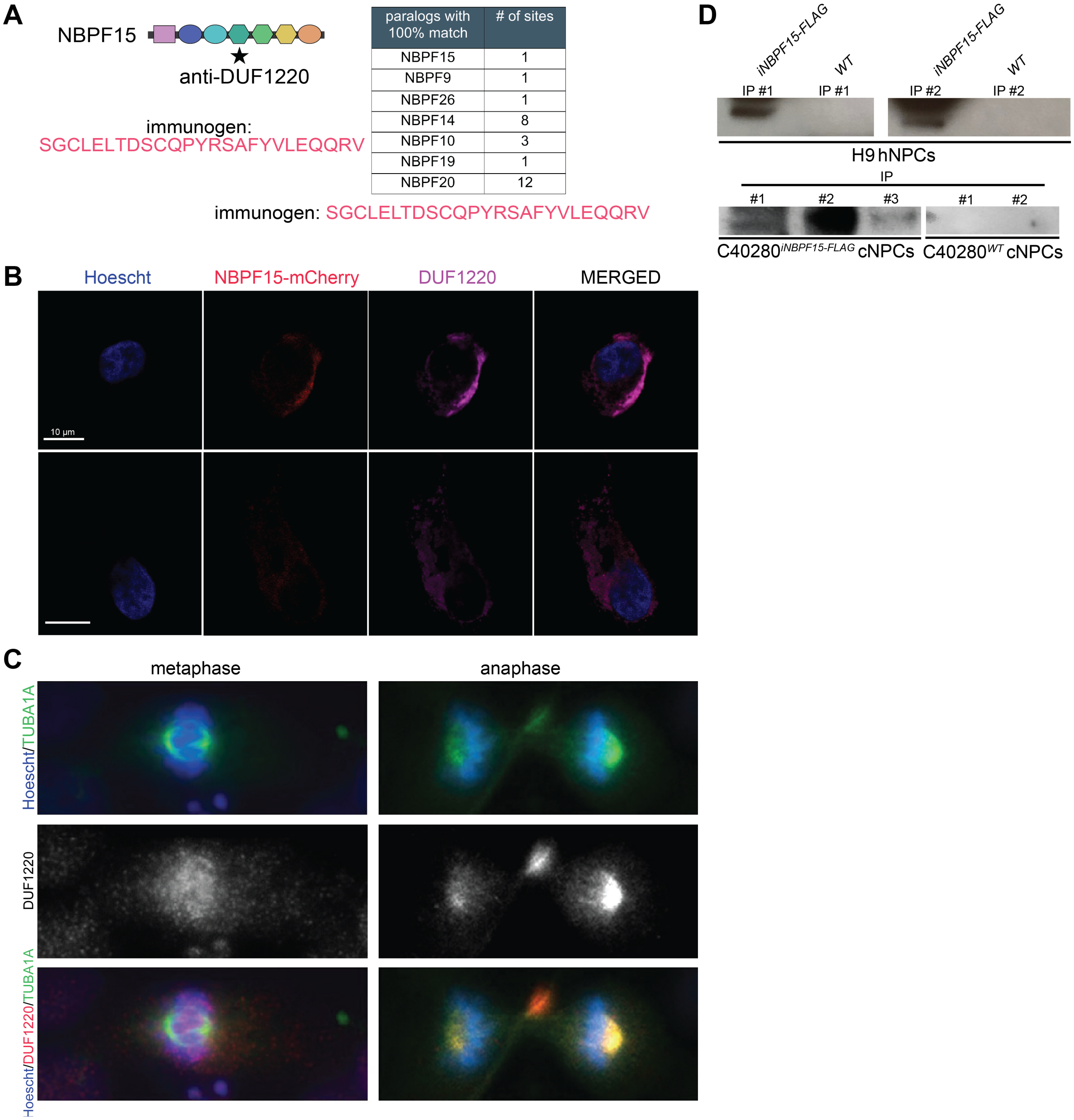


Figure 3.7. scRNAseq analysis of DUF1220-depleted cortical organoids.
(A) UMAP plot showing cell clusters following reciprocal PCA integration of H9 edited and control genotypes, separated by genotype. Four replicate organoids were analyzed per genotype. Cluster annotations: NECs, neuroepithelia cells. DL, deep layer. IPs, intermediate progenitors. (B) Proportions of cells in each cluster for each genotype. (C) Scatterplot showing notable skews in expression level towards one genotype. (D) Violin plots of cell expression level differences between genotypes for specific markers within clusters: proliferation markers, *PCNA* and *MKI67,* within NPC and mitotic clusters, DCX within the immature neuron cluster, and TBR1 and CTIP2 in cortical neuron clusters.



paralogs with 100% match	# of sites
NBPF15	1
NBPF9	1
NBPF26	1
NBPF14	8
NBPF10	3
NBPF19	1
NBPF20	12

Figure S3.1. NBPF IHC and IP-MS.
(A) Immunogen sequence of anti-NBPF-DUF1220 antibody was derived from the HLS1 DUF1220 domain of human NBPF15, and this sequence is found at 100% sequence identity in the NBPF paralogs listed in the table. Some paralogs encode more than one repeat of this sequence. (B) Immunostaining of endogenous NBPF (DUF1220) in electroporated H9 NPCs transiently expressing NBPF15-mCherry confirms binding specificity of DUF1220 antibody for NBPF proteins. (C) Immunofluorescent staining of H9 wildtype NPCs showing co-localization with TUBA1A at spindle poles, microtubules, and midbody (center in anaphase panel). (D) FLAG-immunoprecipitation on NPC bait replicates and controls from 20% sample-M2 magnetic bead fraction for IP-MS submission. Top, human samples; bottom, chimpanzee samples.

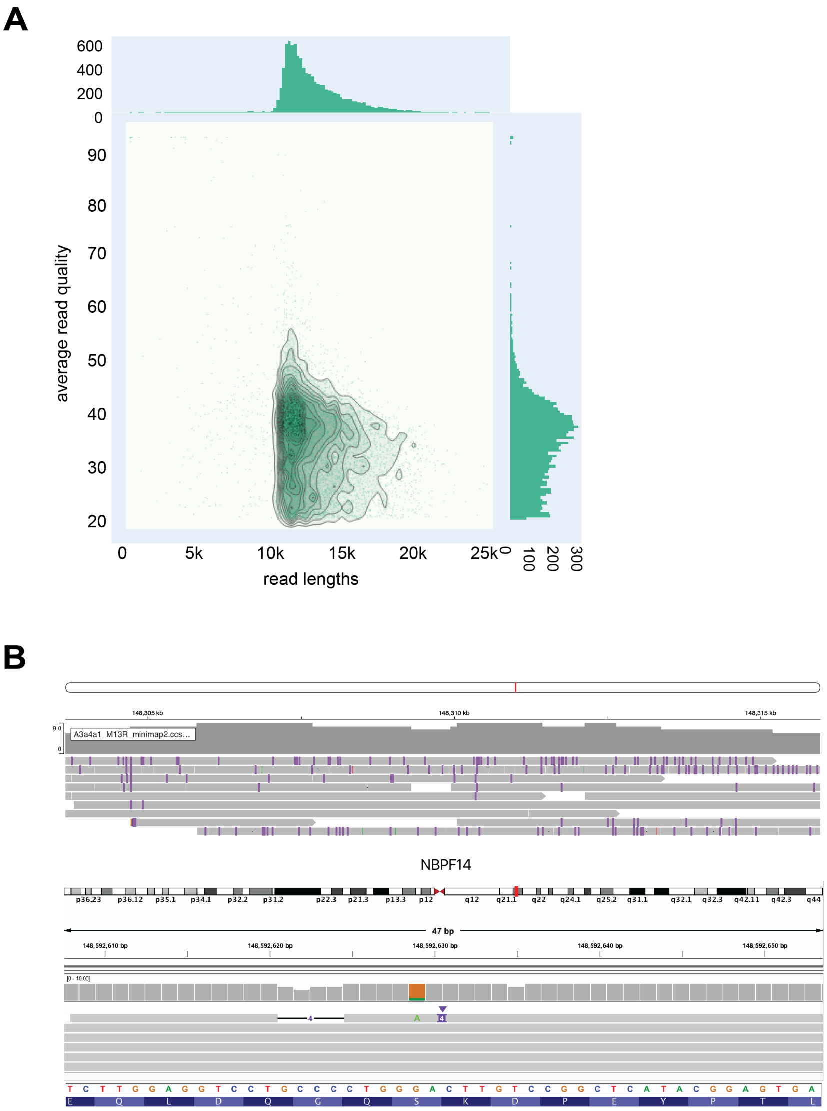


Figure S3.2. PacBio long read HiFi sequencing.
(A) Nanoplot of HiFi read quality distribution by read length. (B) IGV screenshots of Winnowmap alignment of HiFi reads at NBPF14 locus (top) and the edit site in NBPF14 exon 3 (bottom).

Table 3. List of oligo sequences used for *NBPF* cloning and sequencing.

ID	Sequence
NBPF14/10 forward primer	5’ TTCTCCAAGGGGCTCAATCG
M13 reverse primer	5’ GTCATAGCTGTTTCCTG
NBPF14/10 sgRNA sense oligo	5’ CACCG CCTCCTCACTCCGTATGAGC
NBPF14/10 sgRNA antisense oligo	5' AAAC GCTCATACGGAGTGAGGAGG C
NBPF14/10 ssODN	5’ TCACTCTCAGGAACGAGAGCTGACCCAGCTAAGGGAGAAGTTACGGGAAGGGAGAGATGCCTCCCGCTCATTGTATGAGCATCTCCAGGCCCTCCTCACTCCGTATGAGCCGTCATAGCTGTTTCCTGAGGACCTCCAAGAACAGCTGGCTGAGGGGTGTAGACTGGCACAGCACCTTGTCCAAAAGCTCAGCCCAGGTA
pPBhCMV1-NBPF15 forward primer	5’ CGACTGTGCCTTCTAGTTGC
pPBhCMV1-NBPF15 reverse primer	5’ TTGTCTTCCCAATCCTCCCC

Table 4. hNBPF15-FLAG IP-MS top hits in hNPCs.
Log$_2$ (Abundance ratio) > 2.5; *p*-value < 0.001. *PSM values instead of abundance counts were provided.

Protein	**Grouped Abundances: H9^{WT}**	**Grouped Abundances: H9$^{iNBPF15\text{-}FLAG}$**	**Abundance ratio (log2): H9$^{iNBPF15\text{-}FLAG}$/ H9^{WT}**	***p-value***	**CRAPome # experiments found/total**
NBPF1*	83.135	1	0.000461934	2.392E-05	0 / 716
NBPF9*	105.5	1	0.000265628	1.62E-38	1 / 716
NBPF15	476131.373	255984609	10.83597901	2.372E-19	1 / 716
DNAJC7	96182.1367	74751127.5	9.602110647	0.0001526	150 / 716
S100A10	4173.60156	222118.375	5.733892058	3.16E-05	17 / 716
HSPH1	93439.5542	2708513.7	4.857324188	4.012E-10	347 / 716
RPL5	446079.447	12793021.8	4.841912595	0.0009438	395 / 716
PRPS1	5180.38281	135507.703	4.709172344	0.0008227	225 / 716
HSPB1	888778.573	21657933.8	4.606927766	0.0001541	128 / 716
SMARCA5	12737.6045	244658.391	4.263602875	6.258E-05	210 / 716
HMGN3	103743.898	1940174.33	4.225087893	0.0004497	19 / 716
ARHGEF2	90092.9795	1636981.39	4.183479426	0.0001385	78 / 716
RPL23	448208.178	7866677.16	4.133513502	0.0002965	507 / 716
HSPA1B	218854.456	3816118.37	4.124062257	4.439E-06	698 / 716
HIST3H3	2312383.39	39063578.3	4.078371587	9.597E-05	383 / 716
PARP1	223395.166	3685269.27	4.044100164	0.0008533	438 / 716
SLC25A1	11131.9971	177212.508	3.992696095	2.732E-05	175 / 716
HSPA8	5354806.02	84789039.6	3.984971564	5.146E-06	703 / 716
BLVRA	48651.0923	739245.664	3.925509775	0.0001494	98 / 716
EIF3B	18856.5176	272213.18	3.85160175	0.0007739	282 / 716
ABCD3	7015.17236	91161.6445	3.699876494	1.347E-05	148 / 716
HMGA1	2182108.62	27127954.8	3.635985461	0.0009774	174 / 716
RPS28	83253.375	1025677.09	3.622924041	3.58E-06	404 / 716
MAP4	179123.219	2198653.49	3.61759599	0.0002552	312 / 716
RPL21	276973.992	3289728.33	3.57014603	0.0003423	329 / 716
HMGN2	4917460.5	58244907.8	3.566146556	9.701E-05	193 / 716
CKB	165063.53	1822395.25	3.464742585	0.0007687	308 / 716
MAT2A	63155.146	607603.223	3.266157324	0.000429	228 / 716
RNPS1	54549.2344	519052.188	3.250248747	0.0004854	242 / 716
PSMC4	28118.7148	262390.836	3.222114771	1.285E-05	210 / 716
ACTA1	1512511.17	14104328	3.22112023	8.075E-05	656 / 716
ABCF1	107597.569	970633.703	3.173281468	0.0002302	288 / 716
EIF1AX	1717625.72	15122995.8	3.13825636	0.000326	157 / 716

RPS20	246056.029	2092228.15	3.087981409	0.0004356	383 / 716
TMEM33	40680.3945	342525.625	3.073806347	0.0002261	144 / 716
TWISTNB	294120.391	2473113.92	3.071849984	0.0006665	14 / 716
RPL38	238602.813	2005413.84	3.071217033	0.0004644	300 / 716
PTHLH	18609.8477	155587.414	3.063587211	0.0002779	/ 716
RPL17-C18orf32	3565042.75	28795784.6	3.013866352	0.0003701	424 / 716
HIST1H1E	13558727.4	109126506	3.008707874	9.595E-06	605 / 716
CCT3	370729.033	2932396.2	2.983643031	0.0001329	367 / 716
MYEF2	16105.5908	126582.496	2.974444431	4.343E-05	27 / 716
HMGA2	2685060.8	20312657.1	2.919352309	0.0002001	37 / 716
HDLBP	66815.0508	500639.867	2.905528155	6.442E-06	218 / 716
PDHA1	39756.9512	297771.813	2.904928162	1.322E-05	126 / 716
RCN2	66427.3047	489864.969	2.882535842	0.0009628	208 / 716
KPNA4	25714.6094	188214.359	2.871716556	4.361E-05	127 / 716
RPL7A	507829.704	3658563.44	2.848860586	2.323E-05	398 / 716
RPS11	2427383.66	17205212	2.82537161	3.978E-05	340 / 716
RPS23	2280336.19	16106463.4	2.820321309	4.452E-05	413 / 716
PSMC6	41918.5	290198.805	2.791382579	5.56E-05	179 / 716
EIF3D	21623.0234	149029.367	2.784956484	0.0009736	249 / 716
VAPA	55416.0898	379616.539	2.776166026	0.0009316	116 / 716
FRG1	867202.379	5881087.14	2.761642248	0.0002929	138 / 716
MORF4L2	233846.688	1578623.9	2.755032595	0.0006308	76 / 716
ILF3	296070.009	1987188.78	2.74671867	0.000344	379 / 716
RPL28	939669.189	6276102.89	2.739644154	0.0001875	316 / 716
TUFM	617267.316	4069279.47	2.720806055	0.0001046	408 / 716
HNRNPM	2316386.13	15181453.5	2.712362254	0.000737	520 / 716
HSPA9	777408.136	5071337.75	2.705622249	9.624E-05	552 / 716
IMPDH2	146258.004	948911.047	2.697757274	1.228E-05	195 / 716
RPL30	182714.473	1176518.36	2.686861011	9.075E-05	255 / 716
GNB2L1	1057689.28	6803339.13	2.685327145	5.113E-05	Not identified
PABPC1	236542.924	1511069.56	2.675396168	2.211E-05	408 / 716
RPSA	419525.263	2654073.03	2.661378476	7.437E-06	335 / 716
GPX1	24697.5078	155788.375	2.657150209	0.0009395	68 / 716
RPL26	3323896.13	20941700.7	2.655431405	0.000131	312 / 716
TUBB6	126841.652	798012.215	2.653382253	0.0001804	617 / 716
RPL32	1168233.22	7304372.7	2.644432066	3.091E-06	184 / 716
CBX5	243830.916	1522356.81	2.64235357	0.000104	100 / 716
IDH2	168522.16	1035965.46	2.61996568	0.0008435	65 / 716

RPS2	1617969.84	9733437.09	2.588764624	0.0002941	447 / 716
HMGN4	4166813.11	25039268.5	2.587176118	6.577E-05	31 / 716
TMPO	16156.8623	94639.5586	2.550296294	2.268E-06	386 / 716
RPS3	2375040.53	13803511.2	2.539011259	5.191E-07	491 / 716
RPS16	974599.867	5647141.98	2.534638974	0.000312	411 / 716
HSDL1	23591.1328	136621.906	2.533872225	5.017E-06	9 / 716
FASN	1151607.27	6608939.97	2.52077009	1.61E-05	436 / 716
DDX5	2162319.98	12350652.1	2.513935281	0.0002825	528 / 716
NAT10	262284.457	1494934.3	2.510877857	9.643E-06	196 / 716

Table 5. hNBPF15-FLAG IP-MS top hits in cNPCs.
Log$_2$ (Abundance ratio) > 1; adj. *p*-value < 0.05.

Protein	**Grouped Abundances: C40280^{WT}**	**Grouped Abundances: C40280$^{iNBPF15\text{-}FLAG}$**	**Abundance ratio (log$_2$): C40280$^{iNBPF15\text{-}FLAG}$ /C40280^{WT}**	**adjusted *p*-value**	**CRAPome # experiments found/total**
CD38	28.4	171.6	2.79	2.1558E-09	0 /716
NBPF15	23.1	176.9	2.64	2.9563E-07	0 /716
RIMS1	65.4	134.6	2.61	2.7353E-09	1 /716
LOXL3	34.8	165.2	2.48	9.9044E-10	0 /716
IQCD	33.3	166.7	2.33	8.6896E-08	1 /716
RPAP1	37	163	2.14	0.00027963	29 /716
DDA1	32.5	167.5	1.98	0.00078047	4 /716
KCNH4	28.9	171.1	1.95	0.00076639	1 /716
NTPCR	44.1	155.9	1.78	0.00026923	101 /716
CGNL1	59.2	140.8	1.77	0.01218625	9 /716
ARFGAP2	48.4	151.6	1.76	0.00985676	34 /716
STOM	43.9	156.1	1.76	0.01393707	10 /716
KLHL9	35.6	164.4	1.67	0.00036412	0 /716
RILPL1	47.1	152.9	1.66	0.00036412	2 /716
PRSS1	42.6	157.4	1.65	0.00372851	108 /716
H2BC13	55	145	1.62	0.00463814	Not identified
TRIM29	52.9	147.1	1.58	0.00985676	3 /716
H1-0	51.9	148.1	1.57	0.00093686	Not identified
THSD4	56.9	143.1	1.57	0.01269582	2 /716
CDC73	50.7	149.3	1.56	0.01726639	162 /716
FRZB	51.4	148.6	1.53	0.00504721	0 /716
SERPINB9	47.5	152.5	1.53	0.00637089	3 /716
ITGA8	51.9	148.1	1.49	0.00242035	0 /716
PRSS1	40.4	159.6	1.49	0.01218625	108 /716
RDH10	41.8	158.2	1.49	0.03378906	0 /716
A0A2I3TBV7	26.6	173.4	1.49	0.03959285	Not identified
A0A2I3RS64	33.4	166.6	1.42	0.00463814	Not identified
SREK1IP1	55.9	144.1	1.37	0.02769207	72 /716
MT1X	56.2	143.8	1.21	0.0272052	24 /716
FBLN1	62.8	137.2	1.2	0.03508771	3 /716
CTNNBL1	62	138	1.15	0.04003666	61 /716

REFERENCES

Andrews MG, Subramanian L, Kriegstein AR (2020) mTOR signaling regulates the morphology and migration of outer radial glia in developing human cortex. Elife 9. doi: 10.7554/eLife.58737

Andries V, Vandepoele K, Staes K, Berx G, Bogaert P, Van Isterdael G, Ginneberge D, Parthoens E, Vandenbussche J, Gevaert K, van Roy F (2015) NBPF1, a tumor suppressor candidate in neuroblastoma, exerts growth inhibitory effects by inducing a G1 cell cycle arrest. BMC Cancer 15: 391. doi: 10.1186/s12885-015-1408-5

Astling DP, Heft IE, Jones KL, Sikela JM (2017) High resolution measurement of DUF1220 domain copy number from whole genome sequence data. BMC Genomics 18: 614. doi: 10.1186/s12864-017-3976-z

Banani SF, Lee HO, Hyman AA, Rosen MK (2017) Biomolecular condensates: organizers of cellular biochemistry. Nat Rev Mol Cell Biol 18: 285-298. doi: 10.1038/nrm.2017.7

Bernier R, Steinman KJ, Reilly B, Wallace AS, Sherr EH, Pojman N, Mefford HC, Gerdts J, Earl R, Hanson E, Goin-Kochel RP, Berry L, Kanne S, Snyder LG, Spence S, Ramocki MB, Evans DW, Spiro JE, Martin CL, Ledbetter DH, Chung WK, consortium SV (2016) Clinical phenotype of the recurrent 1q21.1 copy-number variant. Genet Med 18: 341-9. doi: 10.1038/gim.2015.78

Castillo PE, Schoch S, Schmitz F, Südhof TC, Malenka RC (2002) RIM1alpha is required for presynaptic long-term potentiation. Nature 415: 327-30. doi: 10.1038/415327a

Chen HL, Yuh CH, Wu KK (2010) Nestin is essential for zebrafish brain and eye development through control of progenitor cell apoptosis. PLoS One 5: e9318. doi: 10.1371/journal.pone.0009318

Crow TJ (2000) Schizophrenia as the price that homo sapiens pays for language: a resolution of the central paradox in the origin of the species. Brain Res Brain Res Rev 31: 118-29. doi: 10.1016/s0165-0173(99)00029-6

Davis JM, Heft I, Scherer SW, Sikela JM (2019) A Third Linear Association Between Olduvai (DUF1220) Copy Number and Severity of the Classic Symptoms of Inherited Autism. Am J Psychiatry 176: 643-650. doi: 10.1176/appi.ajp.2018.18080993

Davis JM, Searles Quick VB, Sikela JM (2015) Replicated linear association between DUF1220 copy number and severity of social impairment in autism. Hum Genet 134: 569-75. doi: 10.1007/s00439-015-1537-6

Davis JM, Searles VB, Anderson N, Keeney J, Dumas L, Sikela JM (2014) DUF1220 dosage is linearly associated with increasing severity of the three primary symptoms of autism. PLoS Genet 10: e1004241. doi: 10.1371/journal.pgen.1004241

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C (2018) NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 34: 2666-2669. doi: 10.1093/bioinformatics/bty149

Dennis MY, Eichler EE (2016) Human adaptation and evolution by segmental duplication. Curr Opin Genet Dev 41: 44-52. doi: 10.1016/j.gde.2016.08.001

Dennis MY, Harshman L, Nelson BJ, 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 HAF, Nuttle X, Hoekzema K, Lindsay-Graves TA, Wilson RK, Eichler EE (2017) The evolution and population diversity of human-specific segmental duplications. Nat Ecol Evol 1: 69. doi: 10.1038/s41559-016-0069

Ditlev JA, Case LB, Rosen MK (2018) Who's In and Who's Out-Compositional Control of Biomolecular Condensates. J Mol Biol 430: 4666-4684. doi: 10.1016/j.jmb.2018.08.003

Dougherty ML, Nuttle X, Penn O, Nelson BJ, Huddleston J, Baker C, Harshman L, Duyzend MH, Ventura M, Antonacci F, Sandstrom R, Dennis MY, Eichler EE (2017) The birth of a human-specific neural gene by incomplete duplication and gene fusion. Genome Biol 18: 49. doi: 10.1186/s13059-017-1163-9

Dumas L, Kim YH, Karimpour-Fard A, Cox M, Hopkins J, Pollack JR, Sikela JM (2007) Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res 17: 1266-77. doi: 10.1101/gr.6557307

Ferent J, Zaidi D, Francis F (2020) Extracellular Control of Radial Glia Proliferation and Scaffolding During Cortical Development and Pathology. Front Cell Dev Biol 8: 578341. doi: 10.3389/fcell.2020.578341

Fiddes IT, Lodewijk GA, Mooring M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, van den Bout A, Bishara A, Rosenkrantz JL, Lorig-Roach R, Field AR, Haeussler M, Russo L, Bhaduri A, Nowakowski TJ, Pollen AA, Dougherty ML, Nuttle X, Addor MC, Zwolinski S, Katzman S, Kriegstein A, Eichler EE, Salama SR, Jacobs FMJ, Haussler D (2018) Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell 173: 1356-1369.e22. doi: 10.1016/j.cell.2018.03.051

Fiddes IT, Pollen AA, Davis JM, Sikela JM (2019) Paired involvement of human-specific Olduvai domains and NOTCH2NL genes in human brain evolution. Hum Genet 138: 715-721. doi: 10.1007/s00439-019-02018-4

Florio M, Albert M, Taverna E, Namba T, Brandl H, Lewitus E, Haffner C, Sykes A, Wong FK, Peters J, Guhr E, Klemroth S, Prüfer K, Kelso J, Naumann R, Nüsslein I, Dahl A, Lachmann R, Pääbo S, Huttner WB (2015) Human-specific gene ARHGAP11B promotes basal progenitor amplification and neocortex expansion. Science 347: 1465-70. doi: 10.1126/science.aaa1975

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531-45. doi: 10.1093/genetics/151.4.1531

Frigerio G, Grimsey N, Dale M, Majoul I, Duden R (2007) Two human ARFGAPs associated with COP-I-coated vesicles. Traffic 8: 1644-55. doi: 10.1111/j.1600-0854.2007.00631.x

Gallego Romero I, Pavlovic BJ, Hernando-Herraez I, Zhou X, Ward MC, Banovich NE, Kagan CL, Burnett JE, Huang CH, Mitrano A, Chavarria CI, Friedrich Ben-Nun I, Li Y, Sabatini K, Leonardo TR, Parast M, Marques-Bonet T, Laurent LC, Loring JF, Gilad Y (2015) A panel of induced pluripotent stem cells from chimpanzees: a resource for comparative functional genomics. Elife 4: e07103. doi: 10.7554/eLife.07103

Gierahn TM, Wadsworth MH, 2nd, Hughes TK, Bryson BD, Butler A, Satija R, Fortune S, Love JC, Shalek AK (2017) Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods 14: 395-398. doi: 10.1038/nmeth.4179

Gilbert CC, Jungers WL (2017) Comment on relative brain size in early primates and the use of encephalization quotients in primate evolution. J Hum Evol 109: 79-87. doi: 10.1016/j.jhevol.2017.04.007

Gupta GD, Pelletier L (2017) Centrosome Biology: Polymer-Based Centrosome Maturation. Curr Biol 27: R836-R839. doi: 10.1016/j.cub.2017.07.036

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, Hoffman P, Stoeckius M, Papalexi E, Mimitou EP, Jain J, Srivastava A, Stuart T, Fleming LM, Yeung B, Rogers AJ, McElrath JM, Blish CA, Gottardo R, Smibert P, Satija R (2021) Integrated analysis of multimodal single-cell data. Cell 184: 3573-3587.e29. doi: 10.1016/j.cell.2021.04.048

Heft IE, Mostovoy Y, Levy-Sakin M, Ma W, Stevens AJ, Pastor S, McCaffrey J, Boffelli D, Martin DI, Xiao M, Kennedy MA, Kwok PY, Sikela JM (2020) The Driver of Extreme Human-Specific Olduvai Repeat Expansion Remains Highly Active in the Human Genome. Genetics 214: 179-191. doi: 10.1534/genetics.119.302782

Herculano-Houzel S (2012) Neuronal scaling rules for primate brains: the primate advantage. Prog Brain Res 195: 325-40. doi: 10.1016/B978-0-444-53860-4.00015-5

Herculano-Houzel S, Collins CE, Wong P, Kaas JH (2007) Cellular scaling rules for primate brains. Proc Natl Acad Sci U S A 104: 3562-7. doi: 10.1073/pnas.0611396104

Herculano-Houzel S, Kaas JH (2011) Gorilla and orangutan brains conform to the primate cellular scaling rules: implications for human evolution. Brain Behav Evol 77: 33-44. doi: 10.1159/000322729

Hill R, Wu H (2009) PTEN, stem cells, and cancer stem cells. J Biol Chem 284: 11755-9. doi: 10.1074/jbc.R800071200

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc Biol Sci 256: 119-24. doi: 10.1098/rspb.1994.0058

Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM (2022) Long-read mapping to repetitive reference sequences using Winnowmap2. Nat Methods 19: 705-710. doi: 10.1038/s41592-022-01457-8

Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM (2020) Weighted minimizer sampling improves long read mapping. Bioinformatics 36: i111-i118. doi: 10.1093/bioinformatics/btaa435

Jiang H, Wang S, Huang Y, He X, Cui H, Zhu X, Zheng Y (2015) Phase transition of spindle-associated protein regulate spindle apparatus assembly. Cell 163: 108-22. doi: 10.1016/j.cell.2015.08.010

Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, Guijarro P, Sidow L, Fleck JS, Han D, Qian Z, Heide M, Huttner WB, Khaitovich P, Pääbo S, Treutlein B, Camp JG (2019) Organoid single-cell genomic atlas uncovers human-specific features of brain development. Nature 574: 418-422. doi: 10.1038/s41586-019-1654-9

Keeney JG, Davis JM, Siegenthaler J, Post MD, Nielsen BS, Hopkins WD, Sikela JM (2015a) DUF1220 protein domains drive proliferation in human neural stem cells and are associated with increased cortical volume in anthropoid primates. Brain Struct Funct 220: 3053-60. doi: 10.1007/s00429-014-0814-9

Keeney JG, Dumas L, Sikela JM (2014) The case for DUF1220 domain dosage as a primary contributor to anthropoid brain expansion. Front Hum Neurosci 8: 427. doi: 10.3389/fnhum.2014.00427

Keeney JG, O'Bleness MS, Anderson N, Davis JM, Arevalo N, Busquet N, Chick W, Rozman J, Hölter SM, Garrett L, Horsch M, Beckers J, Wurst W, Klingenspor M, Restrepo D, de Angelis MH, Sikela JM, Consortium GMC (2015b) Generation of mice lacking DUF1220 protein domains: effects on fecundity and hyperactivity. Mamm Genome 26: 33-42. doi: 10.1007/s00335-014-9545-8

Kono K, Yoshiura S, Fujita I, Okada Y, Shitamukai A, Shibata T, Matsuzaki F (2019) Reconstruction of Par-dependent polarity in apolar cells reveals a dynamic process of cortical polarization. Elife 8. doi: 10.7554/eLife.45559

Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS, Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, Munson KM, Hastie AR, Diekhans M, Hormozdiari F, Lorusso N, Hoekzema K, Qiu R, Clark K, Raja A, Welch AE, Sorensen M, Baker C, Fulton RS, Armstrong J, Graves-Lindsay TA, Denli AM, Hoppe ER, Hsieh P, Hill CM, Pang AWC, Lee J, Lam ET, Dutcher SK, Gage FH, Warren WC, Shendure J, Haussler D, Schneider VA, Cao H, Ventura M, Wilson RK, Paten B, Pollen A, Eichler EE (2018) High-resolution comparative analysis of great ape genomes. Science 360. doi: 10.1126/science.aar6343

Lancaster MA, Renner M, Martin CA, Wenzel D, Bicknell LS, Hurles ME, Homfray T, Penninger JM, Jackson AP, Knoblich JA (2013) Cerebral organoids model human brain development and microcephaly. Nature 501: 373-9. doi: 10.1038/nature12517

Laureys G, Speleman F, Opdenakker G, Benoit Y, Leroy J (1990) Constitutional translocation t(1;17)(p36;q12-21) in a patient with neuroblastoma. Genes Chromosomes Cancer 2: 252-4. doi: 10.1002/gcc.2870020315

Laureys G, Speleman F, Versteeg R, van der Drift P, Chan A, Leroy J, Francke U, Opdenakker G, Van Roy N (1995) Constitutional translocation t(1;17)(p36.31-p36.13;q11.2-q12.1) in a neuroblastoma patient. Establishment of somatic cell hybrids and identification of PND/A12M2 on chromosome 1 and NF1/SCYA7 on chromosome 17 as breakpoint flanking single copy markers. Oncogene 10: 1087-93.

Leyns L, Bouwmeester T, Kim SH, Piccolo S, De Robertis EM (1997) Frzb-1 is a secreted antagonist of Wnt signaling expressed in the Spemann organizer. Cell 88: 747-56. doi: 10.1016/s0092-8674(00)81921-2

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34: 3094-3100. doi: 10.1093/bioinformatics/bty191

Li J, Yen C, Liaw D, Podsypanina K, Bose S, Wang SI, Puc J, Miliaresis C, Rodgers L, McCombie R, Bigner SH, Giovanella BC, Ittmann M, Tycko B, Hibshoosh H, Wigler MH, Parsons R (1997) PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science 275: 1943-7. doi: 10.1126/science.275.5308.1943

Liao YX, Zhang ZP, Zhao J, Liu JP (2018) Effects of Fibronectin 1 on Cell Proliferation, Senescence and Apoptosis of Human Glioma Cells Through the PI3K/AKT Signaling Pathway. Cell Physiol Biochem 48: 1382-1396. doi: 10.1159/000492096

Lin YH, Forman-Kay JD, Chan HS (2018) Theories for Sequence-Dependent Phase Behaviors of Biomolecular Condensates. Biochemistry 57: 2499-2508. doi: 10.1021/acs.biochem.8b00058

Linden SC, Watson CJ, Smith J, Chawner SJRA, Lancaster TM, Evans F, Williams N, Skuse D, Raymond FL, Hall J, Owen MJ, Linden DEJ, Green-Snyder L, Chung WK, Maillard AM, Jacquemont S, van den Bree MBM (2021) The psychiatric phenotypes of 1q21 distal deletion and duplication. Transl Psychiatry 11: 105. doi: 10.1038/s41398-021-01226-9

Liu Z, Yang Y, Gu A, Xu J, Mao Y, Lu H, Hu W, Lei QY, Li Z, Zhang M, Cai Y, Wen W (2020) Par complex cluster formation mediated by phase separation. Nat Commun 11: 2266. doi: 10.1038/s41467-020-16135-6

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151-5. doi: 10.1126/science.290.5494.1151

Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genomics 3: 35-44.

Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20: 544-9. doi: 10.1016/j.tig.2004.09.001

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, McCarroll SA (2015) Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 161: 1202-1214. doi: 10.1016/j.cell.2015.05.002

Maerki S, Olma MH, Staubli T, Steigemann P, Gerlich DW, Quadroni M, Sumara I, Peter M (2009) The Cul3-KLHL21 E3 ubiquitin ligase targets aurora B to midzone microtubules in anaphase and is required for cytokinesis. J Cell Biol 187: 791-800. doi: 10.1083/jcb.200906117

Manzo G (2019) Similarities Between Embryo Development and Cancer Process Suggest New Strategies for Research and Therapy of Tumors: A New Point of View. Front Cell Dev Biol 7: 20. doi: 10.3389/fcell.2019.00020

Martucci LL, Amar M, Chaussenot R, Benet G, Bauer O, de Zélicourt A, Nosjean A, Launay JM, Callebert J, Sebrié C, Galione A, Edeline JM, de la Porte S, Fossier P, Granon S, Vaillend C, Cancela JM (2019) A multiscale analysis in CD38. FASEB J 33: 5823-5835. doi: 10.1096/fj.201800489R

Mehta S, Zhang J (2022) Liquid-liquid phase separation drives cellular function and dysfunction in cancer. Nat Rev Cancer 22: 239-252. doi: 10.1038/s41568-022-00444-7

Mora-Bermúdez F, Badsha F, Kanton S, Camp JG, Vernot B, Köhler K, Voigt B, Okita K, Maricic T, He Z, Lachmann R, Pääbo S, Treutlein B, Huttner WB (2016) Differences and similarities between human and chimpanzee neural progenitors during cerebral cortex development. Elife 5. doi: 10.7554/eLife.18683

Munesue T, Yokoyama S, Nakamura K, Anitha A, Yamada K, Hayashi K, Asaka T, Liu HX, Jin D, Koizumi K, Islam MS, Huang JJ, Ma WJ, Kim UH, Kim SJ, Park K, Kim D, Kikuchi M, Ono Y, Nakatani H, Suda S, Miyachi T, Hirai H, Salmina A, Pichugina YA, Soumarokov AA, Takei N, Mori N, Tsujii M, Sugiyama T, Yagi K, Yamagishi M, Sasaki T, Yamasue H, Kato N, Hashimoto R, Taniike M, Hayashi Y, Hamada J, Suzuki S, Ooi A, Noda M, Kamiyama Y, Kido MA, Lopatina O, Hashii M, Amina S, Malavasi F, Huang EJ, Zhang J, Shimizu N, Yoshikawa T, Matsushima A, Minabe Y, Higashida H (2010) Two genetic variants of CD38 in subjects with autism spectrum disorder and controls. Neurosci Res 67: 181-91. doi: 10.1016/j.neures.2010.03.004

Mönnich M, Borgeskov L, Breslin L, Jakobsen L, Rogowski M, Doganli C, Schrøder JM, Mogensen JB, Blinkenkjær L, Harder LM, Lundberg E, Geimer S, Christensen ST, Andersen JS, Larsen LA, Pedersen LB (2018) CEP128 Localizes to the Subdistal Appendages of the Mother Centriole and Regulates TGF-β/BMP Signaling at the Primary Cilium. Cell Rep 22: 2584-2592. doi: 10.1016/j.celrep.2018.02.043

Nowakowski TJ, Bhaduri A, Pollen AA, Alvarado B, Mostajo-Radji MA, Di Lullo E, Haeussler M, Sandoval-Espinosa C, Liu SJ, Velmeshev D, Ounadjela JR, Shuga J, Wang X, Lim DA, West JA, Leyrat AA, Kent WJ, Kriegstein AR (2017) Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. Science 358: 1318-1323. doi: 10.1126/science.aap8809

O'Bleness M, Searles VB, Dickens CM, Astling D, Albracht D, Mak AC, Lai YY, Lin C, Chu C, Graves T, Kwok PY, Wilson RK, Sikela JM (2014) Finished sequence and assembly of the DUF1220-rich 1q21 region using a haploid human genome. BMC Genomics 15: 387. doi: 10.1186/1471-2164-15-387

O'Bleness MS, Dickens CM, Dumas LJ, Kehrer-Sawatzki H, Wyckoff GJ, Sikela JM (2012) Evolutionary history and genome organization of DUF1220 protein domains. G3 (Bethesda) 2: 977-86. doi: 10.1534/g3.112.003061

O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, Munson J, Hiatt JB, Turner EH, Levy R, O'Day DR, Krumm N, Coe BP, Martin BK, Borenstein E, Nickerson DA, Mefford HC, Doherty D, Akey JM, Bernier R, Eichler EE, Shendure J (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science 338: 1619-22. doi: 10.1126/science.1227764

Ohno S (1970) Evolution by gene duplication. Springer, New York

Parikshak NN, Luo R, Zhang A, Won H, Lowe JK, Chandran V, Horvath S, Geschwind DH (2013) Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell 155: 1008-21. doi: 10.1016/j.cell.2013.10.031

Piatigorsky J, Wistow G (1991) The recruitment of crystallins: new functions precede gene duplication. Science 252: 1078-9. doi: 10.1126/science.252.5009.1078

Poirier K, Saillour Y, Bahi-Buisson N, Jaglin XH, Fallet-Bianco C, Nabbout R, Castelnau-Ptakhine L, Roubertie A, Attie-Bitach T, Desguerre I, Genevieve D, Barnerias C, Keren B, Lebrun N, Boddaert N, Encha-Razavi F, Chelly J (2010) Mutations in the neuronal ß-tubulin subunit TUBB3 result in malformation of cortical development and neuronal migration defects. Hum Mol Genet 19: 4462-73. doi: 10.1093/hmg/ddq377

Pollen AA, Bhaduri A, Andrews MG, Nowakowski TJ, Meyerson OS, Mostajo-Radji MA, Di Lullo E, Alvarado B, Bedolli M, Dougherty ML, Fiddes IT, Kronenberg ZN, Shuga J, Leyrat AA, West JA, Bershteyn M, Lowe CB, Pavlovic BJ, Salama SR, Haussler D, Eichler EE, Kriegstein AR (2019) Establishing Cerebral Organoids as Models of Human-Specific Brain Evolution. Cell 176: 743-756.e17. doi: 10.1016/j.cell.2019.01.017

Popesco MC, Maclaren EJ, Hopkins J, Dumas L, Cox M, Meltesen L, McGavran L, Wyckoff GJ, Sikela JM (2006) Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. Science 313: 1304-7. doi: 10.1126/science.1127980

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, Cagan A, Theunert C, Casals F, Laayouni H, Munch K, Hobolth A, Halager AE, Malig M, Hernandez-Rodriguez J, Hernando-Herraez I, Prüfer K, Pybus M, Johnstone L, Lachmann M, Alkan C, Twigg D, Petit N, Baker C, Hormozdiari F, Fernandez-Callejo M, Dabad M, Wilson ML, Stevison L, Camprubí C, Carvalho T, Ruiz-Herrera A, Vives L, Mele M, Abello T, Kondova I, Bontrop RE, Pusey A, Lankester F, Kiyang JA, Bergl RA, Lonsdorf E, Myers S, Ventura M, Gagneux P, Comas D, Siegismund H, Blanc J, Agueda-Calpena L, Gut M, Fulton L, Tishkoff SA, Mullikin JC, Wilson RK, Gut IG, Gonder MK, Ryder OA, Hahn BH, Navarro A, Akey JM, Bertranpetit J, Reich D, Mailund T, Schierup MH, Hvilsom C, Andrés AM, Wall JD, Bustamante CD, Hammer MF, Eichler EE, Marques-Bonet T (2013) Great ape genetic diversity and population history. Nature 499: 471-5. doi: 10.1038/nature12228

Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet 3: 827-37. doi: 10.1038/nrg928

Qi H, Dong C, Chung WK, Wang K, Shen Y (2016) Deep Genetic Connection Between Cancer and Developmental Disorders. Hum Mutat 37: 1042-50. doi: 10.1002/humu.23040

Qin Y, Tang X, Liu M (2016) Tumor-Suppressor Gene NBPF1 Inhibits Invasion and PI3K/mTOR Signaling in Cervical Cancer Cells. Oncol Res 23: 13-20. doi: 10.3727/096504015X14410238486766

Quick VB, Davis JM, Olincy A, Sikela JM (2016) DUF1220 copy number is associated with schizophrenia risk and severity: implications for understanding autism and schizophrenia as related diseases. Transl Psychiatry 6: e735. doi: 10.1038/tp.2016.11

Raff JW (2019) Phase Separation and the Centrosome: A Fait Accompli? Trends Cell Biol 29: 612-622. doi: 10.1016/j.tcb.2019.04.001

Rebane AA, Ziltener P, LaMonica LC, Bauer AH, Zheng H, López-Montero I, Pincet F, Rothman JE, Ernst AM (2020) Liquid-liquid phase separation of the Golgi matrix protein GM130. FEBS Lett 594: 1132-1144. doi: 10.1002/1873-3468.13715

Roubin R, Acquaviva C, Chevrier V, Sedjaï F, Zyss D, Birnbaum D, Rosnet O (2013) Myomegalin is necessary for the formation of centrosomal and Golgi-derived microtubules. Biol Open 2: 238-50. doi: 10.1242/bio.20123392

Santoro A, Nicolin V, Florenzano F, Rosati A, Capunzo M, Nori SL (2017) BAG3 is involved in neuronal differentiation and migration. Cell Tissue Res 368: 249-258. doi: 10.1007/s00441-017-2570-7

Schubbert S, Shannon K, Bollag G (2007) Hyperactive Ras in developmental disorders and cancer. Nat Rev Cancer 7: 295-308. doi: 10.1038/nrc2109

Shahsavani M, Pronk RJ, Falk R, Lam M, Moslem M, Linker SB, Salma J, Day K, Schuster J, Anderlid BM, Dahl N, Gage FH, Falk A (2018) An in vitro model of lissencephaly: expanding the role of DCX during neurogenesis. Mol Psychiatry 23: 1674-1684. doi: 10.1038/mp.2017.175

Shan Z, Tu Y, Yang Y, Liu Z, Zeng M, Xu H, Long J, Zhang M, Cai Y, Wen W (2018) Basal condensation of Numb and Pon complex via phase transition during Drosophila neuroblast asymmetric division. Nat Commun 9: 737. doi: 10.1038/s41467-018-03077-3

Shew CJ, Carmona-Mora P, Soto DC, Mastoras M, Roberts E, Rosas J, Jagannathan D, Kaya G, O'Geen H, Dennis MY (2021) Diverse Molecular Mechanisms Contribute to Differential Expression of Human Duplicated Genes. Mol Biol Evol 38: 3060-3077. doi: 10.1093/molbev/msab131

Sikela JM, Searles Quick VB (2018) Genomic trade-offs: are autism and schizophrenia the steep price of the human brain? Hum Genet 137: 1-13. doi: 10.1007/s00439-017-1865-9

Smaers JB, Gómez-Robles A, Parks AN, Sherwood CC (2017) Exceptional Evolutionary Expansion of Prefrontal Cortex in Great Apes and Humans. Curr Biol 27: 1549. doi: 10.1016/j.cub.2017.05.015

Smaers JB, Rothman RS, Hudson DR, Balanoff AM, Beatty B, Dechmann DKN, de Vries D, Dunn JC, Fleagle JG, Gilbert CC, Goswami A, Iwaniuk AN, Jungers WL, Kerney M, Ksepka DT, Manger PR, Mongle CS, Rohlf FJ, Smith NA, Soligo C, Weisbecker V, Safi K (2021) The evolution of mammalian brain size. Sci Adv 7. doi: 10.1126/sciadv.abe2101

Speir ML, Bhaduri A, Markov NS, Moreno P, Nowakowski TJ, Papatheodorou I, Pollen AA, Raney BJ, Seninge L, Kent WJ, Haeussler M (2021) UCSC Cell Browser: Visualize Your Single-Cell Data. Bioinformatics. doi: 10.1093/bioinformatics/btab503

Sumara I, Quadroni M, Frei C, Olma MH, Sumara G, Ricci R, Peter M (2007) A Cul3-based E3 ligase removes Aurora B from mitotic chromosomes, regulating mitotic progression and completion of cytokinesis in human cells. Dev Cell 12: 887-900. doi: 10.1016/j.devcel.2007.03.019

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F, Detours V, Vanderhaeghen P (2018) Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. Cell 173: 1370-1384.e16. doi: 10.1016/j.cell.2018.03.067

Sweeney KJ, Clark GD, Prokscha A, Dobyns WB, Eichele G (2000) Lissencephaly associated mutations suggest a requirement for the PAFAH1B heterotrimeric complex in brain development. Mech Dev 92: 263-71. doi: 10.1016/s0925-4773(00)00242-2

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512-26. doi: 10.1093/oxfordjournals.molbev.a040023

Tartaglia M, Kalidas K, Shaw A, Song X, Musat DL, van der Burgt I, Brunner HG, Bertola DR, Crosby A, Ion A, Kucherlapati RS, Jeffery S, Patton MA, Gelb BD (2002) PTPN11 mutations in Noonan syndrome: molecular spectrum, genotype-phenotype correlation, and phenotypic heterogeneity. Am J Hum Genet 70: 1555-63. doi: 10.1086/340847

Tartaglia M, Niemeyer CM, Fragale A, Song X, Buechner J, Jung A, Hählen K, Hasle H, Licht JD, Gelb BD (2003) Somatic mutations in PTPN11 in juvenile myelomonocytic leukemia, myelodysplastic syndromes and acute myeloid leukemia. Nat Genet 34: 148-50. doi: 10.1038/ng1156

Thiruvalluvan A, de Mattos EP, Brunsting JF, Bakels R, Serlidaki D, Barazzuol L, Conforti P, Fatima A, Koyuncu S, Cattaneo E, Vilchez D, Bergink S, Boddeke EHWG, Copray S, Kampinga HH (2020) DNAJB6, a Key Factor in Neuronal Sensitivity to Amyloidogenesis. Mol Cell 78: 346-358.e9. doi: 10.1016/j.molcel.2020.02.022

Tiwary AK, Zheng Y (2019) Protein phase separation in mitosis. Curr Opin Cell Biol 60: 92-98. doi: 10.1016/j.ceb.2019.04.011

Van Bibber NW, Haerle C, Khalife R, Dayhoff GW, Uversky VN (2020) Intrinsic Disorder in Human Proteins Encoded by Core Duplicon Gene Families. J Phys Chem B 124: 8050-8070. doi: 10.1021/acs.jpcb.0c07676

Van Rossum G, Drake Jr FL (1995) Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam

Vandepoele K, Andries V, van Roy F (2009) The NBPF1 promoter has been recruited from the unrelated EVI5 gene before simian radiation. Mol Biol Evol 26: 1321-32. doi: 10.1093/molbev/msp047

Vandepoele K, Andries V, Van Roy N, Staes K, Vandesompele J, Laureys G, De Smet E, Berx G, Speleman F, van Roy F (2008) A constitutional translocation t(1;17)(p36.2;q11.2) in a neuroblastoma patient disrupts the human NBPF1 and ACCN1 genes. PLoS One 3: e2207. doi: 10.1371/journal.pone.0002207

Verde I, Pahlke G, Salanova M, Zhang G, Wang S, Coletti D, Onuffer J, Jin SL, Conti M (2001) Myomegalin is a novel protein of the golgi/centrosome that interacts with a cyclic nucleotide phosphodiesterase. J Biol Chem 276: 11189-98. doi: 10.1074/jbc.M006546200

Waite KA, Eng C (2003) From developmental disorder to heritable cancer: it's all in the BMP/TGF-beta family. Nat Rev Genet 4: 763-73. doi: 10.1038/nrg1178

Wright PE, Dyson HJ (2015) Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev Mol Cell Biol 16: 18-29. doi: 10.1038/nrm3920

Wu H, Zhai LT, Guo XX, Rety S, Xi XG (2020a) The N-terminal of NBPF15 causes multiple types of aggregates and mediates phase transition. Biochem J 477: 445-458. doi: 10.1042/BCJ20190566

Wu X, Cai Q, Feng Z, Zhang M (2020b) Liquid-Liquid Phase Separation in Neuronal Development and Synaptic Signaling. Dev Cell 55: 18-29. doi: 10.1016/j.devcel.2020.06.012

Zhang JJ (2003) Evolution by gene duplication: an update. *Trends in ecology & evolution* 18: 292-298.

Zhang X, Liu D, Lv S, Wang H, Zhong X, Liu B, Wang B, Liao J, Li J, Pfeifer GP, Xu X (2009) CDK5RAP2 is required for spindle checkpoint function. Cell Cycle 8: 1206-16. doi: 10.4161/cc.8.8.8205

Zimmer F, Montgomery SH (2015) Phylogenetic Analysis Supports a Link between DUF1220 Domain Number and Primate Brain Expansion. Genome Biol Evol 7: 2083-8. doi: 10.1093/gbe/evv122

Chapter 4
Biallelic *CSMD1* Variants as a Novel Genetic Basis of Malformations of Cortical Development

4.1 INTRODUCTION

Advances in clinical genomics testing using exome and genome sequencing has led to the discovery of novel genes and genetic variants implicated in neurodevelopmental and neuropsychiatric disorders (NND). One cluster of NND with a growing number of pathogenic variants identified by clinical genomics testing is malformations of cortical development (MCD). MCD is a broad group of heterogeneous disorders that often present with developmental delay, intellectual disability, and epilepsy (Barkovich et al., 2012; Guerrini and Dobyns, 2014). Pathogenic variants in over 100 genes have been implicated in the etiology of several classes of MCD (Guerrini and Dobyns, 2014). Here, we expand this list to include the *CSMD1* gene as a novel genetic basis of MCD.

CSMD1 has predominantly been studied within the context of immune-centered activation as a key regulator of the complement pathway (Escudero-Esparza et al., 2013; Kraus et al., 2006). In brief, the complement system does is comprised of a cluster of soluble proteins that facilitate the innate immune response by recognition and removal of pathogens as well as signal and sequester immune cells to sites of inflammation (Coulthard et al., 2018). New avenues of research implicate emerging patterns of *CSMD1* expression, activation, and function in development and physiology, notably brain development (Coulthard *et al.*, 2018). CSMD1 is an inhibitor of the complement pathway in neural tissues and is highly expressed in neurons with synaptic enrichment (Baum et al., 2020). The complement cascade has documented roles in neural proliferation, neural migration, and synaptic pruning (Coulthard *et al.*, 2018; Schafer et al., 2012)—affected aspects of biology in MCD.

Single nucleotide variants (SNVs) and copy number variants (CNVs) in *CSMD1* have been previously associated with strong risk for schizophrenia (Consortium, 2011; Sekar et al., 2016), Alzheimer's disease (Hong et al., 2016), Parkinson's disease (Ruiz-Martínez et al., 2017), and infertility (Lee et al., 2019). Consistent with gonadal association findings, *Csmd1* knockout (*Csmd1*^{KO}) mouse exhibit loss of germ cells within seminiferous tubules (Lee *et al.*, 2019). Consistent with associations with schizophrenia, cortical neurons differentiated from human embryonic stem cells (hESCs) with biallelic knockout variants (*CSMD1*$^{fs/fs}$) showed enhanced complement deposition (Baum *et al.*, 2020). Despite differences in synaptogenesis and neural circuitry that distinguish human and mouse brains, *Csmd1*^{KO} mouse neurons also showed elevated complement activity with fewer synapses and disruption to complement-dependent circuit formation, without overall structural brain defects (Baum *et al.*, 2020). These findings strengthen the implicated role of CSMD1 in mediating synaptic plasticity. However, human neural developmental phenotypes were not assessed.

Here, we identify novel biallelic variants in *CSMD1* in a cohort of individuals from seven families with overlapping and unique MCD clinical phenotypes, namely microcephaly, global developmental delay, intellectual disability, and polymicrogyria. To investigate CSMD1 function in human neural development, we generated human cerebral organoid models differentiated from *CSMD1*$^{fs/fs}$ hESCs. Forebrain-fated organoid models display elevated proliferation, premature differentiation, and disorganization cytoarchitecture. The lack of cortical structural defects in *Csmd1*^{KO} mouse (Baum *et al.*, 2020) may reflect functional divergence of CSMD1 in the developing cortex between human and mouse. The formation of proper neuronal circuits depends on successful neuronal migration tightly orchestrated by spatiotemporal architecture of the cortex during development (Marin et al., 2010). The secondary insults of this disruption on later synaptogenesis and synaptic plasticity may further underlie human-specific susceptibilities and expressivity of *CSMD1* variants. This represents the first study to propose *CSMD1* as a novel genetic basis of MCD.

4.2 RESULTS

4.2.1 Biallelic variants in *CSMD1* identified in individuals with MCD

Using exome-based sequencing, we identified ten novel variants in *CSMD1* in eight individuals with MCD inherited as compound heterozygous or homozygous from seven families (Table 6). Two are intronic variants and eight are missense variants. *CSMD1* is a long gene, comprised of 70 exons encoding a 3,564 amino acid (388 kD) type-I transmembrane protein (Figure 4.1). CSMD1 is made up of several repetitive elements encoding alternating CUB and sushi domains (Figure 4.1). Identified missense variants localize to CUB and Sushi domains throughout the CSMD1 protein. While no nonsense variants were identified in our cohort, CSMD1 demonstrates high probability of loss-of-function intolerance (pLI) in gnomAD.

Missense variants that lie within CSMD1 CUB domains (E138K, R213L, R621Q, and L1332V) cluster more N-terminally in the protein whereas those that lie within Sushi domains are more evenly distributed (V187I, S188N, Q1782K, P2262A, D2296N, G2980S). All individuals with biallelic *CSMD1* variants have moderate-to-severe intellectual disability (ID) and/or global developmental delay (GDD). All individuals carrying at least one N-terminal variant present with microcephaly and/or seizures (Probands 1-2 and 6-8). Two individuals have polymicrogyria, one displaying corpus callosum agenesis (Proband 1) and the other with cerebellar agenesis (Proband 8). Of note, Proband 3, who is compound heterozygous for two C-terminal variants (c.7285+2T>C; D2296N), showed an unremarkable MRI without seizures and mild ID, suggesting N-terminal variants correlate with increased severity of MCD. However, there is no apparent difference between phenotypic correlation between homozygous versus compound heterozygous genotypes. Shared facial dysmorphisms among multiple individuals include retrognathia, micrognathia, and strabismus. Detailed clinical summaries for each proband are provided below.

Proband 1

Proband 1 is from Alpena, Michigan, USA and of European ancestry. She was seven years old at last examination. Proband 1 was carried to full-term (39 weeks at birth) with pregnancy characterized by intrauterine growth delay. She had neonatal hypotonia and microcephaly. She presented with upper respiratory infections, GERD, drooling, focal epilepsy, G-tube-dependence, tracheostomy-dependence, dysphagia, and repeated ear infections. Bilateral club feet, amblyopia of the left eye, myopia bilateral, eustachian tube dysfunction, and esotropia were noted. She exhibited GDD, moderate-to-severe ID, dysphagia, and focal epilepsy. She had onic clonic, myoclonic-tonic and myoclonic seizures, with minimal response to varying dosages of oxycarbazepine, levetiracetam, and CBD oil. MRI at 12 months indicated diffusely dysplastic cerebral hemispheres, polymicrogyria, and corpus callosum agenesis.

Upon exome sequencing, compound heterozygous variants g.4277477C>T, p.E138K and g.3855604C>A, p.R213L were identified in *CSMD1*. Of note, compound heterozygous variants were also identified in *ABCA1*. Biallelic variants in *ABCA1* cause autosomal recessive disorder called Tangier disease that causes reduced levels of plasma high density lipoproteins (HDL).

Proband 2

Proband 2 is from Monza, Italy and of European ancestry. Prenatal manifestations of dextrocardia and polyhydramnios were noted. Proband 2 was born at 40 weeks of gestation. He had patent foramen ovale at birth. As a newborn, he experienced feeding difficulties and doliocephaly. He has mild ID and displays oppositive behavior. He had anorectal malformation and mandibular hypoplasia, both requiring surgical correction. He has malar hypopplasia, micrognathia (after mandibular hypoplasia surgery), anteverted ears, downslanting of palpebral fissures, deep-set eyes, high nasal bridge, nasal voice, thin lips, and hypoglossia. Age of onset of seizures at 5 years old. Proband 2 was responsive to levetiracetam. New seizures occurred after 18 months without therapy at age 11 years requiring reintroduction of levetiracetam. Exome sequencing identified compound heterozygous *CSMD1* variants c.1862G>A, p.R621Q; c.3994C>G, p.L1332V.

Proband 3

Proband 3 is from Phoenix, Arizona, USA and of European ancestry. Prenatal polyhydramnios and preterm labor were noted. Proband 3 was born at 37 weeks of gestation. Episode of hypoglycemia occurred at birth and possible respiratory arrest requiring NICU stay for 24 hours. He has mild ID, hypotonia, and ADHD. Generally, he has nondysmorphic head features. Brain MRI at 12 months of age indicated mild bilateral periventricular white matter injury, likely remote, with normal brain development for age. MRA was normal. Hypertelorism was noted, but father also has slight hypertelorism. He has strabismus, non-paralytic estropia, eczema, nevus flammeus over the glabella, retractile testes, and inguinal hernia. MRI was unremarkable. Proband 3 sat independently at 10 months, walked unassisted at 2.5 years, and said first words at 1 years of age. He could speak in sentences at 3 years of age. Diffuse joint hypermobility and spiral fracture of his right leg were also noted. Genome sequencing identified compound heterozygous *CSMD1* variants c.7285+2T>C; c.6886G>A, p.D2296N.

Probands 4 and 5

Probands 4 and 5 are siblings from Quetta, Pakistan of Pashtoon ancestry. Both siblings were carried to full-term. Proband 4 was 16 years of age upon last examination. Proband 5 is the younger brother of Proband 4, 8 years of age upon last examination. Both present with severe-to-profound ID, GDD, hypotonia, and restricted speech. Proband 5 can walk while Proband 4 cannot. Exome sequencing of both siblings identified compound heterozygous *CSMD1* missense variants c.8938 G>A, p.G2980S and c.6784C>G, p.P2262A.

Proband 6

Proband 6 is from the United Kingdom and of European ancestry. Proband 6 presents with microcephaly, GDD, and hypotonia. Compound heterozygous *CSMD1* variants c.1098-4T>C and c.861G>T, p.K287N were identified by exome sequencing. Additionally, a heterozygous variant was identified in *PRPS1*. Variants in *PRPS1* cause Arts syndrome, which is characterized by sensorineural hearing impairment, early-onset hypotonia, ID, ataxia, developmental motor delay, and increased infections.

Proband 7

Proband 7 is from Utrecht, the Netherlands and is of Afghani ancestry. Pregnancy was uneventful and proband 7 was born at 37 weeks of gestation. He has severe ID, microcephaly, hypotonia, mildly protruding ears and deep-set eyes, ptosis, broad upper incisors, sandal gap bilateral feet, and mild hypertrichosis on his back. Proband 7 began sitting at 6 months, walking at 1 year, and first words at 2 years of age. Trio exome sequencing identified homozygous *CSMD1* missense variant c.559G>A, p.V187I in Proband 7.

Proband 8

Proband 8 is from Italy and of European ancestry. She was born at gestation week 36 (birth weight, 2150 kg; birth length, 44 cm; birth head circumference, 33 cm). Risk of miscarriage occurred in the first trimester of pregnancy. Mother had gestational diabetes. Proband 8 exhibited failure to thrive. She had frontal bossing, retrognathia, strabismus, and club feet. She began walking at 3 years of age. Seizures were noted at 21 months of age. Proband 8 responded to valproate and remained stable over time. Proband 8 has been seizure-free for last 1.5 years with the following medications: topamax, Depakin, and Tolep. Fronto-temporal anomalies were reported on EEG. She has generalized hypotonia, and MRI identified cerebellar agenesis and polymicrogyria. Compound heterozygous missense variants c.5344C>A, p.Q1782K and c.563G>A, p.S188N in *CSMD1* in Proband 8 were identified by exome sequencing.

4.2.2 *CSMD1$^{fs/fs}$* cortical organoids exhibit features of premature differentiation with disorganized neural rosette morphology

To investigate the function of *CSMD1* in human cortical development, 3D forebrain-fated organoid models were generated by neural differentiation from previously generated *CSMD1*$^{\text{fs/fs}}$ and *CSMD1*$^{+/+}$ control hESCs (Baum *et al.*, 2020). *CSMD1*$^{\text{fs/fs}}$ organoids were consistently smaller throughout development compared to *CSMD1*$^{+/+}$ controls across three independent differentiations (p = 0.0018, two-tailed t test) (Figure 4.2A). However, growth rates were similar with exceptions of significant elevation differences attributed to slower growth rate from day 21 neural differentiation (ND) to day 28ND, followed by a faster growth from day 28ND to day 35ND, in *CSMD1*$^{\text{fs/fs}}$ compared to controls (p = 0.0013, two-tailed t test) (Figure 4.2A).

While consideration of whole organoid size is helpful for pinpointing affected timepoints during cortical development, assessment of individual neural rosette (NR) structures allows us to hone the cellular mechanisms of pathogenesis. A slight reduction in the number of cells per NR at day 28ND was noted, however this finding was not significant (Figure 4.2B). *CSMD1*$^{\text{fs/fs}}$ NRs showed a significant increase in proliferating NPCs (KI67) (p = <0.0001, two-tailed t test) at day 28ND (Figure 4.2C and 4.2D). Further, while *CSMD1*$^{+/+}$ NRs show a uniform distribution of proliferating cells across the pseudostratified neuroepithelium, cells in affected NRs appear to lose their polarity and are more condensed around the central lumen (Figure 4.2C). Immunostaining of adherens junctions (N-cadherin) further highlights altered spatial organization at NR structures in *CSMD1*$^{fs/fs}$ organoids (Figures 4.2E and 4.2F). Cells appear less condensed radially with wider somas. This disrupted spatial disorganization is either directly or indirectly causing premature differentiation of deep layer neurons, as measured by an increase in the fraction of TBR1-expressing cells per NR (p = <0.0056, two-tailed t test) at day 28ND, suggestive of premature differentiation (Figures 4.2G and 4.2H). These findings implicate disrupted neural proliferation, differentiation, and spatial organization in the pathology of *CSMD1*-associated MCD.

4.3 DISCUSSION

As CSMD1 is gaining traction as an important regulator of the complement pathway influencing synaptic activity (Baum *et al.*, 2020; Coulthard *et al.*, 2018; Schafer *et al.*, 2012), our findings expand its functions to include neural development. For the first time, this study implicates biallelic variants in *CSMD1* in the genetic etiology of MCD. Individuals in this study present with overlapping features of ID and/or GDD, polymicrogyria, microcephaly, epilepsy, and facial dysmorphisms.

A variety of pathogenic mechanisms have been implicated in MCD, including abnormal proliferation and/or altered apoptosis, which often causes microcephaly, megalencephaly, and dysplastic malformations, or altered neuronal migration that can lead to malformations such as lissencephaly and polymicrogyria (Guerrini and Parrini, 2010). Here, findings from *CSMD1*$^{fs/fs}$ cerebral organoid models reveal several hallmark features of MCD, including over-proliferation of neural progenitors, premature differentiation, and altered NR spatial and cellular morphology. The cytoskeleton works closely with cell adhesion receptor systems to ensure proper cell polarity for successful neuronal locomotion, including tangential and radial migrations (Govek et al., 2011). Observed alterations to the compact, pseudostratified organization of *CSMD1*$^{fs/fs}$ neuroepithelium may reflect disturbances to cytoskeletal projections and the apicobasal polarity of NPCs, which ultimately can impair migration kinetics during NPC cell division and/or the guidance of migrating neurons to their final destinations in the cortical plate that may lead to premature differentiation. Premature differentiation is a known developmental mechanism of primary microcephaly due to diminishing the number of proliferative cells leading to fewer neurons and may be the primary cause of microcephaly in individuals with biallelic pathogenic *CSMD1* variants.

While our data from cortical organoids offer insight into early-stage developmental defects, we did not assess effects on later-stage development. Future work will be necessary to directly evaluate deficits in neuronal migration and assess consequences on cortical lamination that may underlie CSMD1-related pathology. Further investigation is also required to determine the genetic mechanism of pathogenic *CSMD1* variants, which we propose is loss-of-function. Our finding of over-proliferation in *CSMD1*$^{fs/fs}$ neuroepithelium supports the model that CSMD1 is an inhibitor of complement pathway in early-stage NPCs, as described for neurons. Follow-up functional investigation to address test this model and to determine how complement deposition is mechanistically linked to changes in radial organization and morphology will be important for elucidating *CSMD1*-associated neuropathology.

Taken together, we discover novel human genetics findings that implicate *CSMD1* as the genetic basis for human neurodevelopmental disorders. Our functional findings support the hypothesis that CSMD1 is an important regulator of early neural proliferation and differentiation in corticogenesis.

4.4 MATERIALS AND METHODS

Subject accrual

All subjects or parents/guardians provided informed consent and were enrolled in institutional review board-approved research studies. Followed procedures were in accordance with the ethical standards of the respective institution's committee on human research and were in keeping with international standards. Probands 2-8 were identified through GeneMatcher (Sobreira et al., 2015).

Exome-based sequencing

Exome libraries from genomic DNA of all probands were prepared and captured with the Agilent SureSelectXT Human All Exon 50Mb Kit. Exome libraries were sequenced on an Illumina HiSeq instrument as described previously (Srivastava et al., 2018) at the University of Washington Mendelian Sequencing Center.

Variant calling and filtering

Reads were aligned to the hg38 reference genome (GRCh38.p13) using Burrows-Wheeler Aligner (BWA). Variant calling of single nucleotide variants (SNVs) and copy number variants (CNVs) was performed using GATK. The data were filtered and annotated from the canonical *CSMD1* transcript (ENST00000635120.2) using in-house bioinformatics software. Variants were also filtered against public databases including the 1000 Genomes Project phase 311, Genome Aggregate Database (gnomAD), National Heart, Lung, and Blood Institute Exome Sequencing Project Exome Variant Server (ESP6500SI-V2). Those with a minor allele frequency >3.3% were excluded. Additionally, variants flagged as low quality or putative false positives (Phred quality score 14; 15 < 20, low quality by depth <20) were excluded from the analysis. Variants in genes known to be associated with MCD were selected and prioritized based on predicted pathogenicity. Reported variants were confirmed by Sanger sequencing of *CSMD1* (NM_033225.6) for each individual and respective family members who submitted samples.

Human ESC culture

Human ESCs were cultured using feeder free conditions on Matrigel (Corning with mTeSR-1 (STEMCELL Techonologies). H1 (*CSMD1*$^{+/+}$, 46XY, WA01, WiCell) and H1 (*CSMD1*$^{fs/fs}$, 46XY, WA01, WiCell) ESC lines used in this manuscript were obtained from, and validated by, the Stevens lab at the BROAD Institute of MIT and Harvard where cell line quality was assessed by CNV, karyotype, and morphology analyses. No further validation of ESCs was performed in our lab.

Cerebral organoid generation

Telencephalic cerebral organoids were generated based on previously published protocols (Lancaster et al., 2013), with few modifications to start with low cell density to generate smaller and more consistent embryoid bodies (EBs). Briefly, human and chimpanzee iPSCs were passaged into 96-well V-shaped bottom ultra-low attachment cell culture plates (PrimeSurface® 3D culture, MS-9096VZ) to achieve a starting cell density of 600-1,000 cells per well in 30 μl of mTesRTM1 with 1 nM ROCK inhibitor. After 36 hours, 150 μl of N-2/SMAD inhibition media (cocktail of 1X N-2 supplement (Invitrogen 17502048), 2 μM A-83-01 inhibitor (Tocris Bioscience 2939), and 1 mM dorsomorphin (Tocris Bioscience 309350) in DMEM-F12 (Gibco 11330032)) was added for neural induction. On day 7, EBs were transferred to Matrigel-coated plates to enrich for neural rosettes at a density of 20-30 EBs per well of a 6-well plate, and media was changed to neural differentiation media (0.5X N-2 supplement, 0.5X B-27 supplement (Invitrogen 17504044) with 20 pg/μl bFGF and 1mM dorsomorphin inhibitor in DMEM/F-12). For organoid differentiation EBs were outlined on day 14 using a pipet tip and uplifted carefully with a cell scraper to minimize organoid fusion and tissue ripping. Media was changed once more to N-2/B-27 with bFGF only and plates with uplifted organoids were placed on a shaker in the incubator set at a rotation speed of 90. On day 14, media was changed once more to N-2/B-27 with bFGF only. Media changes were performed every 48 hours.

Immunohistochemistry

Human cortical organoids were fixed in 4% PFA for 24 hours at 4°C, cryoprotected in 15% and 30% sucrose in 1x DPBS for 24 hours at 4°C, then embedded in OCT with quick freezing in -50°C 2-methylbutane, followed by cryosectioning at 16 μm. Antigen retrieval was performed on sections by incubation in heated 10 mM sodium citrate solution (95-100°C) for 20 minutes prior to immunostaining. Sections were then incubated for 1 hour with blocking buffer (5% NDS (Jackson ImmunoResearch) 0.1% Triton X-100, 5% BSA) at room temperature, then overnight with primary antibodies diluted in blocking buffer at 4°C, and for 1-2 hours in secondary dilution at room temperature. Washes performed in PBS. For nuclear staining, samples were incubated at room temperature for 10 minutes in Hoescht (1:1000 dilution in PBS) prior to final washes. Primary antibodies used: mouse anti-PAX6 (1:250, Abcam, MA-109), rabbit anti-KI67 (1:200, Abcam ab16667), rat anti-PH3 (1:250, Abcam ab10543), mouse anti-N-Cadherin (1:200, BD Biosciences 610920), rabbit anti-ZO-1 (1:200, Invitrogen 61-7300), and rabbit anti-TBR1 (1:200, Abcam ab31940). AlexaFluor-conjugated secondaries used: donkey anti-mouse 647 (1:400, Invitrogen A31571), donkey anti-rat 555 (1:400, Invitrogen A48270), and donkey anti-rabbit 488 (1:400, Invitrogen, A21206).

Glass covers were mounted onto all slides with Prolong Gold (Molecular Probes S36972) and incubated for 24 hours at room temperature prior to imaging. Imaging was performed with a Nikon A1ss inverted confocal microscope using NIS-Elements Advanced Research software. Image analysis was performed using Fiji (ImageJ) software. Statistical significance of image quantifications was tested using a two-tailed unpaired *t*-test, and data was plotted as mean ±SEM using GraphPad Prism (v9.3.1).

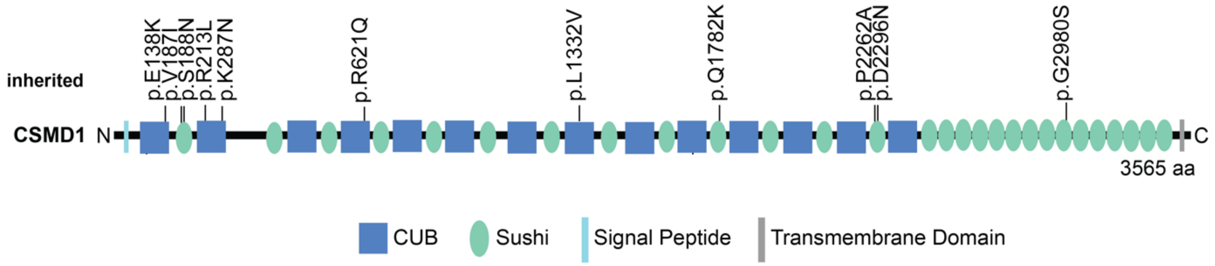


Figure 4.1. Protein map of CSMD1 clinical variants.
Missense variants identified by exome and genome sequencing map to the CUB (blue square) and sushi (green oval) domains of CSMD1. Inherited variants (compound heterozygous or homozygous) are denoted on top.

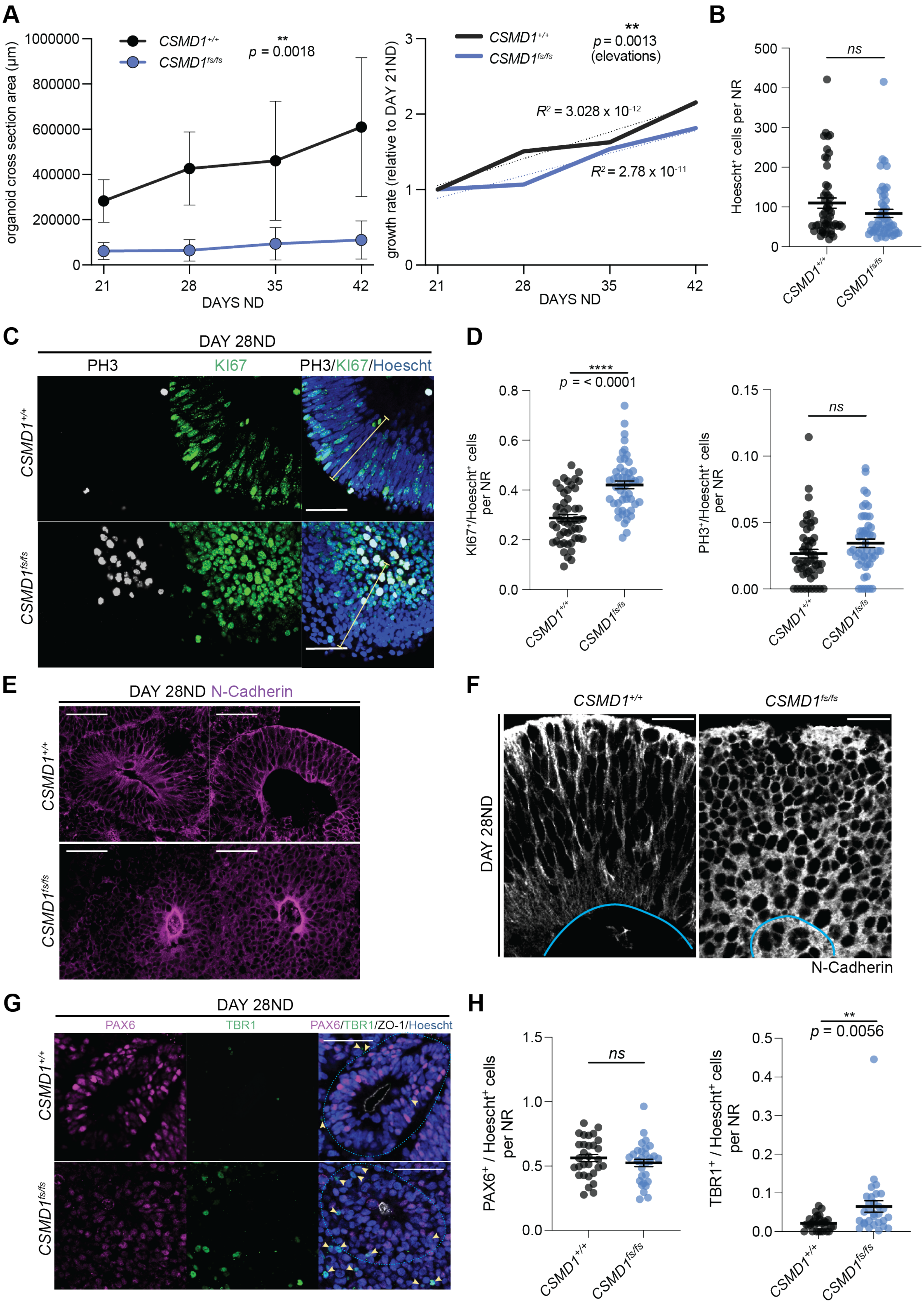

A
organoid cross section area (µm)
1000000
800000
600000
400000
200000
0
CSMD1+/+
CSMD1fs/fs
**
p = 0.0018
21
28
35
42
DAYS ND
growth rate (relative to DAY 21ND)
3
2
1
0
CSMD1+/+
CSMD1fs/fs
**
p = 0.0013
(elevations)
R² = 3.028 x 10⁻¹²
R² = 2.78 x 10⁻¹¹
21
28
35
42
DAYS ND
B
Hoescht+ cells per NR
500
400
300
200
100
0
ns
CSMD1+/+
CSMD1fs/fs
C
DAY 28ND
PH3
KI67
PH3/KI67/Hoescht
CSMD1+/+
CSMD1fs/fs
D
KI67+/Hoescht+ cells per NR
0.8
0.6
0.4
0.2
0.0
****
p = < 0.0001
CSMD1+/+
CSMD1fs/fs
PH3+/Hoescht+ cells per NR
0.15
0.10
0.05
0.00
ns
CSMD1+/+
CSMD1fs/fs
E
DAY 28ND N-Cadherin
CSMD1+/+
CSMD1fs/fs
F
CSMD1+/+
CSMD1fs/fs
DAY 28ND
N-Cadherin
G
DAY 28ND
PAX6
TBR1
PAX6/TBR1/ZO-1/Hoescht
CSMD1+/+
CSMD1fs/fs
H
PAX6+ / Hoescht+ cells per NR
1.5
1.0
0.5
0.0
ns
CSMD1+/+
CSMD1fs/fs
TBR1+ / Hoescht+ cells per NR
0.5
0.4
0.3
0.2
0.1
0.0
**
p = 0.0056
CSMD1+/+
CSMD1fs/fs

Figure 4.2. Characterization of *CSMD1*$^{fs/fs}$ cerebral organoids.

(A) Organoid growth analysis of cross-section area per genotype (*CSMD1*$^{fs/fs}$, *CSMD1*$^{+/+}$) shown as weekly measurements from day 21 to day 42, mean±standard deviation (left). Growth rate determined by mean relative to day 21 from day 21 to day 42 (right). (B) Number of nuclei (Hoescht) per NR in day 28 organoids. (C) Immunostaining of proliferation (PH3, KI67) in day 28 NR per genotype with quantifications (D). *n* = 50 NRs per genotype across three differentiations. (E) Immunostaining of NPCs (PAX6) and early born neurons (TBR1) with quantifications (F). *n* = 30 NRs per genotype across three differentiations. Scale bars, 50 μm. Data shown as mean±SEM. Significance, two-tailed *t* test. N-cadherin immunostaining showing organization of entire NR (G) and segment (H). Scale bars, 50 and 25 μm, respectively.

Table 6. Summary of *CSMD1* clinical findings.
ID, intellectual disability; GDD, global developmental delay; N/A, not available.

Proband	1	2	3	4/5	6	7	8
Country	**USA**	**Italy**	**USA**	**Pakistan (2 siblings)**	**UK**	**Netherlands Afghanistan**	**Italy**
***CSMD1* variants (NM_033225.6)**	g.4277477C>T, p.E138K; g.3855604C>A, p.R213L	c.1862G>A, p.R621Q; c.3994C>G, p.L1332V	c.7285+2T>C; c.6886G>A, p.D2296N	c.8938 G>A, p.G2980S; c.6784C>G, p.P2262A	c.1098-4T>C; c.861G>T, p.K287N	c.559G>A, p.V187I; c.559G>A, p.V187I	c.5344C>A, p.Q1782K; c.563G>A, p.S188N
Brain	Microcephaly; polymicrogyria, CC agenesis	N/A	Unremarkable MRI	N/A	Microcephaly	Microcephaly	Polymicrogyria, cerebellum agenesis
Seizures	**Yes**	**Yes**	No	No	No	No	**Yes**
ID/GDD	Severe ID, GDD, hypotonia	Mild ID	Mild ID	Severe ID, GDD, few words, hypotonia	GDD, hypotonia	Severe ID	Moderate GDD
Dysmorphism	Severe retrognathia, esotropia	Micrognathia, down slanting palpebral fissures, deep-set eyes	strabismus, non-paralytic esotropia	N/A	N/A	ptosis, eyes deep-set	Retrognathia, strabismus (club feet)
2nd genetic hit	Compound heterozygous *ABCA1* Tangier disease				Heterozygous *PRPS1* Sensorineural hearing impairment		

REFERENCES

Barkovich, A.J., Guerrini, R., Kuzniecky, R.I., Jackson, G.D., and Dobyns, W.B. (2012). A developmental and genetic classification for malformations of cortical development: update 2012. Brain *135*, 1348-1369. 10.1093/brain/aws019.

Baum, M.L., Wilton, D.K., Muthukumar, A., Fox, R.G., Carey, A., Crotty, W., Scott-Hewitt, N., Bien, E., Sabatini, D.A., and Lanser, T. (2020). CUB and Sushi Multiple Domains 1 (CSMD1) opposes the complement cascade in neural tissues. bioRxiv.

Consortium, S.P.G.-W.A.S.G. (2011). Genome-wide association study identifies five new schizophrenia loci. Nat Genet *43*, 969-976. 10.1038/ng.940.

Coulthard, L.G., Hawksworth, O.A., and Woodruff, T.M. (2018). Complement: The Emerging Architect of the Developing Brain. Trends Neurosci *41*, 373-384. 10.1016/j.tins.2018.03.009.

Escudero-Esparza, A., Kalchishkova, N., Kurbasic, E., Jiang, W.G., and Blom, A.M. (2013). The novel complement inhibitor human CUB and Sushi multiple domains 1 (CSMD1) protein promotes factor I-mediated degradation of C4b and C3b and inhibits the membrane attack complex assembly. FASEB J *27*, 5083-5093. 10.1096/fj.13-230706.

Govek, E.E., Hatten, M.E., and Van Aelst, L. (2011). The role of Rho GTPase proteins in CNS neuronal migration. Dev Neurobiol *71*, 528-553. 10.1002/dneu.20850.

Guerrini, R., and Dobyns, W.B. (2014). Malformations of cortical development: clinical features and genetic causes. Lancet Neurol *13*, 710-726. 10.1016/S1474-4422(14)70040-7.

Guerrini, R., and Parrini, E. (2010). Neuronal migration disorders. Neurobiol Dis *38*, 154-166. 10.1016/j.nbd.2009.02.008.

Hong, S., Beja-Glasser, V.F., Nfonoyim, B.M., Frouin, A., Li, S., Ramakrishnan, S., Merry, K.M., Shi, Q., Rosenthal, A., Barres, B.A., et al. (2016). Complement and microglia mediate early synapse loss in Alzheimer mouse models. Science *352*, 712-716. 10.1126/science.aad8373.

Kraus, D.M., Elliott, G.S., Chute, H., Horan, T., Pfenninger, K.H., Sanford, S.D., Foster, S., Scully, S., Welcher, A.A., and Holers, V.M. (2006). CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. J Immunol *176*, 4419-4430. 10.4049/jimmunol.176.7.4419.

Lau, W.L., and Scholnick, S.B. (2003). Identification of two new members of the CSMD gene family. Genomics *82*, 412-415. 10.1016/s0888-7543(03)00149-6.

Lee, A.S., Rusch, J., Lima, A.C., Usmani, A., Huang, N., Lepamets, M., Vigh-Conrad, K.A., Worthington, R.E., Mägi, R., Wu, X., et al. (2019). Rare mutations in the complement regulatory gene CSMD1 are associated with male and female infertility. Nat Commun *10*, 4626. 10.1038/s41467-019-12522-w.

Ruiz-Martínez, J., Azcona, L.J., Bergareche, A., Martí-Massó, J.F., and Paisán-Ruiz, C. (2017). Whole-exome sequencing associates novel *CSMD1* gene mutations with familial Parkinson disease. Neurol Genet *3*, e177. 10.1212/NXG.0000000000000177.

Schafer, D.P., Lehrman, E.K., Kautzman, A.G., Koyama, R., Mardinly, A.R., Yamasaki, R., Ransohoff, R.M., Greenberg, M.E., Barres, B.A., and Stevens, B. (2012). Microglia sculpt postnatal neural circuits in an activity and complement-dependent manner. Neuron *74*, 691-705. 10.1016/j.neuron.2012.03.026.

Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al. (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177-183. 10.1038/nature16549.

Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat *36*, 928-930. 10.1002/humu.22844.

Srivastava, A., Srivastava, K.R., Hebbar, M., Galada, C., Kadavigrere, R., Su, F., Cao, X., Chinnaiyan, A.M., Girisha, K.M., Shukla, A., and Bielas, S.L. (2018). Genetic diversity of NDUFV1-dependent mitochondrial complex I deficiency. Eur J Hum Genet *26*, 1582-1587. 10.1038/s41431-018-0209-0.

Chapter 5 [2]

Mechanisms of mRNA Processing Defects in Inherited *THOC6* Intellectual Disability Syndrome

5.1 INTRODUCTION

Intellectual disability (ID) is a clinical feature of neurodevelopmental disorders characterized by limitations in cognitive ability and adaptive behavior (Schalock et al., 2021). In recent years, genetic etiologies of syndromic ID have become increasingly heterogeneous due to broad use of exome-based genetic testing (Anazi et al., 2017; Gieldon et al., 2018; Retterer et al., 2016; Vasudevan and Suri, 2017; Yang et al., 2014). Monogenetic causes account for a substantial portion of syndromic ID, with many following an autosomal-recessive mode of inheritance (Kochinke et al., 2016). ID-associated genes are enriched in diverse biological networks including metabolism, nervous system development, RNA metabolism, transcription, sonic hedgehog signaling, glutamate signaling, peroxisomes, glycosylation, and cilia (Kochinke *et al.*, 2016). THOC6 Intellectual Disability Syndrome (TIDS; OMIM# 613680) is one such recessive disorder, attributed to biallelic pathogenic variants in *THOC6* (Accogli et al., 2018; Amos et al.,

[2] This chapter represents a submitted manuscript. Werren E.A. et al., Mechanisms of mRNA processing defects in inherited THOC6 intellectual disability syndrome.

2017; Anazi et al., 2016; Boycott et al., 2010; Casey et al., 2016; Gupta et al., 2020; Hassanvand Amouzadeh et al., 2020; Kiraz et al., 2022; Mattioli et al., 2018; Ruaud et al., 2022; Zhang et al., 2020). Individuals with TIDS exhibit moderate to severe syndromic ID with microcephaly and multi-organ involvement. Genetic testing is necessary for diagnosis of TIDS, as clinical features overlap other inherited neurodevelopmental disorders (Lemire et al., 2020).

THOC6 is a subunit of the six member THO (suppressors of the transcriptional defects of hpr1 delta by overexpression) complex (Jimeno and Aguilera, 2010). The THO complex is a core component of the transcription/export (TREX) complex. Prior to translation, mRNA transcripts progress through a series of coordinated steps that include mRNA 5' capping, splicing, and 3' end processing to create a messenger ribonucleoprotein complex (mRNP) capable of translocating through the nuclear pore complex into the cytoplasm (Heath et al., 2016; Köhler and Hurt, 2007; Xie and Ren, 2019). TREX is necessary for proper mRNA processing and nuclear export required for gene expression, with a more appreciated role in export based on current literature (Chi et al., 2013; Heath *et al.*, 2016; Luna et al., 2012; Masuda et al., 2005; Peña et al., 2012; Rondón et al., 2010; Wickramasinghe and Laskey, 2015). TREX is recruited to the maturing mRNP during transcription (Heath *et al.*, 2016; Viphakone et al., 2019), in coordination with pre-mRNA binding proteins at the cap-binding complex (CBC), exon-junction complex (EJC), and the polyadenylated 3' end (Cheng et al., 2006; Gromadzka et al., 2016; Merz et al., 2007; Shi et al., 2017). Despite a well-studied role in mRNA export, the precise timing of initial TREX recruitment to maturing mRNPs in mammalian cells is unclear, but mounting evidence suggests mammalian TREX associates to the mRNPs during splicing (Chi *et al.*, 2013; Luo et al., 2001; Masuda *et al.*, 2005). Ultimately, TREX promotes the loading of licensing export factors required for nuclear pore docking and export (Hautbergue et al., 2008; Köhler and Hurt, 2007; Strässer and Hurt, 2001; Taniguchi and Ohno, 2008).

While the general role of TREX in mRNA processing and export is thought to be conserved, there are notable species differences in TREX composition and function that mirror the evolutionary complexity of mRNA processing requirements. In yeast, the TREX dimer is composed of two five-subunit THO monomers, with THOC6 being the notable exception. Yeast TREX monomers dimerize via the coiled coil domains of Thp2 and Mft1, the yeast orthologs of THOC5 and THOC7 (Pühringer et al., 2020). In humans, the THO monomers are composed of six subunits, including THOC6, that form a tetramer with THOC6 serving as the central tethering component (Pühringer *et al.*, 2020). The increased size and molecular complexity of the mammalian TREX tetramer correlates with increased mRNA processing demands that have evolved in organisms with higher transcriptome complexity and mRNP composition, namely expression of long genes with high levels of complex splicing patterns (Singh et al., 2015). For example, introns comprise ~24% of mammalian genomes (Venter et al., 2001). On the level of gene organization, introns are longer and are present in >95% of all human genes (Chen et al., 2006; Lander et al., 2001; Nagasaki et al., 2005). By contrast, introns constitute only 5% of yeast genes, are short relative to the gene length, and mostly limited to one per gene (Chen *et al.*, 2006; Chervitz et al., 1999; Juneau et al., 2006; Nagasaki *et al.*, 2005). Thus, alternative splicing is rare in yeast, but plays a major role in gene expression in mammals, especially humans. Functional differences between a TREX dimer and tetramer may correlate to differences in complex size and number of molecules that can be simultaneously accrued to coordinate mRNP processing through export (Pühringer *et al.*, 2020).

Yeast TREX dimers are configured to bring two bound Sub2 helicases (yeast ortholog to the metazoan DDX39B/UAP56 RNA helicase which couples ATP hydrolysis to mRNA release) in proximity. This formation enables Yra1 (yeast ortholog of ALYREF), a mRNP processing/export factor to bridge the THO monomers by binding the N- and C-terminal of aligned Sub2 components. Components of the mammalian TREX tetramer exhibit the ability to interact with a diversity of mRNP processing and export factors. The TREX tetramer also binds UAP56 molecules, however the tetramer organization means the putative loading sites for ALYREF are doubled, the potential of affording increased affinity of processing adaptors to their target mRNPs (Pühringer *et al.*, 2020). The tetramer structure is predicted to enhance physical support for processing longer mammalian transcripts, allowing TREX to serve as a mRNA chaperone to prevent formation of DNA-RNA hybrid or R-loop structures that can promote genome instability (Luna et al., 2019; Pérez-Calero et al., 2020). In line with the mammalian mRNP processing complexity, the tetramer recruits a portfolio of functionally diverse auxiliary factors, predicted to enhance coordination of mRNA processing steps. Several TREX adaptor proteins in metazoans that have no yeast ortholog include, DDX39A, CHTOP, UIF, LUZP4, POLDIP3, ZC3H11A, ERH, ZC3H18, SRRT, and NCBP3, representing proteins that participate in each step of mRNP processing and export and hinting at a level of coordination required as the complexity of mRNP processing evolved with the emergence of longer transcripts with elevated splicing (Dufu et al., 2010; Heath *et al.*, 2016).

While functions of THO within TREX have primarily focused on mRNP export (Chi *et al.*, 2013; Guria et al., 2011; Maeder et al., 2018), disruption of the tetramer is predicted to disrupt coordination of mRNP processing steps that precede export. TREX-associated functions of CHTOP and ALYREF exemplify this molecular biology. CHTOP and ALYREF bind to THO-bound UAP56 in a mutually exclusive manner (Chang et al., 2013). ALYREF exhibits preferential binding to the 5’ end and 5’ splice sites of mRNA to regulate splicing fidelity, whereas CHTOP preferentially binds the 3’ UTR to regulate polyadenylation site choice along with THOC5 (Viphakone *et al.*, 2019). Lastly, ALYREF, CHTOP, and THOC5 interact with and load the global NXF1-NXT1 licensing heterodimer onto mRNA for export (Hautbergue *et al.*, 2008; Köhler and Hurt, 2007; Strässer and Hurt, 2001; Taniguchi and Ohno, 2008). This organization provides structural logic for how the molecular component required for the many aspects of mRNP processing can be coordinated, but the impact of disrupting a single step on progression of the entire process has not been evaluated.

The conservation of THO dimer functions in mammalian cells is an open question. THOC1, THOC3, THOC5, and THOC7 exhibit high probability of loss-of-function intolerance (pLI) in gnomAD and have not been identified as the genetic basis of developmental disorders, suggesting conserved THO components are likely embryonic lethal. *THOC2* is the genetic basis of an X-linked neurodevelopmental disorder (Kumar et al., 2015). Depletion of THO components THOC1-THOC5 and THOC7 lead to strong nuclear export defects (Chi *et al.*, 2013). This leads to the speculation that dimers retain mRNP functions in mammals, and that THOC6-dependent tetramer functions enhance the efficiency and coordination of these activities. This would also explain the tissue sensitivity in TIDS, where development is disrupted in tissues that disproportionately express long genes. In addition, neural-expressed genes display many isoforms which enhance the mRNP processing burden.

THOC6 evolved as a scaffolding protein to create a larger TREX complex, that has implication for mRNA processing coordination in metazoans relative to yeast. Despite extensive research on TREX, there is a lack of research in neural cells, cells that undergo extensive changes in mRNA processing during differentiation, particularly alternative splicing with increased intron retention (Mauger et al., 2016). Given this, there may be functions of THO within a tetramer conformation mediated by THOC6 that have not been the main focus of study. Here, we investigated this by utilizing a series of models with pathogenic alleles that disrupt THOC6 to assess essential functions of tetramers in features of mRNA processing in neural development. First, we contribute to the *THOC6* pathogenic allele series and TIDS clinical phenotypic spectrum, extending the total number of reported *THOC6* variants in TIDS to 20 and the total number of reported affected individuals to 34. Second, we generated a *Thoc6* mouse model and human induced pluripotent stem cell (iPSC)-derived cell culture models to investigate shared pathogenic mechanisms of mammalian *THOC6.* Given the high penetrance of microcephaly in TIDS, we focus on mRNA processing/export changes and accompanying phenotypes that occur during cortical development by analyzing primary mouse and dorsal forebrain fated human organoids. We propose a model of unproductive selective mRNA processing/export from partial TREX disruption due to loss-of-function *THOC6* alleles leading to dysregulation of proliferation and differentiation. Our findings reveal a broader supportive role across mRNA processing within the context of THOC6 variants than has previously been attributed to THO.

5.2 RESULTS

5.2.1 Biallelic missense and nonsense *THOC6* variants are the genetic basis of TIDS

While initially detected in Hutterite populations (Boycott *et al.*, 2010), a growing number of pathogenic biallelic *THOC6* variants are being discovered across the globe in individuals of diverse ancestry (Accogli *et al.*, 2018; Amos *et al.*, 2017; Anazi *et al.*, 2016; Beaulieu et al., 2013; Casey *et al.*, 2016; Gupta *et al.*, 2020; Hassanvand Amouzadeh *et al.*, 2020; Kiraz *et al.*, 2022; Mattioli *et al.*, 2018; Ruaud *et al.*, 2022; Zhang *et al.*, 2020). As part of an ongoing effort to determine the genetic etiologies of syndromic ID, we discovered nine *THOC6* variants by exome-based genetic testing (Figure 1A). We confirmed six recurrent variants at W100, G190, V234, and G275 amino acids and identified three novel alleles (p.Q47*, p.E188K, and p.R247Q). The clinical phenotype associated with *THOC6* variants affirm penetrance of the core clinical features of TIDS, namely global developmental delay, moderate to severe ID and facial dysmorphisms (Figure 1B). Variable expressivity of cardiac and renal malformations, structural brain abnormalities with and without seizures, urogenital defects, recurrent infections, and feeding complications were also noted, clinical features that highlight the multiorgan involvement of this developmental syndrome (Figure 1C). Detailed clinical summaries for all individuals are provided in Table S1 and S2.

The novel *THOC6* variants are representative of previously described nonsense and missense variants that contribute equally to the severity of TIDS phenotypes. We describe a nonsense *THOC6* c.139C>T, (p.Q47*) variant in exon 2 of proband 1, a missense c.740G>A, (p.R247Q) variant in exon 11 of proband 5, and a missense c.562G>A, (p.E188K) variant in proband 6 (Figures 1A and 1C). THOC6 is comprised of seven WD40 repeat domains (Figure 1D) that form a β-propeller structure when folded. These novel variants, like other clinically relevant *THOC6* variants, map to the WD40 repeats that comprise the beta strand structural regions of THOC6 (Figure 1D). The p.Q47* nonsense variant represents a cluster of three *THOC6* variants in the first WD40 repeat that exhibit a consistent genotype-phenotype correlation for both recessive nonsense and missense variants. The same trend is observed for the novel missense variants that are localized to subsequent THOC6 WD40 domains, which support a loss-of-function (LOF) pathogenic mechanism for both *THOC6* variant types.

A LOF mechanism is also implicated by the clinical consistency observed between biallelic inheritance of pathogenic *THOC6* haplotypes and biallelic inheritance of a single haplotype variant alone. Biallelic inheritance of a triple-variant haplotype (TVH), *THOC6* c.[298T>A;700G>C;824G>A], (p.[W100R;V234L;G275D]) has been reported in seven individuals with clinical features of TIDS (Casey *et al.*, 2016; Gupta *et al.*, 2020; Mattioli et al., 2019; Ruaud *et al.*, 2022). The TVH segregates as a founder haplotype in individuals of European ancestry (Mattioli et al., 2019). In comparison, a homozygous *THOC6* c.824G>A; p.G275D variant was identified in siblings with classic TIDS in family 4 who are of South Asian ancestry. This finding provides additional evidence for the pathogenicity of the TVH *THOC6* c.824G>A; p.G275D variant but does not negate the predicted pathogenicity of the corresponding W100R or V234L TVH variants. Pathogenicity of p.W100R or p.V234L are supported by variants detected in WD40 repeats 2 and 4 (Figure 1D). Comparing biallelic inheritance of TVH and *THOC6* c.824G>A; p.G275D suggests a single THOC6 variant in both alleles are sufficient to comprehensively disrupt THOC6, a baseline deficiency not exacerbated by accumulation of additional LOF variants.

5.2.2 *THOC6* variants show mRNA stability with differential effect on protein abundance

To investigate the genetic mechanism of *THOC6*, the impact of on mRNA nonsense mediated decay (NMD) and protein expression on pathogenicity was tested in embryonic stem cells (ESCs) and iPSCs, collectively referred to as human pluripotent stem cells (hPSCs). hPSCs were reprogrammed from two individuals with TIDS (6:IV:1*,* *THOC6*$^{E188K/E188K}$ and 7:V:2, *THOC6*$^{W100*/W100*}$) and their respective unaffected heterozygous parent (6:IV:2, *THOC6*$^{E188K/+}$ and 7:V:1, *THOC6* $^{W100*/+}$) (Figure 1E), preserving the shared genetic background between affected and unaffected conditions. Consistent with a LOF mechanism, reduction in protein expression due to mRNA nonsense mediated decay was predicted. However, *THOC6* mRNA transcripts remain relatively stable between genotypes, as assessed by Actinomycin D treatment where transcription is inhibited, compared to the unstable mRNA, *FOS*, that is quickly degraded (Figure 1F) (Moon et al., 2012). This finding is consistent regardless of *THOC6* variant, with the c.299G>A, (p.W100*) and c.562G>A, (p.E188K) variants exhibiting similar decay rates as wildtype transcripts (Figure S1C and S1D). This finding could reflect defective NMD from failure of *THOC6*-affected mRNA to be exported to the cytoplasm. Nevertheless, the impact on protein expression is divergent between nonsense and missense variants. Significant reductions in THOC6 abundance were detected in

THOC6$^{W100*/W100*}$ and *THOC6*$^{E188K/E188K}$ iPSCs relative to the *THOC6*$^{+/+}$ control, with the most remarkable reduction for *THOC6*$^{W100*/W100*}$. Significant abundance differences were also noted for heterozygous unaffected iPSCs relative to wildtype controls (Figure 1G). Full-length THOC6 in *THOC6*$^{W100*/W100*}$ samples represent a minority readthrough product. Increased frequency of this rare event was promoted by treatment with 30 μM Ataluren, which extends translation by skipping premature termination codons, leading to an increase in THOC6 detectable by Western blot in *THOC6*$^{W100*/W100*}$ iPSCs (Figure 1G). No truncated product was observed by Western blotting in *THOC6*$^{W100*/W100*}$ and *THOC6*$^{W100*/+}$ iPSCs, suggesting THOC6 reduction is due to rapid degradation of an unstable, truncated protein. Stable expression from missense *THOC6* alleles suggests variant THOC6 is likely functionally inactive.

5.2.3 *THOC6* variants interfere with TREX functions

Based on the solved crystal structure (Pühringer *et al.*, 2020), LOF *THOC6* variants are predicted to impair TREX tetramer formation. The WD40 repeat domains of THOC6 form beta strands predicted to provide the structural interface for TREX core tetramer formation. Pathogenic *THOC6* variants disrupt residues conserved in mammals, though several are not conserved across other metazoan species, mirroring TREX composition variability across species and the evolving function for THOC6 in TREX (Figure 2A). Evaluation of the TREX crystal structure indicates that pathogenic THOC6 variants are positioned at the TREX core tetrameric interface, where THOC5-THOC7 interaction is responsible for dimerization and THOC6-THOC5 interaction tetramerizes the complex (Figure 2B, 2C, and 2D). The allelic series of *THOC6* LOF variants implicate a pathogenic mechanism where LOF missense variants are predicted to perturb THOC6 β-propeller folding and/or interactions with THOC5 and THOC7 that are required for TREX tetramer assembly, and nonsense variants would produce a similar outcome due to low protein abundance (Pühringer *et al.*, 2020). Likewise, *THOC6* variants do not alter the protein abundance of other THO/TREX members (Figure 2E). Consistent with normal abundance of functional THO protein, subcellular localizations at nuclear speckle domain were observed by immunohistochemistry (Figures S1E and S1F). Conversely, ALYREF association with the THO subcomplex is diminished in *THOC6*$^{E188K/E188K}$ and *THOC6*$^{W100*/W100*}$ as assessed by co-immunoprecipitation with THOC5 and THOC6 in patient-derived iPSC lines (Figure 2F). These findings suggest a THOC6-dependent association of ALYREF to THO, with implications for the affinity of other adaptors due to the potential disruption of TREX tetramer formation.

5.2.4 *Thoc6* is required for mouse embryogenesis

To investigate *Thoc6* pathogenic mechanisms in mammalian neural development *in vivo*, we used CRISPR/Cas9 genome editing to introduce an insertion variant in mouse *Thoc6* exon 1 that resulted in a premature termination codon predicted to ablate Thoc6 expression (p.P6Lfs*8, herein referred to as *Thoc6*^{fs}) (Figure 3A). *Thoc6*$^{+/fs}$ male and female mice do not display phenotypic abnormalities, but their intercrosses yielded no homozygous offspring. Analysis at selected embryonic days (E) of gestation confirmed *Thoc6*$^{fs/fs}$ littermates die *in utero*. Differences in embryonic morphology between wildtype (WT) and *Thoc6*$^{fs/fs}$ embryos were noted starting at E7.5, the earliest day of analysis. By E9.5, *Thoc6*$^{fs/fs}$ embryos were smaller with delayed development; however, the difference in the developing neocortex was particularly pronounced (Figure 3B). No body turning differences were observed. Consistent with embryonic lethality, THOC6 was undetectable in E8.5 *Thoc6*$^{fs/fs}$ mouse embryos relative to control littermates (Figure 3C). In E9.5 *Thoc6*$^{fs/fs}$ embryos, a developmental timepoint with high Thoc6 expression, Thoc6 is detectable by Western blot at greatly diminished levels (Figure 3C). Embryonic lethality was confirmed by E11.5, indicating one functional allele of *Thoc6* is essential for mouse embryonic development (Figures 3C and 3D).

Forebrain tissues of E8.5-E10.5 *Thoc6* littermates were characterized to identify neurodevelopmental changes in *Thoc6*$^{fs/fs}$ pups. Immunohistochemistry of the telencephalic vesicles revealed a consistently thinner neuroepithelium (PAX6) in *Thoc6*$^{fs/fs}$ compared to *Thoc6*$^{+/+}$ littermate controls (Figures 3E,3F, S2B, and S2D). While E9.5 neuroepithelium showed a relative increase in mitotically active cells (PH3), widespread apoptosis (Cleaved Caspase-3 (C.CASP3)) was noted (Figures 3E and 3F). These findings are consistent with proliferative defects noted in *Thoc6*$^{fs/fs}$ tissue, suggesting THOC6 is important for corticogenesis.

5.2.5 Global mRNA export is not altered in THOC6 models of human *in vitro* neural development

Given the prominent link between the THO subcomplex and RNA export in current literature, we first sought to investigate the impact of THOC6 variants on RNA nuclear export functions in human neural progenitor cells (hNPCs) differentiated from hPSCs (Figure 4A). NPCs are the embryonic cell population frequently implicated in the developmental mechanism of primary microcephaly, a TIDS clinical feature. Defects in mRNA export are typically observed as differential accumulation of polyadenylated (polyA+) mRNA in the nucleus, enriched at nuclear speckle domains (Bahar Halpern et al., 2015). Standard oligo-dT fluorescent *in situ* hybridization (FISH) was performed on hNPCs to visualize polyA+ mRNA signal in nuclear and cytoplasmic cellular fractions (Figures S3A, S3B, and S3C). Comparison of the nuclear-to-cytoplasmic (N/C) polyA+ signal intensity ratios across genotypes indicates a slight reduction in *THOC6* affected samples, suggesting a trend towards nuclear reduction in affected hNPCs relative to *THOC6*$^{+/+}$ and heterozygous unaffected control hNPCs (Figure S3C). The significance of this modest export finding is evident when the N/C polyA+ signal intensity ratios are compared to *THOC6*$^{+/+}$ hNPCs treated with wheat germ agglutinin (WGA), a potent inhibitor of all nuclear pore transport and the positive control for export changes (Mor et al., 2010) (Figure S3A), Relative to all *THOC6* genotypes, WGA treated hNPCs have a significantly higher N/C ratio (Figure S3C) attributed to strong polyA+ mRNA accumulation in the nucleus (Figure S3B). Although bulk mRNA export is largely unaltered in *THOC6*-affected hNPCs, a THOC6 dependent tetramer TREX export cannot be ruled out for specific polyA+ or in mRNP processing functions upstream of export—TREX functions that have not been explored in hNPCs.

5.2.6 THOC6 depletion reveals TREX function in pre-mRNA splicing in hNPCs

Proper mRNP processing, including mRNA splicing, is required for TREX-dependent RNA nuclear export, linking these steps in mRNA biogenesis. Co-transcriptional recruitment of TREX to the 5’ end of maturing mRNPs coupled with described TREX associations with splicing factors (Viphakone *et al.*, 2019) led us to investigate features of mRNA processing for vulnerability to loss of THOC6. To capture RNA processing differences caused by loss of THOC6 function, we performed RNA-sequencing (RNAseq) on ribosomal (r)RNA-depleted RNA extracted from wildtype and heterozygous unaffected and homozygous affected hNPCs (Table S4). Principal components analysis (PCA) of RNAseq data demonstrates reproducibility across, and distinct transcriptomic differences between the affected hNPC replicates compared to the unaffected control hNPC replicates (Figure S4A). Genotype driven differential expression and splicing changes were assessed. To investigate a THOC6-dependent role for TREX in splicing, comparative splicing analysis was carried out using the rMATS pipeline on biallelic *THOC6*$^{E188K/E188K}$ and *THOC6*$^{W100*/W100*}$ samples versus heterozygous controls (Table S5) (Shen et al., 2014). A combined total of 3,796 significant alternative splicing (AS) events were detected in affected hNPCs, representing the major AS types: skipped/cassette exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), retained intron (RI), and mutually exclusive exon (MXE). The most overrepresented AS events observed in affected cells were SE (56%, 2136 of 3796) and RIs (21%, 784 of 3796). The high frequency of RIs is notable and unique relative to splicing defects identified in LOF models of other splicing factors, as well as for normal splicing patterns in neural development (Figure 4B) (Chai et al., 2021; Ellis et al., 2012; Jin et al., 2020; Licatalosi et al., 2008; Llorian et al., 2010; Weyn-Vanhentenryck et al., 2018). AS events in affected hNPCs show comparable inclusion and exclusion of AS junctions, with a slight trend towards inclusion due in part to the high frequency of RIs (Figure 4B). SE and RI splicing events occur by distinct molecular mechanisms, mediated by EJC pathways. Detection of defects in both splicing categories suggests the THO tetramer serves as a molecular platform for coordinating complex splicing events, as opposed to regulation of a specific subset of splicing events controlled by association with and function of individual RNA splicing factors. In agreement with this finding, we did not find consistent motif enrichment for specific RNA-binding proteins at AS junctions. These findings implicate a novel role for THOC6-dependent TREX splicing in mRNP processing in hNPCs.

Since AS motif enrichment analysis of THOC6-affected and control hNPCs transcriptomic data did not reveal trans-regulatory elements responsible for the differential splicing patterns, it was posited that cis-elements may underlie these differences. A maximum entropy model that assesses short sequence motif distributions was used to test the strength of the donor (5') and acceptor (3') of AS events (Yeo and Burge, 2004). A general trend towards weaker splice sites were detected at differential SE, RI, and A3SS events in affected cells (unpaired two-tailed *t* test, Figure 4C). The SE, RI, and A3SS events were enriched in genes with a disproportionately high number of isoforms that show dependence on weak, alternative/cryptic splice sites to facilitate isoform diversity (Figure 4D) (Wang et al., 2015). RI events in affected hNPCs also had weaker splice sites compared to controls, suggesting that THOC6 deficiency induces mis-splicing at weak splice sites. In addition, the AS SE, RI, and A3SS events in affected *THOC6* hNPCs impacted exons/introns that are significantly longer than nonsignificant events (Figure 4E). Likewise, the length of introns retained in RI events were significantly longer, with a 1.4-fold increase in length quantified for significant RI events (P=<0.0001, unpaired two-tailed *t* test, Figure 4E). Lastly, no positional bias was observed for AS events (Figure S4D and S4F). To validate our bioinformatic analysis, AS inclusion trends in select, top, shared events were validated by qRT-PCR, demonstrating a high correlation (unpaired two-tailed *t* test, Figure 4F). Together, the detected RNA processing signature across diverse SE and RI events at weak splices sites suggest impaired splicing fidelity from loss of THOC6.

To investigate the role of RNA misprocessing in ID pathology, we intersected our AS events with the genes that are known to cause syndromic ID, deposited in the SysID database (SysIDdb). 152 genes with significant AS events included or excluded in >10% of transcripts in hNPCS were detected in nonsense and missense affected genotypes (Figure 4G). 185 AS genes in *THOC6*$^{W100*/W100*}$ and 105 AS genes in *THOC6*$^{E188K/E188K}$ hNPCs are known genes causative for syndromic ID represented in the SysIDdb (Figure 4G). Aberrantly spliced ID genes were identified in both THOC6 affected genotypes, consistent with a role for THOC6 in ID. 37 ID genes (1.3% of SysIDdb) are AS in both affected genotypes, identifying genes for shared mechanisms that may preferentially contribute to TIDS pathology. To identify biological mechanisms implicated by THOC6-dependent AS, biological pathway enrichment analysis was performed on mis-spliced genes in affected cells. Genes with differential splicing were significantly enriched for functions in RNA splicing, cell projection organization, membrane trafficking, organelle organization, mitosis cell cycle, and DNA damage response (Figure 4H). RNA processing is tightly controlled by feedback loops (e.g., auto-repression by poison exons or intron retention), which would explain how effects on *cis* elements may lead to changes in *trans* factors (i.e., AS events in splicing regulatory factors).

5.2.7 THOC6-dependent mRNP processing required for hNPC proliferation and differentiation

Alternative mRNA processing events, such as SE and RI in mRNA splicing, can impact gene expression through different mechanisms; isoform ratio differences versus inclusion of premature termination codons that initiate NMD. To investigate the impact of AS events on expression, we performed differential gene expression analysis on *THOC6*$^{E188K/E188K}$ versus *THOC6*$^{W100*/+}$, and *THOC6*$^{W100*/W100*}$ versus *THOC6*$^{W100*/+}$ dyads (Figure S5B). Among the 336 differentially expressed genes (DEGs) in *THOC6*$^{E188K/E188K}$ hNPCs, only 13 had splicing defects (p = 5.3x10^{-3}, Fisher's exact test) compared to 46 mis-spliced DEGs (of 661 DEGs, p = 4.2x10^{-7}, Fisher's exact test) in *THOC6*$^{W100*/W100*}$ hNPCs, indicating a subtle effect of mis-splicing on expression (Figure 5A). Notably, there are nearly double the number of AS genes (ASGs) (435 E188K; 741 W100*) and DEGs (336 E188K; 661 W100*) in *THOC6*$^{W100*/W100*}$ hNPCs, suggesting that the nonsense genotype has a greater impact on splicing and gene expression (Figure 5A). Relevant for TIDS pathology, 20% (68 of 336; p = 1.5x10^{-5}, Fisher's exact test) of *THOC6*$^{E188K/E188K}$ DEGs and 18% (118 of 661; p = 9.7x10^{-6}, Fisher's exact test) of *THOC6*$^{W100*/W100*}$ DEGs are syndromic ID genes, which conveys important information pertinent for understanding the pathogenic mechanisms of TIDS (Figure 5A).

Retained intron AS events are often subject to NMD in protein-coding genes (PCGs), while intron inclusion in lncRNAs alters nuclear export and conformation. Given the high number of RI events detected in *THOC6* affected hNPCs, we tested the correlation between differential expression and RI events. For this analysis, gene-level mRNA abundance fold-change in affected hNPCs was correlated to the change in intron inclusion within transcripts from PCGs or non-coding RNA loci, referred to as percentage spliced-in (ΔPSI). We observed a trend in both *THOC6*-affected genotypes that lower gene expression correlates with greater intron inclusion (slope, *p* = 0.0045 for *THOC6*$^{W100*/W100*}$and *p* = 0.0002 for *THOC6*$^{E188K/E188K}$, simple linear regression) (Figure 5B). Conversely, the quadrant representing differential intron exclusion and elevated expression was prominently represented by lncRNAs, where intron exclusion is implicated in impaired lncRNA function. The three significantly dysregulated ASGs represented in both affected genotypes were *MEG3*, *PAX6*, and *POSTN*. Consistent with the observed trend between RI events and expression, analysis of all DEGs revealed that PCGs make up the largest portion of DEGs (with a greater portion, 96.45% affected in downregulated genes), while the portion of lncRNAs is highest in upregulated genes (6.29%; compared to 1.98% of non-significant genes), reflecting the molecular differences between these distinct mRNA subtypes.

Additional mRNA characteristics that may account for a portion of the observed differential expression are gene length and isoform number. In affected hNPCs, significantly more transcripts from long genes with on average of less than 10 annotated isoforms were identified compared to non-significant genes (Figures 5D and 5E). The trend towards DEGs with fewer transcript isoforms in affected hNPCs suggest alternatively spliced transcripts are more stable in affected cells (Figure 5C). These findings again reflect a requirement for the larger THOC6-dependent TREX tetramer complex function in facilitating mRNP processing of long mRNAs with high expression in brain.

To identify biological pathways predicted to contribute to *THOC6* neuropathology, DAVID analysis was performed to identify biological categories defined by DEGs in*THOC6* affected hNPCs (Figure 5D). Downregulated genes are enriched in integrin cell adhesion, extracellular matrix interactions, PI3K-AKT signaling, and TGF-β signaling pathways, which are critical for brain development (Figure 5F). PI3K-AKT/mTOR signaling regulates cortical NPC proliferation, differentiation, and apoptosis (Andrews et al., 2020; Li et al., 2017). Over 30 genes attributed to the PI3K-AKT/mTOR signaling pathway were downregulated in affected cells, accounting for the significant enrichment (p = <1 x 10^{-13}). *HAPLN1*, *MYC*, *BMPR1B*, *DCN*, *FBN1*, *INHBA*, *ID4*, *THBS1*, *TGFB2,* DEGs enriched in the TGF-β signaling pathway (p = <0.001), have direct implications for TGF-β signaling in neural induction, differentiation, and NPC fate specification in TIDS developmental mechanisms (Meyers and Kessler, 2017; Vogel et al., 2010). Complementary pathways enriched with upregulated DEGs implicate multipotency (*OCT4*, *PAX6*), proliferation, neuron differentiation and WNT signaling pathways (p = <1 x 10^{-6}, p = <0.001, p = <0.001, and p = <0.001, respectively). WNT signaling is known to promote NPC self-renewal expansion during corticogenesis (Harrison-Uy and Pleasure, 2012; Qu et al., 2013). Shared dysregulation of mTOR, TGF-β, and WNT signaling, coupled with upregulation of multipotency factors in affected genotypes, suggests defects in hNPC multipotency and neural differentiation underlie TIDS pathogenesis.

To identify transcription factor networks dysregulated in *THOC6*-affected hNPCs, transcription factor motif enrichment analyses was performed. Significant enrichment of MEF2, LHX3, and SRF target genes was observed in heterozygous controls compared to both affected genotypes (Figure S5D). Using a second analysis tool, ChEA3, differential expression of SOX, FEZF, FOX, and GLI target genes, and downregulation of HEYL, TWIST, FOX, MEOX2, PRRX2, and MKX target genes were enriched in affected hNPCs (Figure S5E). These transcription factor networks are important for neuronal differentiation and fate specification (Jalali et al., 2011; Tsui et al., 2013; Wang et al., 2011), concordant with GSEA findings. Together, these results suggest that gene expression programs that modulate timing of the switch from neural proliferation to differentiation are altered in TIDS.

To refine specific candidate genes implicated in shared TIDS neuropathology, DEGs between affected *THOC6* genotypes were intersected. 12 genes were upregulated and 117 were downregulated in affected hNPCs, with notable lncRNAs represented. Significant enrichment was detected in Integrin 1 pathway and extracellular matrix protein interaction networks (Figure S5C). Using mRNA obtained from three additional replicate differentiations of hNPCs per genotype, significant upregulation of *MEG3*, *MEG8*, *ESRG*, and *NEAT1* lncRNAs was confirmed by qRT-PCR (Figures 5G). RNA FISH confirmed increased expression of *MEG3* in affected hNPCs compared to controls, with elevated signal observed in both nuclear and cytoplasmic fractions (Figure 5H). Upregulation of functional lncRNAs *NEAT1* and *MEG3* has been linked to activation of WNT activation and suppression of TGF-β signaling, respectively (Cui et al., 2019; Mondal et al., 2015). Concordant with these findings, the protein level of WNT and TGF-β signaling components in *THOC6*-affected hNPCs exhibit a corresponding differential up- and down-expression relative to controls. Specifically, WNT signaling components WNT7A and TP53 showed increased protein expression, with higher abundance detected in affected hNPCs (Figure 5I). TGF-β pathway proteins HAPLN1 and TGFB2 also showed reduced protein expression in affected hNPCs together with high CEMIP and DKK2 (Figure 5I). We propose that loss of THOC6 leads to lncRNA-mediated dysregulation of key developmental signaling pathways which has implications for the balance of proliferation and differentiation during neural development.

5.2.8 Apoptotic upregulation and retained intron enrichment in *Thoc6$^{fs/fs}$* E9.5 mouse forebrain

To investigate the conserved THOC6-dependent TREX functions that account for divergent phenotypic outcomes between mammalian models, mRNP processing was assessed in E9.5 mouse brain using complementary RNAseq experiments to those performed in hNPCs (Figures 6A and S6A). Three biological replicates were analyzed per genotype (*Thoc6$^{+/+}$*, *Thoc6$^{fs/+}$*, *Thoc6$^{fs/fs}$*). Fewer significant AS events (FDR <0.05) were detected in *Thoc6$^{fs/fs}$* E9.5 brain than in affected hNPCs, but the pattern of AS events was recapitulated, with the majority of AS events categorized as SE (45%) and RI (26%) (Figure 6B). Greater than 40 PSI was quantified in the *Thoc6$^{fs/fs}$* transcriptome and retained and excluded intron events in *Cenpt*, *Admts6*, and *Fam214b* were validated (Figures 6C and S6B). Maximum entropy model analysis of splice junctions revealed significantly weaker 3’ splice site strengths for SE events and weaker 5’ splice sites associated with RI events in the *Thoc6$^{fs/fs}$* mouse model. While this signature of splice site weakness is more modest in mouse than in human THOC6 models, these findings suggest a conserved role of THOC6-dependent TREX tetramer in coordinating mRNA processing that precedes TREX export functions (Figures 6D and S6C).

Notably, biological pathway and network enrichment analysis of AS genes identified mRNA processing, pre-miRNA processing, de-adenylation of mRNA, central nervous system development, forebrain development, multicellular growth, response to oxidative stress, cytoskeletal organization, and neuron projection (Figure S6D) — several of the biological categories associated with hNPCs ASGs. These shared findings suggest selective conservation of mRNP processing mechanisms by *THOC6* in mouse and human forebrain.

To assess the correlation between THOC6 mRNP processing defects and expression, differential expression analysis of *Thoc6*$^{fs/fs}$ forebrain mRNA sequencing data was performed compared to *Thoc6*$^{+/+}$ controls. Of note, *Thoc6* mRNA is downregulated (two-fold, *p*=<0.0001), consistent with NMD (Figures 6E and S6F). In this model, 5x more genes were upregulated (144 genes) than downregulated (27 genes). Nevertheless, downregulated genes may covey important pathology. First, downregulated genes functionally converge on neurogenesis, proliferation, and differentiation pathways (Figure 6F). Upregulated genes are implicated in the hypoxic response, HIF-1 signaling pathway, and glycolysis—biological categories indicative of increased apoptosis in affected cells (Figure 6F). To investigate if altered transcription factor networks contribute to pathway dysregulation, we performed GSEA transcription factor motif enrichment analysis. *HIF1*, *NRSF*, *SMAD3*, and *STAT3* target genes were enriched in *Thoc6*$^{fs/fs}$ E9.5 forebrain (Figure S6G). HIF-1 and STAT3 can induce apoptosis in response to hypoxia (Greijer and van der Wall, 2004; Zhou et al., 2021), which is consistent the observed elevation of apoptosis in *Thoc6*$^{fs/fs}$ E9.5 neuroepithelium (Figures 3E and 3F). SMAD3 signaling is activated by TGF-β to promote cortical differentiation (Vogel *et al.*, 2010), suggesting shared disruption of TGF-β signaling in both human and mouse model systems.

DEGs shared between mouse and human model systems are consistent with conserved TIDS molecular pathology (Figure 6G). More *Thoc6*$^{fs/fs}$ DEGs overlapped with *THOC6*$^{W100*/W100*}$ (23 genes) than *THOC6*$^{E188K/E188K}$ (9 genes) samples, and include genes involved in neurogenesis, hypoxic response, and synapse regulation. Validation of *Ier3*, *Islr2, Wnt7a*, *Kcnt2*, *Anax2*, and *Vegfa* DEGs shared across affected models were confirmed by qRT-PCR in three additional E9.5 forebrain biological replicates for *Thoc6*$^{+/+}$, *Thoc6*$^{fs/+}$, and *Thoc6*$^{fs/fs}$ samples (Figures 6E and S6H). Overlapping affected human and mouse molecular mechanisms suggest shared pathology. However, the extent of upregulation of genes in response to increased apoptosis is exacerbated in mouse, highlighting species-specific phenotypic differences due to loss of THOC6.

5.2.9 Delayed differentiation and elevated apoptosis in *THOC6*-affected forebrain organoids

THOC6 pathogenesis in human cortical development was investigated using dorsal forebrain-fated organoids, neural differentiated from iPSC lines (Figure 7A). Forebrain organoids recapitulate the cellular heterogeneity and developmental dynamics of early corticogenesis (Qian et al., 2016). Within each organoid, several neural rosette (NR) structures develop stochastically to recapitulate features of *in vivo* ventricular zone development, including hNPC proliferation and differentiation to cortical neuron fates (Figure 7A). NR morphology was evaluated in cortical organoids at 28 days of neural differentiation (ND) from three independent differentiations per genotype. To minimize the effect of inter-cell line NR variability, the following analyses focus on heterozygous unaffected and homozygous affected comparisons. To characterize the NR proliferative niche, the maximum thickness of the NR neuroepithelial center was measured as defined by N-cadherin immunostaining and pseudostratified NR cytoarchitecture by Hoescht staining (Figures 7B, 7C, S7A, and S7B). *THOC6*-affected organoids show significantly thinner pseudostratified neuroepithelium (*p* = <0.0001, two-tailed *t* test), concordant with reduced NR size composed of fewer cells (*p* = <0.0001, two-tailed *t* test) (Figures 7B, 7C, S7A, and S7B). Given the stark upregulation of apoptosis observed in *Thoc6*$^{fs/fs}$ E9.5 mouse forebrain, we investigated the contribution of apoptosis as a mechanism of reduced NR size in the *THOC6*-affected organoids (Figures 7B, 7D, S7A, and S7B). A significantly higher proportion of affected NR cells expressed the apoptotic marker C.CASP3, providing evidence for shared mechanism of altered corticogenesis (*p* = <0.0001, two-tailed *t* test) (Figure 7D).

To assess alterations in the timing of differentiation in affected NRs, we performed EdU-pulse labeling at day 21ND for 24 hours to label mitotically active cells, followed by organoid immunohistochemistry analysis at day 28ND (Figures 7E, 7F, S7C, and S7D). To assess the balance of multipotency and differentiation EdU, KI67, and DCX co-labeled cells per NR were quantified. A significant increase in cells co-stained with the proliferation marker KI67 and EdU per affected NR were detected at day 28ND (*p* = <0.0001, two-tailed *t* test), indicating affected NPCs remain mitotically active longer than control NPCs (Figure 7F). This finding paired with elevated mRNA and protein expression of OCT4 in affected hNPCs (data not shown) supports retention of multipotency model. Consistent with this finding, we observed a significant reduction in the fraction of EdU cells co-labeled with the migrating neuron marker doublecortin (DCX) in affected NRs (*p* = <0.0001, two-tailed *t* test) (Figure 7F). Paired with the prolonged proliferation dynamics, this suggests a disruption to the differentiation timeline in affected organoids.

To investigate effects of reduced NR growth on organoid size, we measured whole organoid cross section areas weekly from day 21 to day 42. Compared to the steady size increase of *THOC6*$^{+/+}$ organoids, affected organoids showed a slower growth rate (E188K/E188K: *p* = 0.0122; W100*/W100*: *p* = 0.0362) (Figure 7G). Together, our findings implicate a pathogenic mechanism of delayed differentiation due to reduced NPC proliferative capacity and elevated apoptosis with subsequent cortical growth impairment in affected organoids.

5.3 DISCUSSION

There is a growing cohort of individuals with TIDS due to variant discovery resulting from exome based genetic testing (Accogli *et al.*, 2018; Amos *et al.*, 2017; Anazi *et al.*, 2016; Anazi *et al.*, 2017; Beaulieu *et al.*, 2013; Boycott *et al.*, 2010; Casey *et al.*, 2016; Gupta *et al.*, 2020; Hassanvand Amouzadeh *et al.*, 2020; Kiraz *et al.*, 2022; Mattioli *et al.*, 2018; 2019; Ruaud *et al.*, 2022; Zhang *et al.*, 2020). However, our understanding of the molecular functions of THOC6 are limited, despite robust research on the TREX complex (Chi *et al.*, 2013; Dias et al., 2010; Guria *et al.*, 2011; Heath *et al.*, 2016; Katahira et al., 2009; Katahira et al., 2013; Luna *et al.*, 2012; Maeder *et al.*, 2018; Mancini et al., 2010; Masuda *et al.*, 2005; Peña *et al.*, 2012; Pühringer *et al.*, 2020; Rondón *et al.*, 2010; Tran et al., 2014; Viphakone *et al.*, 2019; Wickramasinghe and Laskey, 2015; Zuckerman et al., 2020). Information implicating important species differences in TREX composition and function that are THOC6-dependent serve to highlight this gap in knowledge (Chávez et al., 2000; Jimeno et al., 2002; Meinel et al., 2013; Strässer et al., 2002). Our findings support a novel *THOC6* LOF model of TIDS pathogenesis whereby pathogenic variants impair the formation of the THO/TREX tetramer complex that facilitates multivalent protein-mRNA interactions to support coordination of mRNP processing and export functions (Figure 7H). The resulting molecular impact preferentially disrupts processing of long mRNAs, with complex splicing patterns that are compounded by genetic feedback loops that regulate splicing. The neurological features of TIDS and the enrichment of long genes expressed in the brain further implicate a critical role for the THOC6-dependent TREX tetramer in providing a platform for enhanced coalescence of mRNP processing cofactors and maintenance of mRNA structural integrity during splicing of long mRNA transcripts important for brain development. Cell-type and organism specific requirements for splicing and gene length are predicted to inform the variation in tolerance that underlies distinct human phenotypic presentations and interspecific differences.

The crystal structure of human THO-UAP56 was recently solved, implicating several putative consequences of THOC6-dependent TREX tetramer disruption (Figure 7H) (Pühringer *et al.*, 2020). Defects in TREX tetramer assembly are not predicted to disrupt formation of stable functional dimers, allowing THOC6-depleted models to discriminate between dimer and tetramer functions. The tetramer affords greater surface area for mRNP processing and permits enhanced co-adaptor loading of known and potentially species-specific TREX cofactors in species with a substantial splicing burden. Certainly, we see evidence for reduced association of ALYREF with THO complexes by THOC6 co-immunoprecipitation in *THOC6*-affected iPSCs, providing indirect evidence for altered TREX tetramer formation with impaired binding of tetramer-associated mRNP processing cofactors (Figure 2F). Tetramer formation juxtaposes two UAP56 helicases on each end of dimers (Chang *et al.*, 2013) to support a greater number of cofactors with broad mRNP processing capabilities (e.g., CHTOP in 3' end processing and ALYREF in splicing) (Viphakone *et al.*, 2019) known to compete for the limited number of UAP56 binding sites (Hautbergue et al., 2009; Heath *et al.*, 2016; Izumikawa et al., 2018). The tetramer also affords a greater surface area to maintain the structural integrity of long mRNA transcripts during mRNP processing and export. Through this process, TREX serves as a mRNA chaperone to prevent formation of DNA-RNA hybrid or R-loop structures that can promote genome instability (Luna *et al.*, 2019; Pérez-Calero *et al.*, 2020). Thus, the TREX tetramer helps ensure mRNP quality control and evasion of degradation by the nuclear exosome (Fan et al., 2017). The tetramer may also enable multiple TREX complexes to simultaneously bind several mRNP regions (Pühringer *et al.*, 2020) to facilitate compaction and/or protection of longer transcripts with elevated splicing.

Additional insights into dimer versus tetramer functions come from functional investigation of THO monomer components. For instance, THOC2 and THOC5 are required for the maintenance of ESC pluripotency, whereas TREX tetramer cofactors ALYREF and UAP56 impact differentiation (Wang et al., 2013). This phenotypic difference may reflect cell- and/or species-specific TREX dimer versus tetramer mRNP processing and export functions, with THO monomers being essential for viability and tetramer cofactors critical for differentiation and development. Enhanced mRNA processing efficiency provided by UAP56, ALYREF, CHTOP, and other export adaptors associated with the TREX tetramer are susceptible to *THOC6* pathogenic variants. Likewise, the diversity of expressed isoforms increases phylogenetically (yeast versus mammals), as well as during mammalian differentiation (Chen *et al.*, 2006; Mazin et al., 2021; Nagasaki *et al.*, 2005), especially in the brain, highlighting the vulnerability of neural development to *THOC6* loss. (Mauger *et al.*, 2016; Wang et al., 2008). Neuronal expression of long genes is imperative for proper neuronal differentiation and synaptogenesis (Zylka et al., 2015). ALYREF is predicted to interact with other UAP56/TREX-bound adaptors associated to the same mRNP (Pühringer *et al.*, 2020), suggesting that the tetramer facilitates increased mRNP compaction required for proper export, especially of longer mRNAs. It is possible that enhanced selectivity of mRNP processing and export co-evolved with TREX conformation, contributing to organismal complexity and distinguishing mammalian and yeast cells.

In addition to its prominent role in nuclear RNA export, TREX has been implicated in splicing functions that correlate to increased export efficiency of spliced mRNAs (Masuda *et al.*, 2005; Viphakone *et al.*, 2019). As such, a significant reduction in global polyA+ RNA nuclear export was predicted to result from loss of THOC6-mediated TREX function. However, the absence of global export defects provides evidence to suggest that the dimers may be present and functional for RNA export in *THOC6*-affected hNPCs. Given this finding, alternative pathogenic mechanisms were investigated that may represent tetramer functions. Significant splicing changes implicate a pathogenic mechanism where THOC6-dependent disruption of TREX tetramer formation indirectly disrupts coordination of multiple steps of mRNP processing, including splicing, upstream of polyA+ mRNA packaging and export. This interpretation is supported by the diversity of mRNP processing functions attributed to tetramer-associated cofactors. UAP56 and ALYREF play important roles in mediating pre-mRNA splicing decisions (Shen et al., 2007; Viphakone *et al.*, 2019). This finding does not rule out the possibility that THOC6 plays a direct role in pre-mRNA splicing outside of mediating TREX core tetrameric assembly on the EJC. That THO member THOC5 can interact with unspliced transcripts (Chi *et al.*, 2013), and WD40-repeat domains facilitate splicing factor interactions with pre-mRNA are evidence in support of this possibility (Jin *et al.*, 2020).

Our findings implicating THOC6-dependent TREX tetramers as indirect facilitators of splicing by coordinating the mechanics of mRNP processing is also supported by enrichment of aberrant splicing events at weaker splice sites. Weak splice sites are most often utilized by transcripts during alternative splicing, and genes with elevated isoform diversity from alternative splicing are more susceptible to disruption of the overall integrity of mRNP processing in *THOC6*-affected hNPCs. Long genes that are highly expressed in the brain particularly rely on such infrastructure to ensure pro-neural gene expression. This may also be supported by the observed enrichment of retained introns in THOC6-deficient cells. Indeed, retained introns in unaffected tissues typically have weaker 5' and 3' splice sites compared to other splice junctions (Monteuuis et al., 2019). In addition, intron retention increases during mammalian differentiation (Braunschweig et al., 2014), again suggesting differentiated cells may be more susceptible to loss of THOC6-mediated TREX tetramer functions by miscoordination affecting splicing outcomes.

Given the conserved splicing and nuclear export functions in mammals, the phenotypic discrepancy between *Thoc6*$^{fs/fs}$ mouse embryonic lethality and human biallelic pathogenic *THOC6* variants is notable. Superficially, this finding suggests that humans are more tolerant to *THOC6* variants. Alternatively, this may represent differences in the pathogenic mechanisms of each variant or interspecies sensitivities to splicing defects. In addition, the downstream genetic mechanisms of *Thoc6* and *THOC6* variants are divergent. Human pathogenic *THOC6* variants reside in the WD40 repeat domains whereas the mouse *Thoc6^{fs}* variant is located upstream of the first WD40 repeat domain. Nonsense *THOC6* variant transcripts are stable and readthrough permits limited THOC6 expression. Conversely, *Thoc6^{fs}* transcripts are subject to NMD, impairing THOC6 expression. This suggests that minimal THOC6 expression may be sufficient for human embryogenesis. Splicing requirements may also account for interspecific phenotypic differences. While SE and RI events were the predominant splicing defects in both model systems, AS events affect more protein coding transcripts and lncRNAs in THOC6-affected hNPCs. Despite less AS events in *Thoc6$^{fs/fs}$* cells, lethality in the mouse could be attributed to aberrant alternative splicing of specific transcripts important for mouse embryogenesis. Additionally, previous findings indicate that RI events account for a substantial portion of splicing variation in the primate prefrontal cortex, a trend that is most pronounced in humans (Mazin et al., 2018). Although intron retention is a known mechanism of mouse neuronal gene regulation by initiating RNA exosome-mediated degradation (Yap et al., 2012), it is possible that human cells are more tolerant than mouse cells to elevated intron retention. Further investigation of these interspecific differences is important for generating translationally relevant discoveries.

The number of ID genes that are mis-spliced in *THOC6* affected hNPCs relative to controls implicate shared underlying developmental mechanisms of ID pathology. However, developmental impact of individual defects on TIDS neuropathology is complicated by the compounding effects of constitutive THOC6 LOF models. In addition to trends shared with the mouse model, we show that biallelic THOC6 LOF is responsible for disruption of key TGF-β and Wnt signaling pathways via a mechanism that involves dysregulation of signaling components and lncRNAs resulting in delayed hNPC differentiation, prolonged retention of multipotency, and enhanced apoptosis. This is exemplified by intron retention and upregulation of *MEG3* in affected hNPCs. *MEG3* is linked to the regulation of TGF-β signaling and other EZH2 common target genes (Mondal *et al.*, 2015). Our findings suggest that RI events alter *MEG3* subcellular localization, expression, and downstream WNT signaling that increases multipotency and disrupts the balance of proliferation and differentiation in affected hNPCs. A shift towards cytoplasmic localization of lncRNAs has evolved in human cells, which is important for the maintenance of stem cell pluripotency (e.g., cytoplasmic *FAST* binds E3 ubiquitin ligase β-TrCP to block its interaction with β-catenin and enable activation of Wnt signaling) (Azam et al., 2019; Guo et al., 2020). Given the increased diversity of lncRNA functions in human developmental biology, mouse cells may be less tolerant to lncRNA dysregulation than human cells. In addition, *MEG3* is also upregulated by CREB (Zhao et al., 2006) whose target genes are affected in *Thoc5* conditional knockout mouse cortical neurons (Maeder *et al.*, 2018), potentially reflecting a shared mechanism of THO dysregulation in neural cells. While our analyses from mouse and human organoid models of *Thoc6/THOC6* disruption provide insight into the molecular pathology of early neural development, later analysis of synaptic physiology will be important to elucidate mechanisms of neuronal dysfunction in TIDS.

Altogether, our findings expand the TIDS clinical population and provide novel functional insight into the pathogenic mechanisms of biallelic LOF variants in *THOC6* using comparative mammalian model systems. Functional studies with THOC6 enable us to assess TREX tetramer function while retaining THO subcomplex formation, and our findings provide novel insight into TREX splicing functions separate from export. Future work is needed to dissect the direct and indirect effects of THOC6 loss and confirm endogenous tetramer disruption under native protein conditions. In addition, the well-known role of several TREX members in determination of polyadenylation site choice necessitates research focused on characterizing global aberrant alternative polyadenylation changes that could be contributing to dysregulation. Follow-up investigation at later-stage cortical organoids, and with use of unbiased single-cell RNAseq profiling, will allow for more detailed assessment of the developmental consequences of observed defects for cortical lamination and cell type composition. Lastly, it will be important to determine if alterations in mRNA processing/export also underlie synaptic defects—a morphological basis of ID, and a prominent clinical feature shared between *THOC2* and *THOC6*-associated neurodevelopmental disorders.

ACKNOWLEDGEMENTS

The authors would like to especially thank all the family members who participated in this study and gave permission to use genetic and clinical information. We thank Stephanie Moon for guidance on mRNA stability assay, Sumin Kim for reagents and protocols for oligo-dT FISH, and the University of Michigan Advanced Genomics, Transgenic Animal Model, and Microscopy Cores for sequencing and equipment. This research was supported by NSF Doctoral Dissertation Research Improvement Grant 1919671: F055138, Wenner-Gren Foundation for Anthropological Research N02820: 9967, Leakey Foundation Dissertation Fieldwork Grant, University of Michigan Rackham Candidate Research Grant, the Joan B. Kessler Award, and the University of Michigan Pandemic Research Recovery Award to E.A.W, as well as the National Institutes of Health award numbers R01AWD010411 to S.L.B, A.S., and K.M.G, R00HD099403 to A.E.S, F31NS122207 and T32NS077888 to G.R.L, and T32GM008056 to K.J.

AUTHOR CONTRIBUTIONS

A.E.S. generated the mouse model and human cell lines. E.A.W. performed human cell culture and organoid experiments and all bioinformatics analyses. G.R.L., K.J., B.B., S.L.R, C.D.P, and A.E.S. performed mouse experiments. D.P., S.L.R., and C.D.P. contributed to immunohistochemistry analyses. E.A.W. and G.R.L. performed the molecular biology experiments. E.A.W. created the figures. E.A.W., G.L., S.L.B, and A.E.S. wrote the manuscript. A.E.S., S.L.B, A.S., S.B., S.M., R.P., M.H, R.J.H, E.K, A.O, J.D., A.K., K.C., E.J.P., R.J.L., R.R.L., T.L.E., C.G., K.M.G., K.B., and A.S. contributed to the clinical work.

METHODS

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human subjects

All subjects or parents/guardians provided informed consent and were enrolled in institutional review board-approved research studies. In all cases, the procedures followed were in accordance with the ethical standards of the respective institution's committee on human research and were in keeping with international standards. Probands 1-3 and 5 were identified through GeneMatcher (Sobreira et al., 2015). Details for all subjects are provided in Table 7.

Animal models

All mice were maintained according with the National Institutes of Health Guidelines for the Care and Use of Laboratory Animals and were approved by the Case Western Reserve Institutional Animal Care and Use Committee. CRISPR genome editing was performed in the University of California, San Diego Transgenic and Knockout Mouse Core. C57BL/6JN hybrid mice (Jackson Laboratory, 005304) were used for CRISPR editing of the *Thoc6* locus. Founder mice with the *Thoc6*$^{fs/+}$ allele were intercrossed with C57BL/6JN mice (Jackson Laboratory, 005304) for line maintenance. All *ex vivo* analyses were performed on tissue collected from mice of both sexes at embryonic day (E) 8.5-10.5. Sex-dependent differences were not assessed.

Litters were genotyped by allele-specific polymerase chain reaction (AS-PCR). Genomic DNA was prepared from mouse tissue samples as previously described (Truett et al., 2000). AS-PCR for each allele was assembled using the standard GoTaq DNA polymerase (Promega) protocol. Reaction conditions were executed as recommended by the manufacturer.

Human ESC/iPSC culture

Human ESC and iPSC lines were cultured using feeder free conditions on Matrigel (Corning with mTeSR-1 (STEMCELL Technologies). Lines used in this manuscript include: H9 (*THOC6*$^{+/+}$, 46XX, ESCs, WA09, WiCell), AS0035 (*THOC6*$^{+/+}$, 46XX, iPSCs, New York Stem Cell Foundation (NYSCF) Diabetes iPSC Panel), AS0041 (*THOC6*$^{+/+}$, 46XY, iPSCs, NYSCF Diabetes iPSC Panel), KMC6002 (*THOC6*$^{E188K/+}$, 46XY, iPSCs), KMC6003 (*THOC6*$^{E188K/E188K}$, 46XY, iPSCs), KMC7001 (*THOC6*$^{W100*/+}$, 46XY, iPSCs), KMC7002 (*THOC6*$^{W100*/W100*}$, 46XY, iPSCs). Passaging was performed using mTeSR-1 supplemented 1 nM ROCK inhibitor (BD Biosciences 562822) to prevent differentiation. Both manual and chemical dissociation with Versene (Gibco 15040066) were performed for splitting. Sanger sequencing validation of genotypes (Figures S5.1B and S5.1C) and CNV microarray analysis (Illumina Bead Array, analysis with Genome Studio v2.0) were performed on all lines to ensure no pathogenic changes were acquired during culturing. No further validation of iPSCs lines was performed in our lab.

METHOD DETAILS

Whole exome sequencing and analysis

Exome libraries from genomic DNA of all BBIS-affected probands were prepared and captured with the Agilent SureSelectXT Human All Exon 50Mb Kit for Probands 1 & 4-7 and Individual 5, the Agilent SureSelectXT Clinical Research Exome kit for Proband 3, and the TrueSeq Rapid Exome Kit for Proband 2. Further, exome libraries were sequenced on an Illumina HiSeq or NextSeq instrument as described previously (Srivastava et al., 2018).

Reads were aligned to the human genome version 19 (hg19/NCBI build 37) using Burrows-Wheeler Aligner (BWA). Variant calling of single nucleotide variants (SNVs) and copy number variants (CNVs) was performed using GATK, VEP, CoNIFER, and NextGENE Software. Average depth of coverage was calculated across all targeted regions. The data were filtered and annotated from the canonical *THOC6* transcript (ENST00000326266.8 and ENSP00000326531.8) using in-house bioinformatics software (Müller et al., 2017). Variants were also filtered against public databases including the 1000 Genomes Project phase 311, Exome Aggregation Consortium (ExAC) v.0.3.1, Genome Aggregate Database (gnomAD), National Heart, Lung, and Blood Institute Exome Sequencing Project Exome Variant Server (ESP6500SI-V2). Those with a minor allele frequency >3.3% were excluded. Additionally, variants flagged as low quality or putative false positives (Phred quality score 14; 15 < 20, low quality by depth <20) were excluded from the analysis. Variants in genes known to be associated with ID were selected and prioritized based on predicted pathogenicity.

Sanger sequencing

All variants discovered by WES were confirmed with Sanger sequencing of *THOC6* for each individual and respective family members who submitted samples except Proband 1 where high-coverage WES of *THOC6* in the proband and parents was deemed sufficient to report without Sanger confirmation. Chromatograms were analyzed using Sequencer and Geneious Prime Software (v.2022.1.1).

Cerebral organoid generation

Telencephalic cerebral organoids were generated based on previously published protocols (Lancaster *et al.*, 2013), with few modifications to start with low cell density in order to generate smaller and more consistent embryoid bodies (EBs). Briefly, human and chimpanzee iPSCs were passaged into 96-well V-shaped bottom ultra-low attachment cell culture plates (PrimeSurface® 3D culture, MS-9096VZ) to achieve a starting cell density of 600-1,000 cells per well in 30 μl of mTesR^{TM}1 with 1 nM ROCK inhibitor. After 36 hours, 150 μl of N-2/SMAD inhibition media (cocktail of 1X N-2 supplement (Invitrogen 17502048), 2 μM A-83-01 inhibitor (Tocris Bioscience 2939), and 1 mM dorsomorphin (Tocris Bioscience 309350) in DMEM-F12 (Gibco 11330032)) was added for neural induction. On day 7, EBs were transferred to Matrigel-coated plates to enrich for neural rosettes at a density of 20-30 EBs per well of a 6-well plate, and media was changed to neural differentiation media (0.5X N-2 supplement, 0.5X B-27 supplement (Invitrogen 17504044) with 20 pg/μl bFGF and 1mM dorsomorphin inhibitor in DMEM/F-12). For organoid differentiation EBs were outlined on day 14 using a pipet tip and uplifted carefully with a cell scraper to minimize organoid fusion and tissue ripping. Media was changed once more to N-2/B-27 with bFGF only and plates with uplifted organoids were placed on a shaker in the incubator set at a rotation speed of 90. On day 14, media was changed once more to N-2/B-27 with bFGF only. Prior to day 14, media changes were performed every 48 hours. After day 14, daily media changes were performed until collection. For monolayer NPC differentiation, neural rosettes were scored and uplifted on day 14, dissociated in Accutase (Gibco A1110501), and re-plated on poly-L-ornithine (PLO)/Laminin-coated plates for NPC expansion, selection, and passaging.

15 µg/mL PLO (Sigma-Aldrich P4957) diluted in DPBS (Gibco 14040-133); 10 µg/mL laminin (Sigma-Aldrich L2020) diluted in DMEM/F-12.

Western blot analysis and immunoprecipitation

ESCs, iPSCs, and NPCs used for western blot analysis were pelleted and lysed in RIPA buffer supplemented with 1:50 protease inhibitor cocktail (Sigma-Aldrich P8340) and 1:100 phosphatase inhibitor cocktail 3 (Sigma-Aldrich P0044) using mortar and pestle coupled with end-over-end rotation for 30 minutes to 1 hr at 4°C. Protein concentration was quantified by BCA (Thermo Scientific Pierce A53227). Lysis samples were then incubated at a 1:3 ratio with 4x Laemmli sample buffer (Bio-Rad) supplemented with 10% BME and incubated at 95°C on a heat block for 5 minutes for denaturation. For co-immunoprecipitation, primary antibodies anti-THOC5 and anti-THOC6 (1:50 dilution in 1x PBS with Tween-20) were incubated overnight at 4°C with Dynabeads Protein G (Invitrogen, 10003D). Beads were washed and cell lysis (35 µg of protein) was added for incubation overnight at 4°C with rotation. IP samples were prepared according to manufacturer's instructions with elution in Laemmli sample buffer with 10% BME. For promotion of readthrough of premature termination codons, ataluren (eMolecules NC1485023) was dissolved in DMSO added to ESC/iPSC media at a final concentration of 30 μM for 48 hours as previously described (Roy et al., 2016). Protein was then extracted as described above.

Samples were loaded into 4-20% SDS-polyacrylamide gels (Bio-Rad) and proteins were separated by electrophoresis at 30V for ~4 hours room temperature. Separated proteins were then transferred to PVDF membranes (Millipore) overnight using a wet transfer system (Bio-Rad) at 4°C. For immunoblotting, membranes were incubated in 5% milk blocking buffer (1x TBS-T) followed by primary antibody incubation overnight at 4°C with rotation. Membranes were washed 3 times for 5 minutes in 1x TBS-T and then incubated with secondary antibodies for 1-2 hours at room temperature. Membranes underwent final washes before developing using West Femto Substrate (ThermoFisher 34095) with film exposure. Primary antibodies used: anti-THOC6 (1:1000, Abnova H00079228-A01), anti-THOC1 (1:200, Bethyl Laboratories A302-839A), anti-THOC2 (1:200, Bethyl Laboratories A303-630A), anti-THOC5 (1:200, Bethyl Laboratories A302-120A), anti-ALYREF (1:200, Sigma Aldrich, A9979), anti-CHTOP (1:200, Invitrogen PA5-55929), anti-β-actin (1:250, Abcam ab6276). Secondary antibodies used: donkey anti-rabbit HRP-conjugated (1:5000, Cytiva NA9340V) and goat anti-mouse HRP-conjugated (1:1000, Invitrogen 32430).

Immunofluorescence and Single-molecule Fluorescence *in situ* Hybridization

Human NPCs were fixed in 4% paraformaldehyde (PFA) for 20 minutes. Human cortical organoids and mouse embryos were fixed in 4% PFA for 24 hours at 4°C, cryoprotected in 15% and 30% sucrose in 1x DPBS for 24 hours at 4°C, then embedded in OCT with quick freezing in -50°C 2-methylbutane, followed by cryosectioning for immunostaining. Mouse embryos were sectioned at 13 μm and organoids at 16 μm. Samples for immunostaining were incubated for 1 hour with blocking buffer (5% NDS (Jackson ImmunoResearch) 0.1% Triton X-100, 5% BSA) at room temperature, then overnight with primary antibodies diluted in blocking buffer at 4°C, and for 1-2 hours in secondary dilution at room temperature. Washes performed in PBS. For nuclear staining, samples were incubated at room temperature for 10 minutes in Hoescht or DAPI (1:1000 dilution in PBS) prior to final washes. For EdU labeling detection, the Click-IT EdU imaging kit (Invitrogen C10337) was used according to the manufacturer's instructions. After incubation with the Click-IT reaction cocktail, sections were blocked and immunostained as described above. Some antibodies required antigen retrieval via incubation in heated 10 mM sodium citrate solution (95-100°C) for 20 minutes prior to immunostaining. Primary antibodies used: mouse anti-PAX6 (1:250, Abcam, MA-109), rabbit anti-KI67 (1:200, Abcam ab16667), rat anti-PH3 (1:250, Abcam ab10543), rabbit anti-cleaved caspase3 (1:100-1:400, Cell Signaling 9661), mouse anti-N-Cadherin (BD Biosciences 610920), goat anti-DCX (1:400, Santa Cruz Biotechnology A1313), rat anti-CTIP2 (1:500, Abcam ab18465), rabbit anti-PAX6 (1:100, BioLegend) and goat anti-SOX1 (1:100, R&D Biosystems), and mouse anti-KI-67 (1:500, Cell Signaling). AlexaFluor-conjugated secondaries used: donkey anti-mouse 647 (1:400, Invitrogen A31571), donkey anti-rat 555 (1:400, Invitrogen A48270), and donkey anti-rabbit 488 (1:400, Invitrogen, A21206).

Embryos and organoids used for RNA Fluorescence *in situ* Hybridization (FISH) were fixed and cryoprotected as indicated above using RNAse-free PBS. RNAse-zap treatment of sectioning equipment was performed prior to cryosectioning. NPCs for RNA FISH were fixed in RNAse-free 4% PFA then permeabilized in PBS-TritonX (0.1%) for 15 minutes. Hybridizations were then performed overnight at 37°C with a final concentration of 2 ng/µl of Cy3-conjugated oligo-dT(30-mer) probe, *MALAT1* (Quasar-670, Stellaris VSMF-2211-5), and/or *MEG3* (Quasar-570, Stellaris VSMF-20346-5). Saline-sodium citrate washes were performed before and after hybridization, followed by nuclear staining with RNAse-free Hoescht-PBS wash (1:1000 dilution) and final wash in RNAse-free PBS.

Glass covers were mounted onto all slides with Prolong Gold (Molecular Probes S36972) and incubated for 24 hours at room temperature prior to imaging. Imaging was performed with a Nikon A1ss inverted confocal microscope using NIS-Elements Advanced Research software. Image analysis was performed using Fiji (ImageJ) software. For oligo-dT FISH, Z-series images were taken every 0.2 μm across entire width of cells for each genotype using same laser intensity settings and collapsed by max intensity using Z project tool in Fiji for quantification of nuclear and cytoplasmic fractions of polyA intensity by automated quantitation with CellProfiler (v4.2.1). Hoechst signal was used to segment nuclei and the oligo-dT signal to segment cell body. Three differentiation replicates per genotype. 3D surface plots were made in Fiji.

WGA inhibition of nuclear export

Confluent NPCs were incubated with digitonin at 30 μg/mL diluted in DMSO and WGA conjugated to Alexa Fluor 488 (Invitrogen, W11261) at 5 μg/mL diluted in DPBS for 5 minutes, as previously described (Mor *et al.*, 2010). Cells were washed to remove digitonin and WGA only was added to media at 5 μg/mL for 1 hour. Control NPCs were only treated with digitonin. Cells were fixed and prepped for oligo-dT FISH as described above.

RNA sequencing and bioinformatics analysis

Total RNA was extracted from cultured hNPCs (two biological replicates per genotype) using TRIzol Reagent (Invitrogen 15596026) followed by DNAse column treatment using PureLink RNA extraction kit (Invitrogen 12183018A). Total RNA from dissected E9.5

mouse forebrain tissue (three biological replicates per genotype) was extracted using Picopure RNA isolation kit (Applied Biosystems) according to manufacturer's recommendations. hNPC and E9.5 mouse forebrain RNA samples were ribo-depleted followed by 151 bp paired-end sequencing on the Illumina NovaSeq 300 cycle, ~20-30 million reads per sample. Library preparation and sequencing was conducted by the Advanced Genomics Core (AGC) at the University of Michigan. Invitrogen ERCC spike-ins (#4456740) were added for sequencing controls at starting concentrations according to the manufacturer's instructions. FASTQ files were trimmed with Cutadapt v4.1 using default parameters. Read quality was assessed by FASTQC v0.11.9. MultiQC version 1.7 was used to visualize FASTQC outputs and compare samples. ERCC spike-in FASTA and GTF annotation files were merged with human GRCh38.p13 reference genome FASTA with GTF release 39 or mouse GRCm39 reference genome FASTA with GTF release M28. FASTQ reads were then mapped to merged files using STAR alignment with parameter '--outSAMtype BAM SortedByCoordinate' (Dobin et al., 2013). Count analysis was performed on sorted BAM files using RSEM with paired-end alignment specified (Li and Dewey, 2011). Differential expression analysis was carried out using DESeq2 v1.34.0 (Love et al., 2014) in R v4.1.2 (R Core Team, 2020). ERCC spike-in counts were used to estimate size factors for each sample for DESEq2 analysis. Genes were considered dysregulated if FDR < 0.05 and fold-change > 2 or < -2. Volcano and PCA plots were made using ggplot2 and pcaExplorer packages in R.

Alternative splicing analysis was performed on sorted BAM files using rMATS v4.1.2 (Park et al., 2013) with the following parameters: '-t paired --readLength 150 --variable-read-length --nthread 4' (Shen *et al.*, 2014). AS events were called if FDR < 0.05

and ΔPSI > 10%. Events with less than 5 average reads were filtered out using the MASER package in R (Veiga, 2022). To calculate splice site strength at 5' and 3' splice sites in AS transcripts identified by rMATS, maximum entropy modeling was carried out using MaxEntScan (Yeo and Burge, 2004). The required input is a 9-mer sequence at 5' splice sites (3 bases in exon and 6 bases in downstream intron) and a 23-mer at 3' splice site (20 bases of intron and 3 bases of downstream exon). Scores were plotted in GraphPad Prism (v9.0.0).

DAVID (david.ncifcrf.gov/tools) and Metascape (metascape.org) (Zhou et al., 2019) analyses were performed to identify enriched biological pathways based on Benjamini-Hochberg multiple hypothesis corrections of the *p*-values. To identify potential transcription factors responsible for expression differences, Gene Set Enrichment Analysis (GSEA v4.2.3) against the MSigDB transcription factor motif gene set (c4.tftv7.5.1.symbols.gmt) and ChIP-X Enrichment Analysis v3 (ChEA3) were performed (Keenan et al., 2019). Ensembl BioMart tool (http://useast.ensembl.org/biomart) was used to obtain coding sequence length, transcript number per gene, gene type, and sequences for AS events. The GeneOverlap v1.32 R package was used to identify overlapping DE and AS hits between affected genotypes. Primary and candidate syndromic ID genes were obtained from the SysID database (https://www.sysid.dbmr.unibe.ch).

qRT-PCR and mRNA half-live analysis

Reverse transcription for cDNA synthesis was performed using 1 μg of total RNA with Invitrogen Superscript III kit (18080051) according to manufacturer's instructions. For validation of AS events, standard PCR was performed as described above. Abundances of *THOC6*, *FOS*, *GAPDH*, *MEG3*, *MEG8*, *ESRG*, and *NEAT1* mRNA was determined by quantitative real-time PCR (qPCR) using the Applied Biosystems 7500 system with 7500 Software v2.3 and Radiant Green 2X qPCR Mix Lo-ROX 2X qPCR Mix (Alkali Scientific inc., QS1020) according to manufacturer's instructions. Cycler parameters used: cDNA activation (1 cycle at 95°C for 2 minutes), denaturation (40 cycles 95°C for 5 seconds) and annealing/extension (40 cycles at 60°C for 30 seconds). The ΔΔCt method was used to analyze data with *GAPDH* as a reference gene. ΔΔCt values obtained by subtracting mean *THOC6*$^{+/+}$ ΔCt values for each sample. Data shown represent mean values of three qPCR technical replicates per sample for three biological replicates per genotype (independent differentiations for NPCs). Melt curve analysis was performed on all primers to ensure temperature peaks at ~80-90°C. *GAPDH* and *FOS* primer sequences were obtained from (Moon *et al.*, 2012). *NEAT1* was obtained from (Cui *et al.*, 2019) and *MEG3* was obtained from (Mondal *et al.*, 2015). All others were designed using NCBI primer blast.

mRNA decay analysis was performed using transcription inhibition by Actinomycin D (ActD) based on (Moon *et al.*, 2012). Human ESCs/iPSCs were first passaged into five 12-well plates. Each plate had the following lines: *THOC6*$^{+/+}$ (H9 ESCs), *THOC6*$^{+/+}$ (AS0041 iPSCs), *THOC6*$^{W100*/W100*}$, *THOC6*$^{W100*/+}$, *THOC6*$^{E188K/E188K}$, *THOC6*$^{E188K/+}$. Once confluent, ActD was added to media of all four plates at 10 μg/mL (Sigma-Aldrich A9415). After 30 minutes, media was removed from one plate and 1 mL of TRIzol Reagent was added directly to each well (*t* = 0). Cells were uplifted in TRIzol by pipetting and transferred to a fresh tube. Tubes were immediately frozen in TRIzol at -80°C. This was repeated every 30 minutes to obtain the following time points 30 minutes post-ActD treatment: t = 0.5, 1, 1.5, and 2 hrs. Extractions were performed in batches per time point based on protocol described above. Standard curve analysis was performed to validate primers (Figure S5.1C). This experiment was repeated to capture longer decay window using the following time points: t = 0, 2, 4, 8 (Figure S5.1D). ΔΔCt values obtained by subtracting mean t = 0 ΔCt for each genotype. Abundances of *THOC6*, *GAPDH*, and *FOS* (positive control for rapid decay) mRNA was determined.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical significance of all quantifications from microscopy images, western blot images, gel images, and qRT-PCR abundances was tested using a student's two-tailed *t*-test and data was plotted using GraphPad Prism (v9.3.1) as mean ±SEM or mean ±SD, as specified in figure legends. Simple linear regression was performed in qRT-PCR standard curve analysis, organoid growth curves, and intron retention analysis. Statistical significance of gene overlaps were tested using Fisher's exact test via GeneOverlap package function testGeneOverlap() in R. Benjamini-Hochberg multiple hypothesis corrections were performed in pathway enrichment analyses.

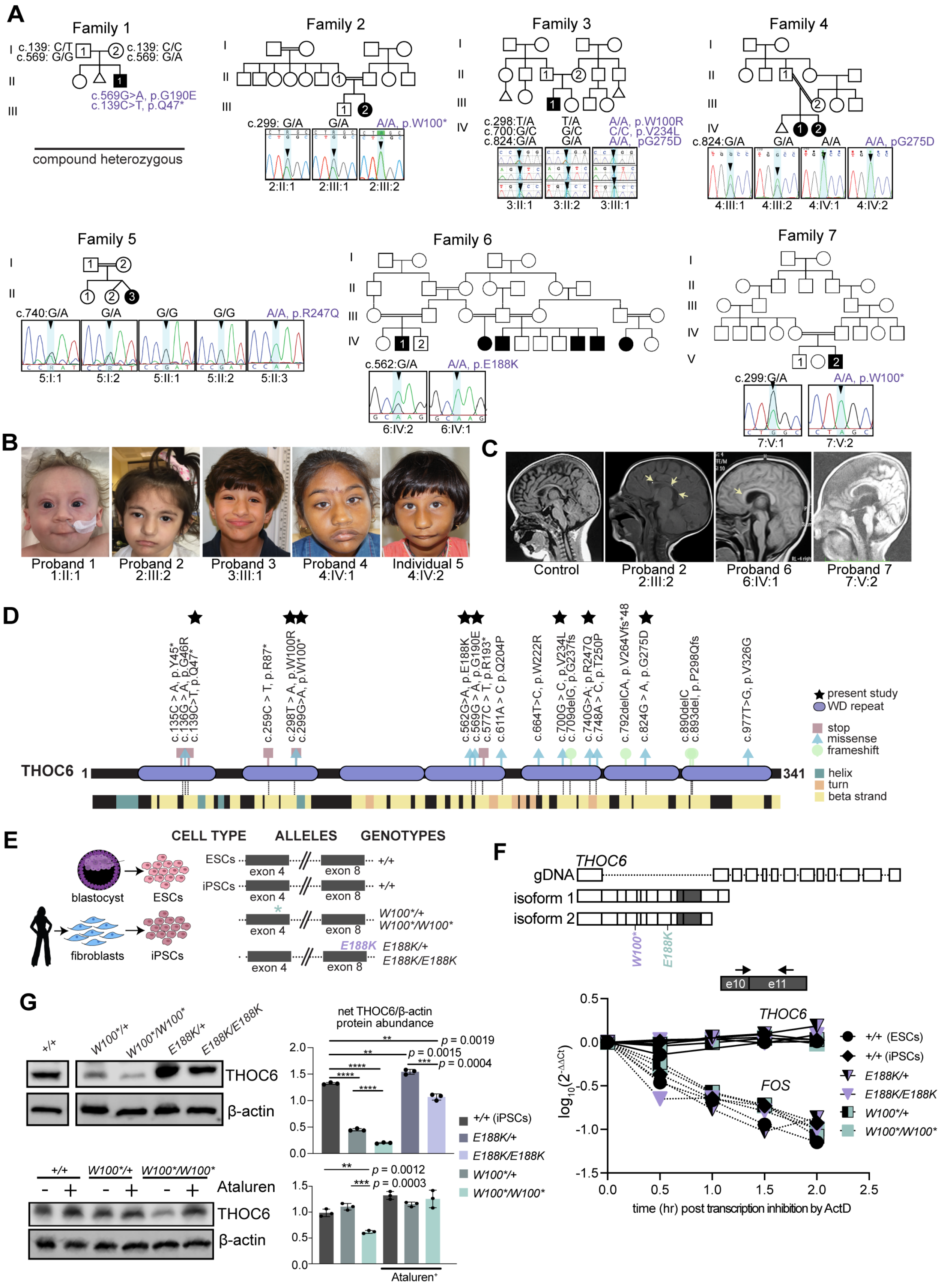
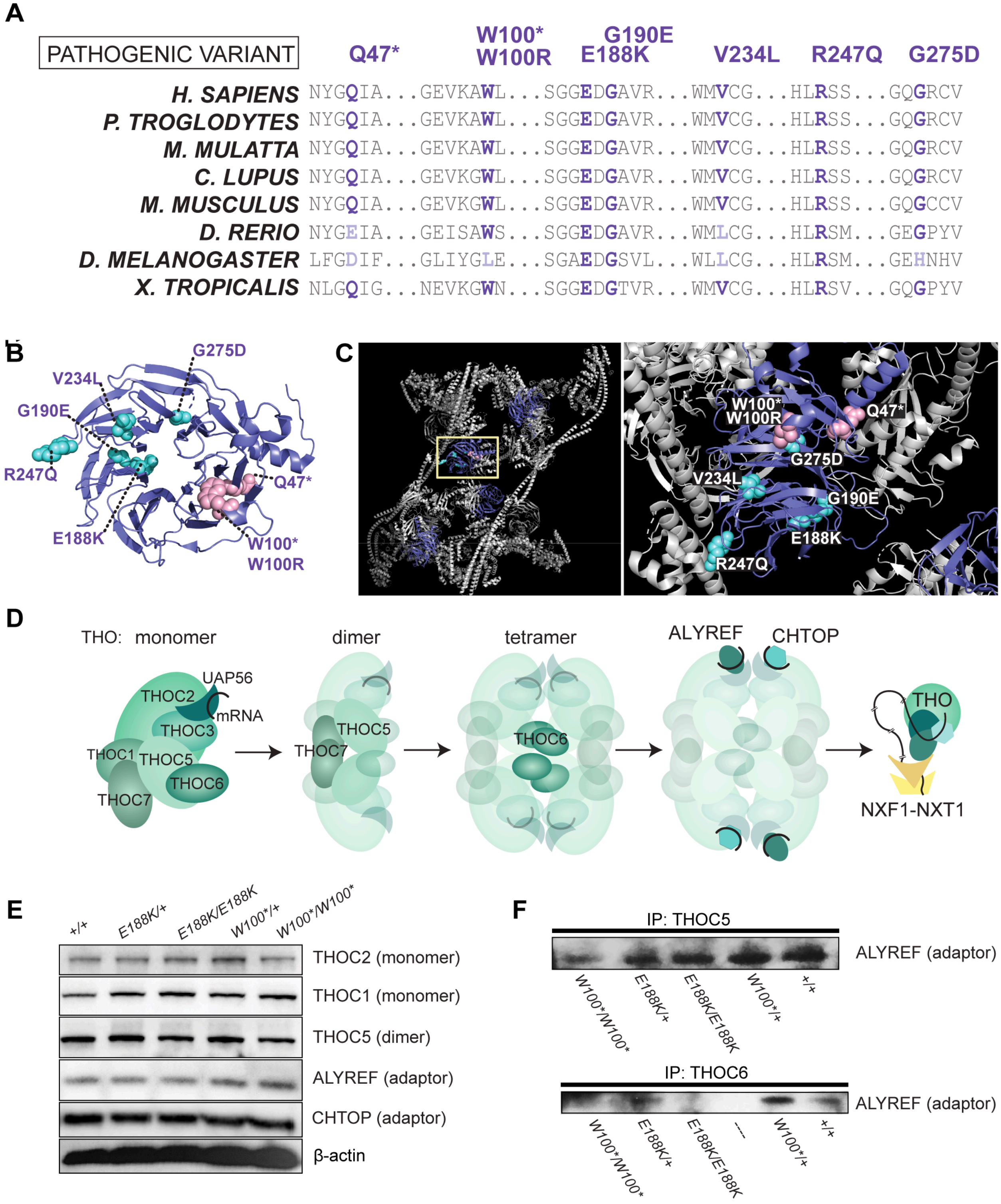

A
Family 1
I c.139: C/T c.569: G/G
c.139: C/C c.569: G/A
II
III
c.569G>A, p.G190E
c.139C>T, p.Q47*
compound heterozygous
Family 2
c.299: G/A
G/A
A/A, p.W100*
2:II:1
2:III:1
2:III:2
Family 3
c.298:T/A c.700:G/C c.824:G/A
T/A G/C G/A
A/A, p.W100R C/C, p.V234L A/A, pG275D
3:II:1
3:II:2
3:III:1
Family 4
c.824:G/A
G/A
A/A
A/A, pG275D
4:III:1
4:III:2
4:IV:1
4:IV:2
Family 5
c.740:G/A
G/A
G/G
G/G
A/A, p.R247Q
5:I:1
5:I:2
5:II:1
5:II:2
5:II:3
Family 6
c.562:G/A
A/A, p.E188K
6:IV:2
6:IV:1
Family 7
c.299:G/A
A/A, p.W100*
7:V:1
7:V:2
B
Proband 1 1:II:1
Proband 2 2:III:2
Proband 3 3:III:1
Proband 4 4:IV:1
Individual 5 4:IV:2
C
Control
Proband 2 2:III:2
Proband 6 6:IV:1
Proband 7 7:V:2
D
c.135C > A, p.Y45*
c.136G > A, p.G46R
c.139C>T, p.Q47*
c.259C > T, p.R87*
c.298T > A, p.W100R
c.299G>A, p.W100*
c.562G>A, p.E188K
c.569G > A, p.G190E
c.577C > T, p.R193*
c.611A > C p.Q204P
c.664T>C, p.W222R
c.700G > C, p.V234L
c.709delG, p.G237fs
c.740G>A, p.R247Q
c.748A > C, p.T250P
c.792delCA, p.V264Vfs*48
c.824G > A, p.G275D
c.890delC
c.893del, p.P298Qfs
c.977T>G, p.V326G
THOC6 1
341
★ present study
WD repeat
stop
missense
frameshift
helix
turn
beta strand
E
CELL TYPE
ALLELES
GENOTYPES
blastocyst
ESCs
fibroblasts
iPSCs
ESCs
iPSCs
exon 4
exon 8
+/+
+/+
W100*/+
W100*/W100*
E188K
E188K/+
E188K/E188K
F
THOC6
gDNA
isoform 1
isoform 2
W100*
E188K
e10
e11
THOC6
FOS
log10(2^-ΔΔCt)
0.5
0.0
-0.5
-1.0
-1.5
0.0
0.5
1.0
1.5
2.0
2.5
time (hr) post transcription inhibition by ActD
+/+ (ESCs)
+/+ (iPSCs)
E188K/+
E188K/E188K
W100*/+
W100*/W100*
G
+/+
W100*/+
W100*/W100*
E188K/+
E188K/E188K
THOC6
β-actin
net THOC6/β-actin protein abundance
p = 0.0019
p = 0.0015
p = 0.0004
2.0
1.5
1.0
0.5
0.0
+/+ (iPSCs)
E188K/+
E188K/E188K
W100*/+
W100*/W100*
+/+
W100*/+
W100*/W100*
Ataluren
THOC6
β-actin
p = 0.0012
p = 0.0003
Ataluren+

Figure 5.1. Biallelic pathogenic variants in *THOC6* cause syndromic intellectual disability.
(A) Pedigree drawings of segregating TIDS phenotypes in families 1-7, with generations listed on the left-hand side. Females are represented as circles and males are denoted by squares. Miscarriages are denoted by small triangles. Affected family members are indicated by solid black coloring while unaffected are unfilled. Consanguineous partnerships are represented by double lines. Chromatograms from Sanger sequencing of THOC6 confirmation of genotypes are provided for each tested family member in families 1-7. (B) Facial photographs of Probands 1-4 and Individual 5. (C) Sagittal brain MRI showing corpus callosum dysgenesis (Probands 2, 6, & 7) and cortical and cerebellar atrophy (Proband 6) compared to control (left). (D) Canonical THOC6 protein map consisting of 341 amino acids. WD40 repeat domains 1-7 are denoted by purple rectangles. Location of known pathogenic variants are annotated relative to linear protein map (top) and secondary structure (below). Variants reported in present study are distinguished by a black star. Missense (blue triangle), nonsense (red square), and frameshift (green circle). (E) Schematic of patient and control-derived human cell types and respective genotypes. (F) Decay of *THOC6* mRNA (solid line) following ActD transcriptional inhibition compared to *FOS* mRNA decay (dotted line) in human ESC/iPSCs across genotypes. Values calculated relative to *GAPDH* reference mRNA. (G) Western blot of human ESC/iPSCs indicating reduced THOC6 protein expression in $THOC6^{W100*/W100*}$ iPSCs compared to unaffected controls. Confirmation of readthrough by ataluren treatment (30 μM). Abundance quantifications relative to β-actin control (right). Data represented as mean ±SEM. P-value, two-tailed unpaired *t* test. ****, p-value <0.0001.


A
PATHOGENIC VARIANT
Q47*
W100*
W100R
G190E
E188K
V234L
R247Q
G275D
H. SAPIENS NYGQIA...GEVKAWL...SGGEDGAVR...WMVCG...HLRSS...GQGRCV
P. TROGLODYTES NYGQIA...GEVKAWL...SGGEDGAVR...WMVCG...HLRSS...GQGRCV
M. MULATTA NYGQIA...GEVKAWL...SGGEDGAVR...WMVCG...HLRSS...GQGRCV
C. LUPUS NYGQIA...GEVKGWL...SGGEDGAVR...WMVCG...HLRSS...GQGRCV
M. MUSCULUS NYGQIA...GEVKGWL...SGGEDGAVR...WMVCG...HLRSS...GQGCCV
D. RERIO NYGEIA...GEISAWS...SGGEDGAVR...WMLCG...HLRSM...GEGPYV
D. MELANOGASTER LFGDIF...GLIYGLE...SGAEDGSVL...WLLCG...HLRSM...GEHNHV
X. TROPICALIS NLGQIG...NEVKGWN...SGGEDGTVR...WMVCG...HLRSV...GQGPYV
B
G275D
V234L
G190E
R247Q
Q47*
E188K
W100*
W100R
C
W100*
W100R
Q47*
G275D
V234L
G190E
E188K
R247Q
D
THO: monomer
UAP56
THOC2
mRNA
THOC3
THOC1
THOC5
THOC6
THOC7
dimer
THOC5
THOC7
tetramer
THOC6
ALYREF
CHTOP
THO
NXF1-NXT1
E
+/+
E188K/+
E188K/E188K
W100*/+
W100*/W100*
THOC2 (monomer)
THOC1 (monomer)
THOC5 (dimer)
ALYREF (adaptor)
CHTOP (adaptor)
β-actin
F
IP: THOC5
ALYREF (adaptor)
W100*/W100*
E188K/+
E188K/E188K
W100*/+
+/+
IP: THOC6
ALYREF (adaptor)
W100*/W100*
E188K/+
E188K/E188K

W100*/+
+/+

Figure 5.2. Genetic mechanism of biallelic pathogenic *THOC6* variants.
(A) Amino acid alignment showing conservation of affected residues for pathogenic variants in present clinical study. Variants mapped to THOC6 folded β-propeller structure (B) and THO/TREX complex (C). (D) Schematic of TREX core tetrameric assembly mediated by THOC6 with functional implications for mRNA processing and export based on published crystal structure (Pühringer *et al.*, 2020). (E) Steady-state protein abundance for THO/TREX complex members and (F) ALYREF abundance following co-immunoprecipitation with THOC5 and THOC6 across genotypes.

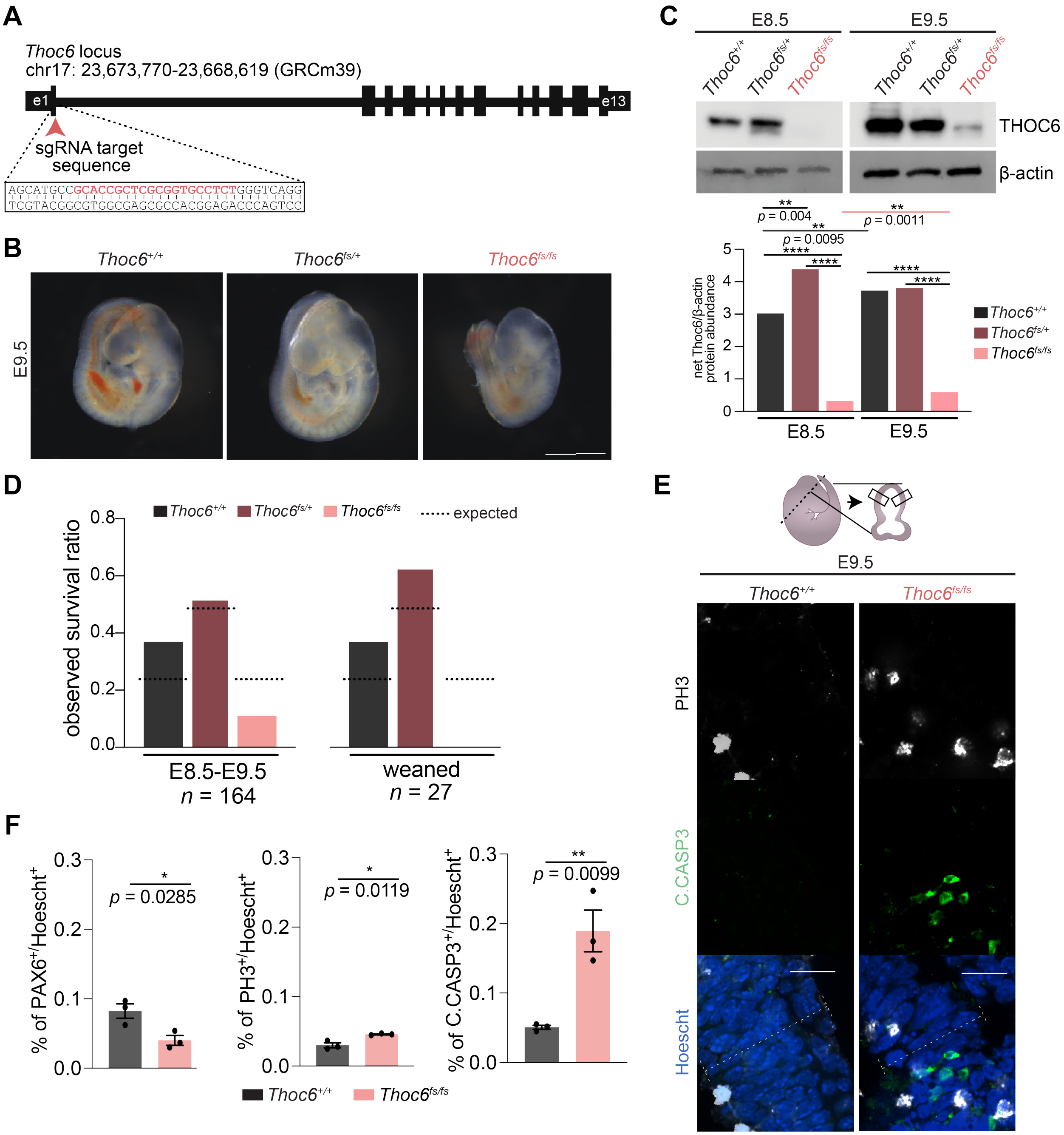


Figure 5.3. Generation of *Thoc6*$^{fs/fs}$ mouse model.

(A) CRISPR/Cas9 editing strategy to introduce frameshift variants in mouse *Thoc6*. (B) Representative images of isolated *Thoc6*$^{+/+}$, *Thoc6*$^{fs/+}$, and *Thoc6*$^{fs/fs}$ whole embryos at E9.5 prior. Scale bar: 50 μm. (C) Western blot analysis with quantifications of E8.5 and E9.5 mouse embryos showing increased expression during development. Ablation of THOC6 protein in *Thoc6*$^{fs/fs}$ is observed at E8.5, with presence of band suggesting read-through product at E9.5. β-actin, loading control. (D) Litter ratio analysis for E8.5-9.5 (left) and weaned (right) *Thoc6*$^{fs/fs}$ mice. Ratios are consistent with embryonic lethality of homozygous frameshift mice. n = 164 mice (left); n = 27 mice (right). (E) Immunostaining of markers PH3 and C.CASP3 in E9.5 mouse forebrain. Illustration highlights sectioning and quantification approach. (F) Quantifications of fractions of PAX6, PH3, and C.CASP3-expressing cells in E9.5 neuroepithelium. Measurements were combined from one rostral and one caudal section (from two lateral segments depicted by solid black boxes in E) per three embryo replicates per genotype. Data shown as mean ±SEM. Significance, two-tailed *t* test.

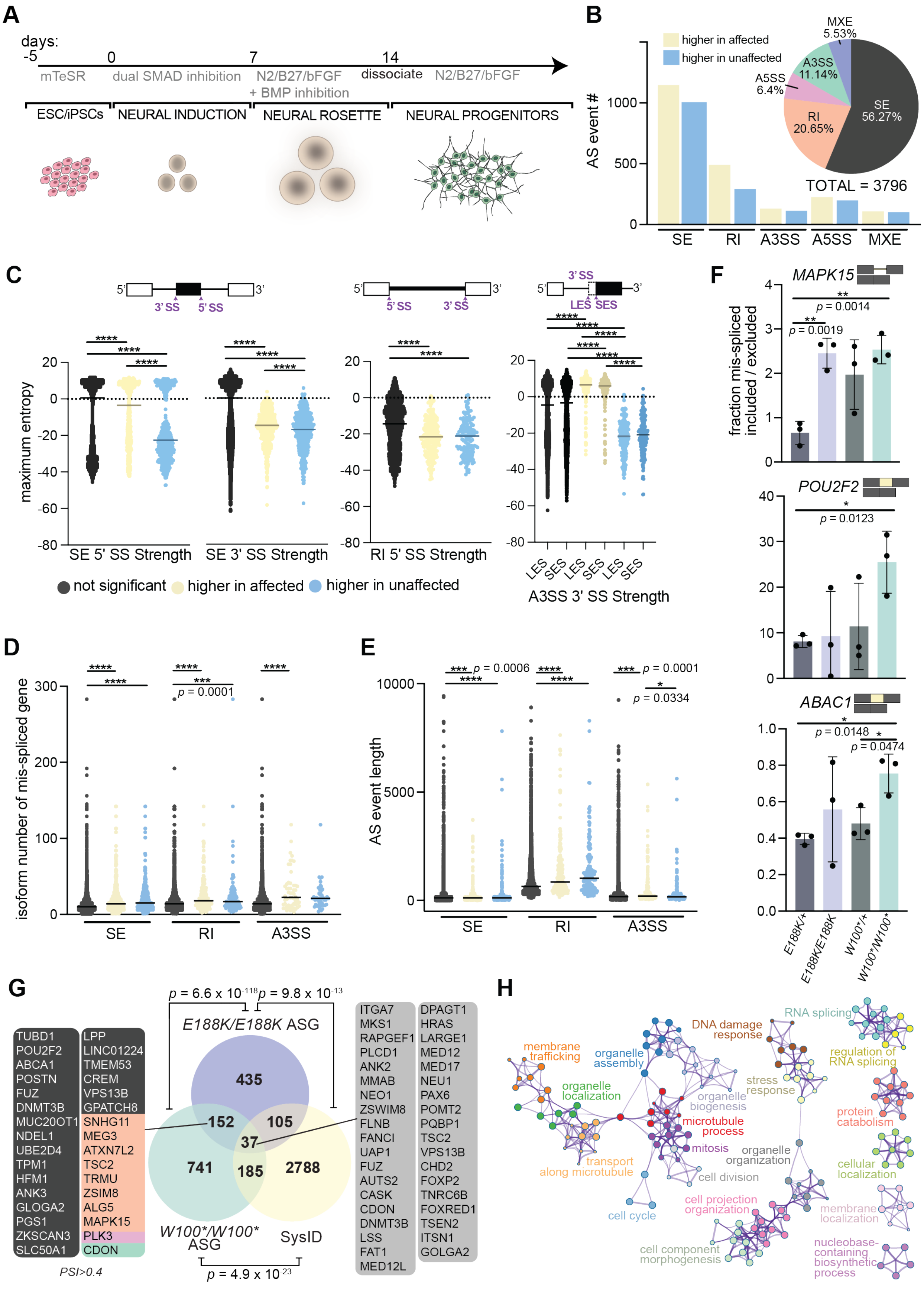

A
days:
-5
0
7
14
mTeSR
dual SMAD inhibition
N2/B27/bFGF
+ BMP inhibition
dissociate
N2/B27/bFGF
ESC/iPSCs
NEURAL INDUCTION
NEURAL ROSETTE
NEURAL PROGENITORS
B
higher in affected
higher in unaffected
AS event #
0
500
1000
SE
RI
A3SS
A5SS
MXE
MXE 5.53%
A3SS 11.14%
A5SS 6.4%
RI 20.65%
SE 56.27%
TOTAL = 3796
C
5'
3'
3' SS
5' SS
LES
SES
maximum entropy
20
0
-20
-40
-60
-80
****
***
SE 5' SS Strength
SE 3' SS Strength
RI 5' SS Strength
LES
SES
A3SS 3' SS Strength
● not significant ● higher in affected ● higher in unaffected
D
isoform number of mis-spliced gene
0
100
200
300
****
***
p = 0.0001
SE
RI
A3SS
E
AS event length
0
5000
10000
*** p = 0.0006
****
*** p = 0.0001
*
p = 0.0334
SE
RI
A3SS
F
MAPK15
fraction mis-spliced
included / excluded
**
p = 0.0014
**
p = 0.0019
0
1
2
3
4
POU2F2
*
p = 0.0123
0
10
20
30
40
ABAC1
*
p = 0.0148
*
p = 0.0474
0.0
0.2
0.4
0.6
0.8
1.0
E188K/+
E188K/E188K
W100*/+
W100*/W100*
G
TUBD1
POU2F2
ABCA1
POSTN
FUZ
DNMT3B
MUC20OT1
NDEL1
UBE2D4
TPM1
HFM1
ANK3
GLOGA2
PGS1
ZKSCAN3
SLC50A1
LPP
LINC01224
TMEM53
CREM
VPS13B
GPATCH8
SNHG11
MEG3
ATXN7L2
TSC2
TRMU
ZSIM8
ALG5
MAPK15
PLK3
CDON
PSI>0.4
p = 6.6 x 10^-118
p = 9.8 x 10^-13
E188K/E188K ASG
435
152
105
37
741
185
2788
W100*/W100*
ASG
SysID
p = 4.9 x 10^-23
ITGA7
MKS1
RAPGEF1
PLCD1
ANK2
MMAB
NEO1
ZSWIM8
FLNB
FANCI
UAP1
FUZ
AUTS2
CASK
CDON
DNMT3B
LSS
FAT1
MED12L
DPAGT1
HRAS
LARGE1
MED12
MED17
NEU1
PAX6
POMT2
PQBP1
TSC2
VPS13B
CHD2
FOXP2
TNRC6B
FOXRED1
TSEN2
ITSN1
GOLGA2
H
membrane trafficking
organelle assembly
DNA damage response
RNA splicing
regulation of RNA splicing
stress response
organelle localization
organelle biogenesis
microtubule process
protein catabolism
mitosis
organelle organization
cellular localization
transport along microtubule
cell division
cell cycle
cell projection organization
membrane localization
cell component morphogenesis
nucleobase-containing biosynthetic process

Figure 5.4. Characterization of alternative splicing events in *THOC6* affected hNPCs.
(A) Differentiation protocol to derive human neural progenitor cells from affected and unaffected ESC/iPSCs. (B) Combined rMATS summary results for AS events in *THOC6*$^{W100*/W100*}$ and *THOC6*$^{E188K/E188K}$ hNPCs relative to *THOC6*$^{W100*/+}$ control hNPCs. Event type (pie chart) and inclusion status (bar chart). Yellow, higher inclusion in affected. Blue, higher inclusion in unaffected. (C) Significant splice site strength score differences at mis-spliced events in affected hNPCs based on maximum entropy model. (D) transcript number per AS gene and (E) AS event length in *THOC6*$^{W100*/W100}$ and *THOC6*$^{E188K/E188K}$ vs. *THOC6*$^{W100*/+}$ NPCs. (F) RT-PCR AS validations of SE (*ABAC1*, *POU2F2*) and RI (*MAPK15*) events in three additional biological replicates of hNPCs per genotype with quantified mis-spliced ratios. Data shown as mean ±SEM. P-values (C-F), two-tailed unpaired *t* test. ****, *p* = < 0.0001. (G) Venn diagram of overlap of *THOC6* $^{W100*/W100*}$ and *THOC6*$^{E188K/E188K}$ AS genes and all syndromic intellectual disability genes included in the SysID database. Overlap significance tested by Fisher's exact test. ASG, alternatively spliced genes. Metascape analysis on combined significant mis-spliced events (FDR <0.05) in *THOC6*$^{E188K/E188K}$ and *THOC6*$^{W100*/W100*}$ NPCs (H).

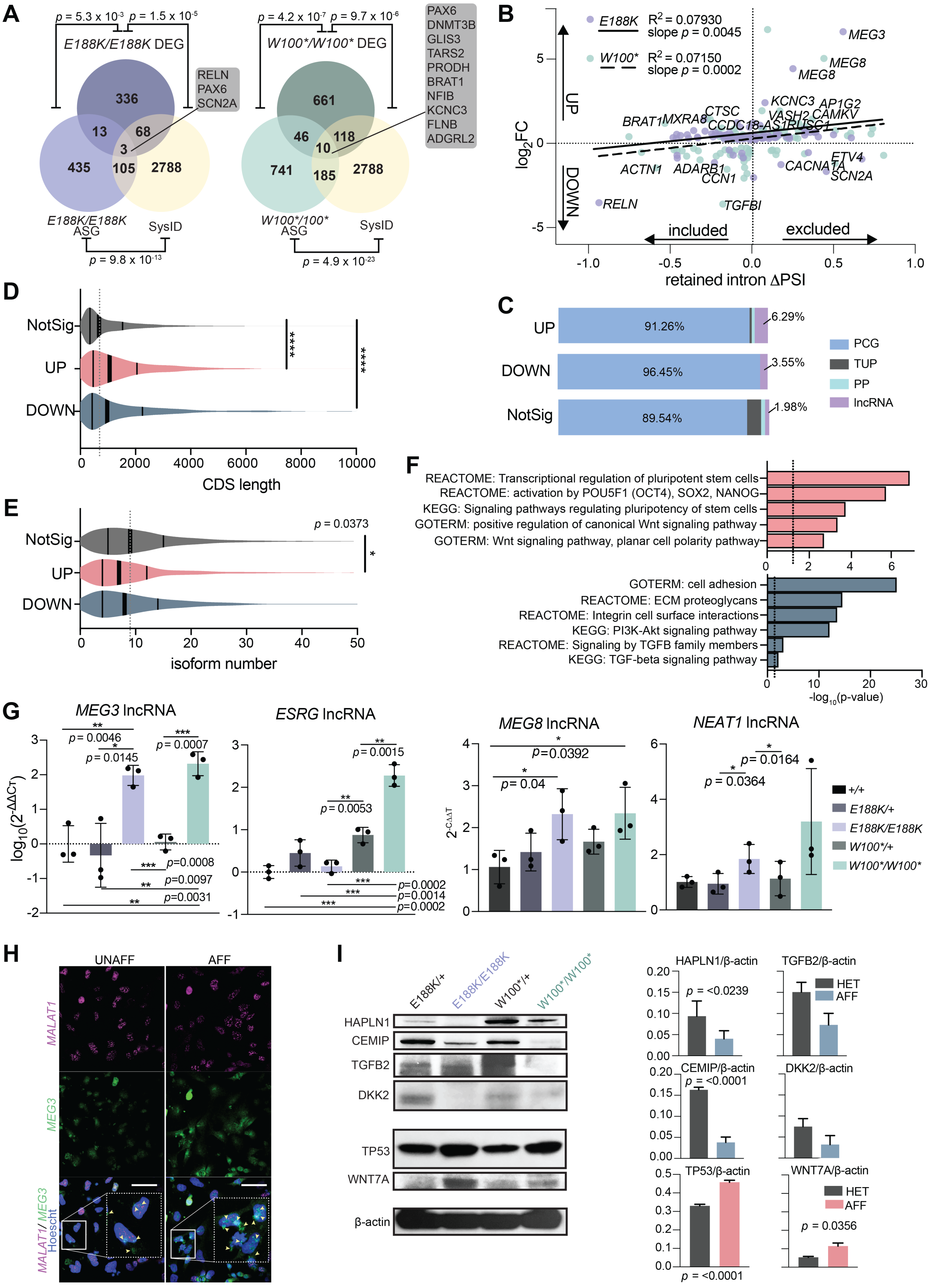

A
p = 5.3 x 10^-3
p = 1.5 x 10^-5
E188K/E188K DEG
336
13
68
3
435
105
2788
RELN
PAX6
SCN2A
E188K/E188K ASG
SysID
p = 9.8 x 10^-13
p = 4.2 x 10^-7
p = 9.7 x 10^-6
W100*/W100* DEG
661
46
118
10
741
185
2788
PAX6
DNMT3B
GLIS3
TARS2
PRODH
BRAT1
NFIB
KCNC3
FLNB
ADGRL2
W100*/100* ASG
SysID
p = 4.9 x 10^-23
B
E188K R^2 = 0.07930 slope p = 0.0045
W100* R^2 = 0.07150 slope p = 0.0002
MEG3
MEG8
MEG8
KCNC3
AP1G2
VASH2
CAMKV
CTSC
BRAT1
MXRA8
CCDC18-AS1
RUSC1
ACTN1
ADARB1
CCN1
CACNA1A
ETV4
SCN2A
RELN
TGFBI
UP
DOWN
log2FC
5
0
-5
included
excluded
-1.0
-0.5
0.0
0.5
1.0
retained intron ΔPSI
D
NotSig
UP
DOWN
****
****
0
2000
4000
6000
8000
10000
CDS length
C
UP
DOWN
NotSig
91.26%
96.45%
89.54%
6.29%
3.55%
1.98%
PCG
TUP
PP
lncRNA
E
NotSig
UP
DOWN
p = 0.0373
*
0
10
20
30
40
50
isoform number
F
REACTOME: Transcriptional regulation of pluripotent stem cells
REACTOME: activation by POU5F1 (OCT4), SOX2, NANOG
KEGG: Signaling pathways regulating pluripotency of stem cells
GOTERM: positive regulation of canonical Wnt signaling pathway
GOTERM: Wnt signaling pathway, planar cell polarity pathway
0
2
4
6
GOTERM: cell adhesion
REACTOME: ECM proteoglycans
REACTOME: Integrin cell surface interactions
KEGG: PI3K-Akt signaling pathway
REACTOME: Signaling by TGFB family members
KEGG: TGF-beta signaling pathway
0
10
20
30
-log10(p-value)
G
MEG3 lncRNA
log10(2^-ΔΔCT)
**
p= 0.0046
*
p= 0.0145
***
p= 0.0007
***
p=0.0008
**
p=0.0097
**
p=0.0031
ESRG lncRNA
**
p= 0.0015
**
p= 0.0053
***
p=0.0002
***
p=0.0014
***
p=0.0002
MEG8 lncRNA
2^-CΔΔT
*
p=0.0392
*
p= 0.04
NEAT1 lncRNA
*
p= 0.0164
*
p = 0.0364
+/+
E188K/+
E188K/E188K
W100*/+
W100*/W100*
H
UNAFF
AFF
MALAT1
MEG3
MALAT1/ MEG3 Hoescht
I
E188K/+
E188K/E188K
W100*/+
W100*/W100*
HAPLN1
CEMIP
TGFB2
DKK2
TP53
WNT7A
β-actin
HAPLN1/β-actin
p = <0.0239
TGFB2/β-actin
HET
AFF
CEMIP/β-actin
p = <0.0001
DKK2/β-actin
TP53/β-actin
p = <0.0001
WNT7A/β-actin
HET
AFF
p = 0.0356

Figure 5.5. Differential expression analysis in affected hNPCs.

(A) Venn diagram of gene overlap of *THOC6* $^{W100*/W100*}$ and *THOC6*$^{E188K/E188K}$ affected genes and all syndromic intellectual disability genes included in the SysID database. Overlap significance tested by Fisher's exact test. DEG, differentially expressed genes; ASG, alternatively spliced genes. (B) Linear regression analysis of $\log_2$foldchange and Δ percent transcripts spliced in (PSI) for significant retained intron events in affected cells. Purple dots indicate *THOC6*$^{E188K/E188K}$ hits and green dots indicate *THOC6* $^{W100*/W100*}$. Best fit line, R^2, and slope p-value for *THOC6*$^{E188K/E188K}$ (solid line) and *THOC6* $^{W100*/W100*}$ (dotted line). (C) Percentage of gene type by condition for combined DEGs in affected hNPCs. NotSig, not significant; Up, upregulated. Down, downregulated; lncRNA, long non-coding RNA; PP, processed pseudogene; TUP, transcribed unprocessed pseudogene; PCG, protein coding gene. Violin plots of coding sequence (CDS) length (D) and isoform number (E) of combined DEGs in affected cells compared to non-significant genes. (F) DAVID biological pathway enrichment analysis of combined upregulated genes (top, red) and downregulated genes (bottom, blue) in *THOC6* affected hNPCs. (G) qPCR relative abundance quantifications ($2^{-\Delta\Delta Ct}$) for *MEG3*, *ESRG*, *MEG8*, and *NEAT1* in hNPCs*.* Three technical replicates of three biological replicates per genotype. (H) RNA FISH probing for *MEG3* and *MALAT1* in affected and unaffected hNPCs. Cell inset showing *MEG3* expression and localization differences with yellow arrows in merged image. Scale bar 50 μm. (I) Protein abundance of top downregulated and upregulated genes across genotypes. Genes labeled with $\log_2$FC > 1 or < -1 and PSI > 0.1 or < -0.1. Data shown as mean ±SEM. Significance, two-tailed unpaired *t* test.

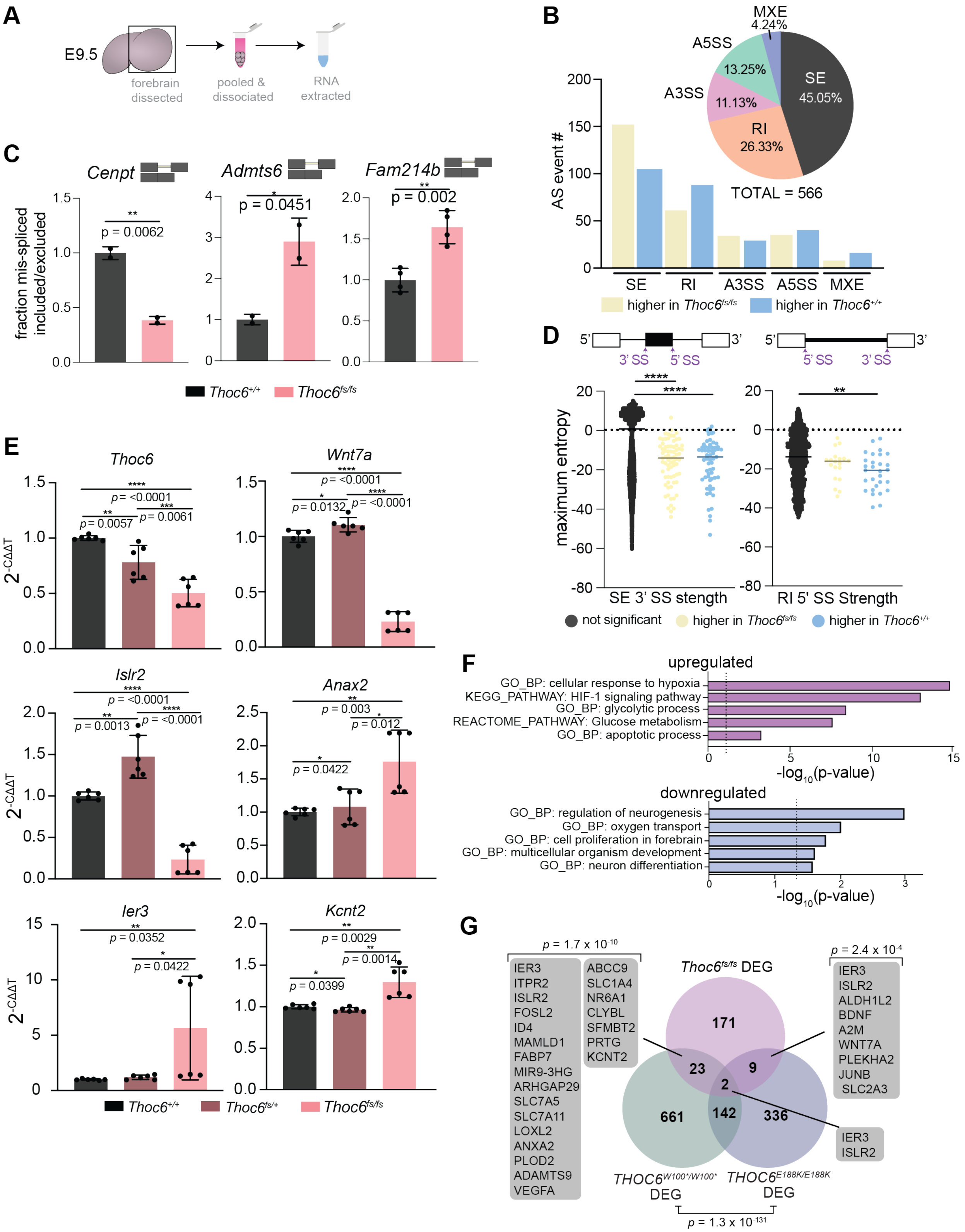

A
E9.5
forebrain dissected
pooled & dissociated
RNA extracted
B
MXE 4.24%
A5SS 13.25%
A3SS 11.13%
RI 26.33%
SE 45.05%
TOTAL = 566
AS event #
200
150
100
50
0
SE
RI
A3SS
A5SS
MXE
higher in Thoc6fs/fs
higher in Thoc6+/+
C
Cenpt
Admts6
Fam214b
fraction mis-spliced included/excluded
1.5
1.0
0.5
0.0
**
p = 0.0062
4
3
2
1
0
*
p = 0.0451
2.0
1.5
1.0
0.5
0.0
**
p = 0.002
Thoc6+/+
Thoc6fs/fs
D
5'
3'
3' SS
5' SS
maximum entropy
20
0
-20
-40
-60
-80
****
****
**
SE 3' SS stength
RI 5' SS Strength
not significant
higher in Thoc6fs/fs
higher in Thoc6+/+
E
Thoc6
2-CΔΔT
1.5
1.0
0.5
0.0
****
p= <0.0001
**
p= 0.0057
***
p= 0.0061
Wnt7a
****
p= <0.0001
*
p= 0.0132
****
p= <0.0001
Islr2
2.0
****
p= <0.0001
**
p= 0.0013
****
p= <0.0001
Anax2
2.5
**
p= 0.003
*
p= 0.012
*
p= 0.0422
Ier3
15
10
5
0
**
p= 0.0352
*
p= 0.0422
Kcnt2
**
p= 0.0029
**
p= 0.0014
*
p= 0.0399
Thoc6+/+
Thoc6fs/+
Thoc6fs/fs
F
upregulated
GO_BP: cellular response to hypoxia
KEGG_PATHWAY: HIF-1 signaling pathway
GO_BP: glycolytic process
REACTOME_PATHWAY: Glucose metabolism
GO_BP: apoptotic process
0
5
10
15
-log10(p-value)
downregulated
GO_BP: regulation of neurogenesis
GO_BP: oxygen transport
GO_BP: cell proliferation in forebrain
GO_BP: multicellular organism development
GO_BP: neuron differentiation
0
1
2
3
-log10(p-value)
G
p = 1.7 x 10-10
IER3
ITPR2
ISLR2
FOSL2
ID4
MAMLD1
FABP7
MIR9-3HG
ARHGAP29
SLC7A5
SLC7A11
LOXL2
ANXA2
PLOD2
ADAMTS9
VEGFA
ABCC9
SLC1A4
NR6A1
CLYBL
SFMBT2
PRTG
KCNT2
Thoc6fs/fs DEG
171
23
9
2
661
142
336
THOC6W100*/W100* DEG
THOC6E188K/E188K DEG
p = 1.3 x 10-131
p = 2.4 x 10-4
IER3
ISLR2
ALDH1L2
BDNF
A2M
WNT7A
PLEKHA2
JUNB
SLC2A3
IER3
ISLR2

Figure 5.6. Characterization of mRNA processing defects in *Thoc6$^{fs/fs}$* mouse E9.5 forebrain.
(A) Cartoon of E9.5 mouse forebrain total RNA sample preparation. (B) rMATS summary results for AS events in *Thoc6$^{fs/fs}$* E9.5 forebrain. Event type (pie chart) and inclusion status (bar chart). Yellow, higher inclusion in *Thoc6$^{fs/fs}$*. Blue, higher inclusion in *Thoc6$^{+/+}$*. (C) Quantifications from RT-PCR validating top AS events *Cenpt*, *Admts6*, *Fam214b* in 2-4 biological replicates. Significance, two-tailed unpaired *t* test. Data are mean ±SEM. (D) Significant splice site strength score differences at mis-spliced events in *Thoc6$^{fs/fs}$* samples based on maximum entropy model. (E) RT-qPCR validations of *Thoc6*, *Wnt7a*, *Islr2*, *Ier3*, *Kcnt2*, *Anax2* mRNA abundance on two additional biological replicates of E9.5 forebrain per genotype. Three technical replicates analyzed per sample. Significance, two-tailed unpaired *t* test. Data are mean ±SEM. (F) DAVID analysis showing significantly enriched biological pathways among upregulated genes (top, magenta) and downregulated genes (bottom, blue) in *Thoc6$^{fs/fs}$* E9.5 forebrain. (G) Venn diagram of overlap of DEG in affected hNPCs and *Thoc6$^{fs/fs}$* mouse E9.5 forebrain. Overlap significance tested by Fisher's exact test.

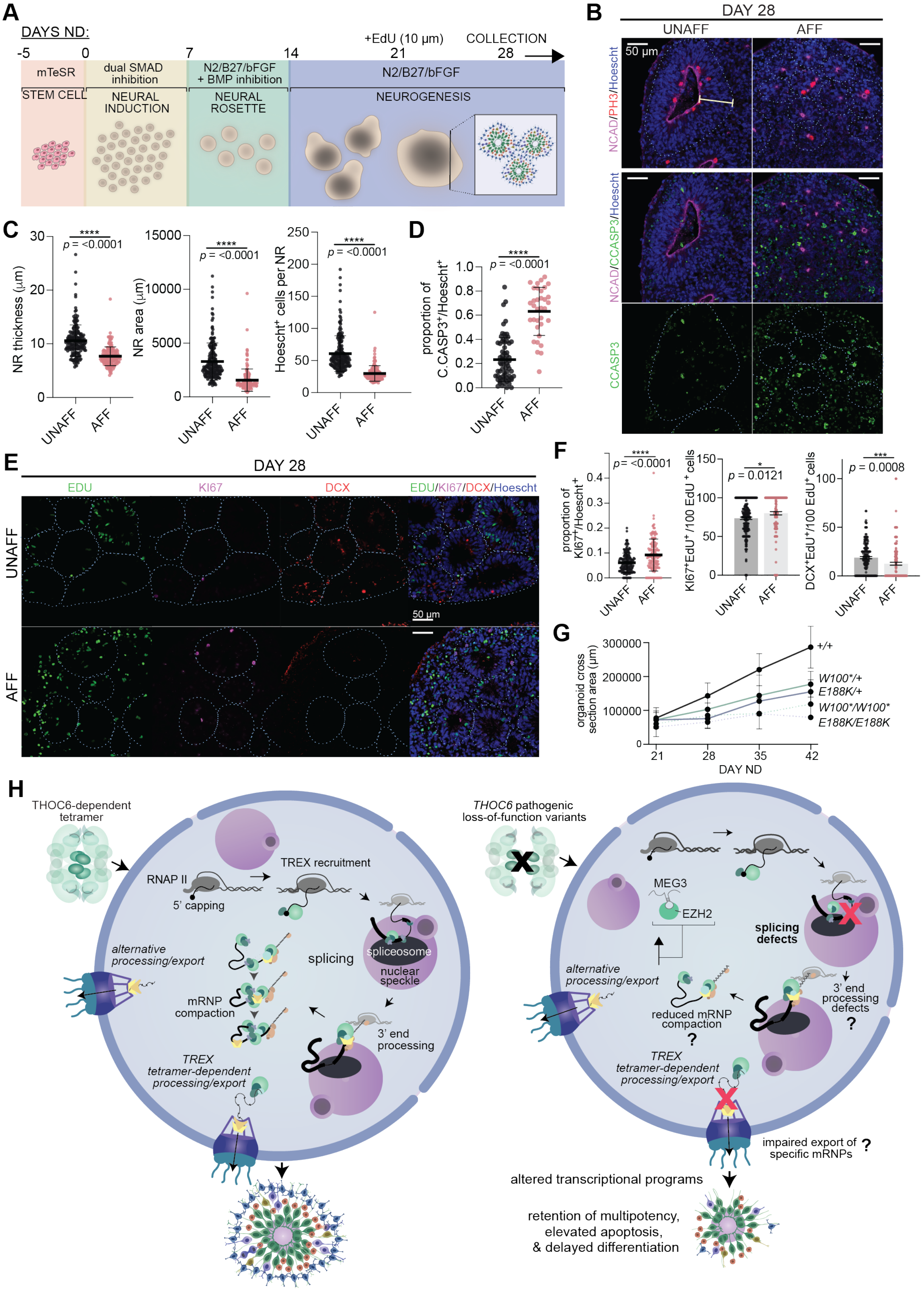
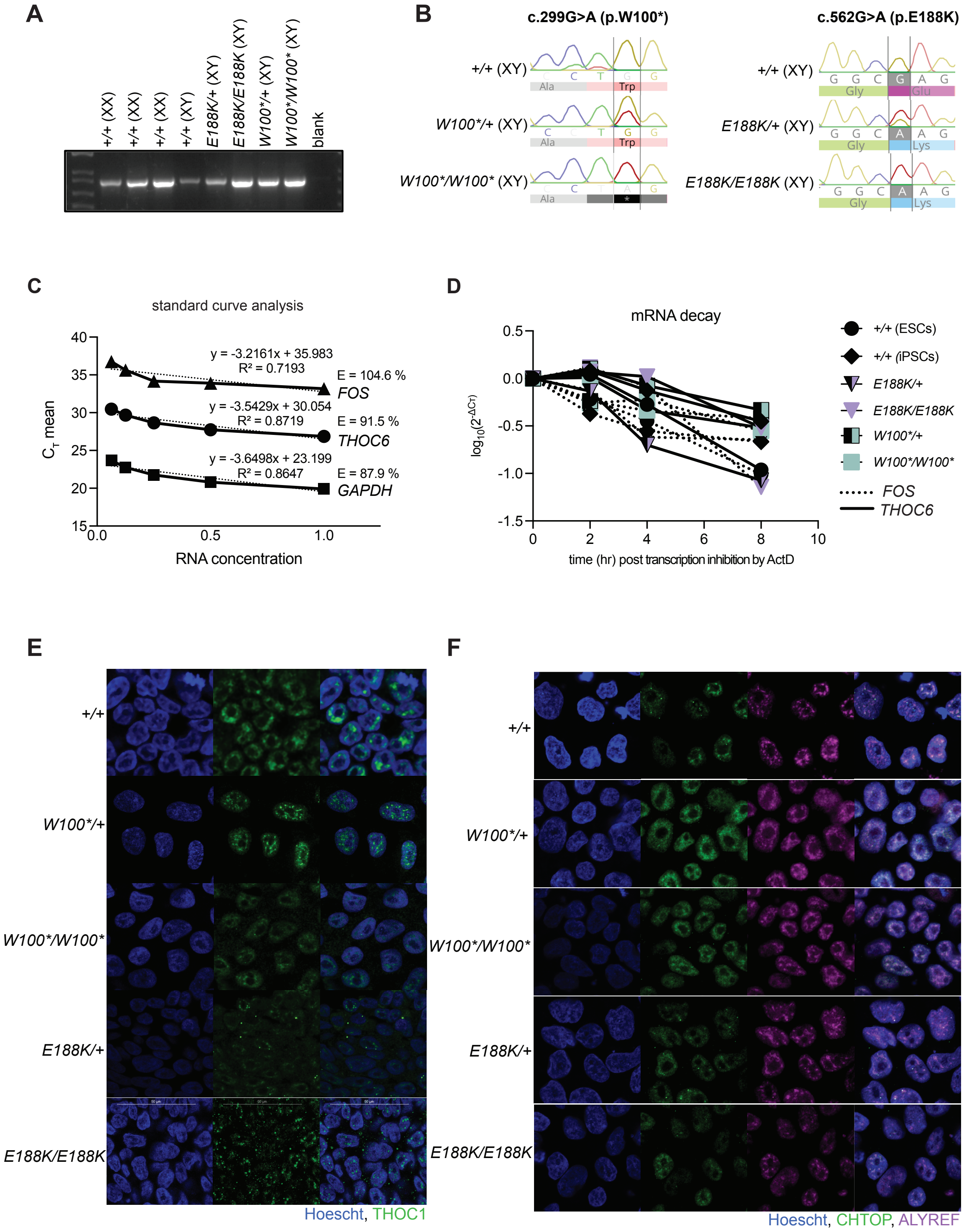

A
DAYS ND:
-5
0
7
14
+EdU (10 μm)
21
COLLECTION
28
mTeSR
dual SMAD inhibition
N2/B27/bFGF + BMP inhibition
N2/B27/bFGF
STEM CELL
NEURAL INDUCTION
NEURAL ROSETTE
NEUROGENESIS
B
DAY 28
UNAFF
AFF
50 μm
NCAD/PH3/Hoescht
NCAD/CCASP3/Hoescht
CCASP3
C
NR thickness (μm)
NR area (μm)
Hoescht+ cells per NR
****
p = <0.0001
UNAFF
AFF
D
proportion of C.CASP3+/Hoescht+
****
p = <0.0001
E
DAY 28
EDU
KI67
DCX
EDU/KI67/DCX/Hoescht
UNAFF
AFF
50 μm
F
proportion of KI67+/Hoescht+
p = <0.0001
KI67+EdU+/100 EdU+ cells
p = 0.0121
DCX+EdU+/100 EdU+ cells
p = 0.0008
G
organoid cross section area (μm)
DAY ND
+/+
W100*/+
E188K/+
W100*/W100*
E188K/E188K
H
THOC6-dependent tetramer
RNAP II
5' capping
TREX recruitment
spliceosome
nuclear speckle
splicing
alternative processing/export
mRNP compaction
3' end processing
TREX tetramer-dependent processing/export
THOC6 pathogenic loss-of-function variants
MEG3
EZH2
splicing defects
3' end processing defects
?
reduced mRNP compaction
?
impaired export of specific mRNPs ?
altered transcriptional programs
retention of multipotency, elevated apoptosis, & delayed differentiation

Figure 5.7. Modeling of *THOC6* variant pathogenesis in human cerebral organoids.
(A) Cerebral organoid differentiation protocol. ND, neural differentiation. (B) Immunostaining of PH3, N-Cadherin, apoptosis marker cleaved caspase3 (CCASP3), and Hoescht in day 28 human cerebral organoids differentiated from unaffected and affected iPSCs, highlighting differences in neural rosette morphology. 40x magnification; Scale bar: 50 µm. (C-D) Quantifications of area, thickness, Hoescht+ cells, and fraction of CCASP3+ cells per NR for THOC6W100*/+ and THOC6E188K/+ controls and THOC6W100*/W100* and THOC6E188K/E188K affected organoids. NR (organoid) number analyzed across one differentiation replicate per genotype: unaffected, n = 67 (15); affected, n = 34 (10). (E) Immunostaining of EDU, KI67, DCX to assess timing of differentiation in day 28 organoids with quantifications (F). NR (organoid) number analyzed across three differentiation replicates per genotype: unaffected, n = 187 (87); affected, n = 157 (67). (G) Growth rate of organoids across genotypes measured by cross section area (µm) from days 21-42. (C-D, F-G) Data shown as mean ±SEM. Significance, two-tailed unpaired t test. (H) Schematic of proposed model of THOC6 pathogenesis.



A
+/+ (XX)
+/+ (XX)
+/+ (XX)
+/+ (XY)
E188K/+ (XY)
E188K/E188K (XY)
W100*/+ (XY)
W100*/W100* (XY)
blank
B
c.299G>A (p.W100*)
+/+ (XY)
W100*/+ (XY)
W100*/W100* (XY)
Ala
Trp
*
c.562G>A (p.E188K)
+/+ (XY)
E188K/+ (XY)
E188K/E188K (XY)
Gly
Glu
Lys
C
standard curve analysis
y = -3.2161x + 35.983
R² = 0.7193
E = 104.6 %
FOS
y = -3.5429x + 30.054
R² = 0.8719
E = 91.5 %
THOC6
y = -3.6498x + 23.199
R² = 0.8647
E = 87.9 %
GAPDH
C_T mean
RNA concentration
40
35
30
25
20
15
0.0
0.5
1.0
D
mRNA decay
log10(2^-ΔCT)
time (hr) post transcription inhibition by ActD
0.5
0.0
-0.5
-1.0
-1.5
0
2
4
6
8
10
+/+ (ESCs)
+/+ (iPSCs)
E188K/+
E188K/E188K
W100*/+
W100*/W100*
FOS
THOC6
E
+/+
W100*/+
W100*/W100*
E188K/+
E188K/E188K
Hoescht, THOC1
F
+/+
W100*/+
W100*/W100*
E188K/+
E188K/E188K
Hoescht, CHTOP, ALYREF

Figure S5.1. Genetic mechanism of biallelic pathogenic *THOC6* variants.
See also Figures 5.1 and 5.2. Confirmation of genotypes in human ESC/iPSC lines by PCR (A) and Sanger sequencing (B). (C) Standard curve analysis using 5 two-fold serial dilutions of control cDNA to confirm primer quality for FOS, THOC6, and GAPDH for mRNA stability assay. (D) mRNA decay curve for extended timeframe capturing THOC6 and FOS RNA decay after 2.5 hrs following Actinomysin D treatment. Values were not normalized to GAPDH because control transcripts are also degraded after 2.5 hrs of transcription inhibition. Immunostaining to assess subcellular localization of THO/TREX complex members THOC1 (E), and CHTOP and ALYREF (F) in human ESC/iPSCs with biallelic pathogenic THOC6 variants compared to heterozygous and wildtype unaffected controls.

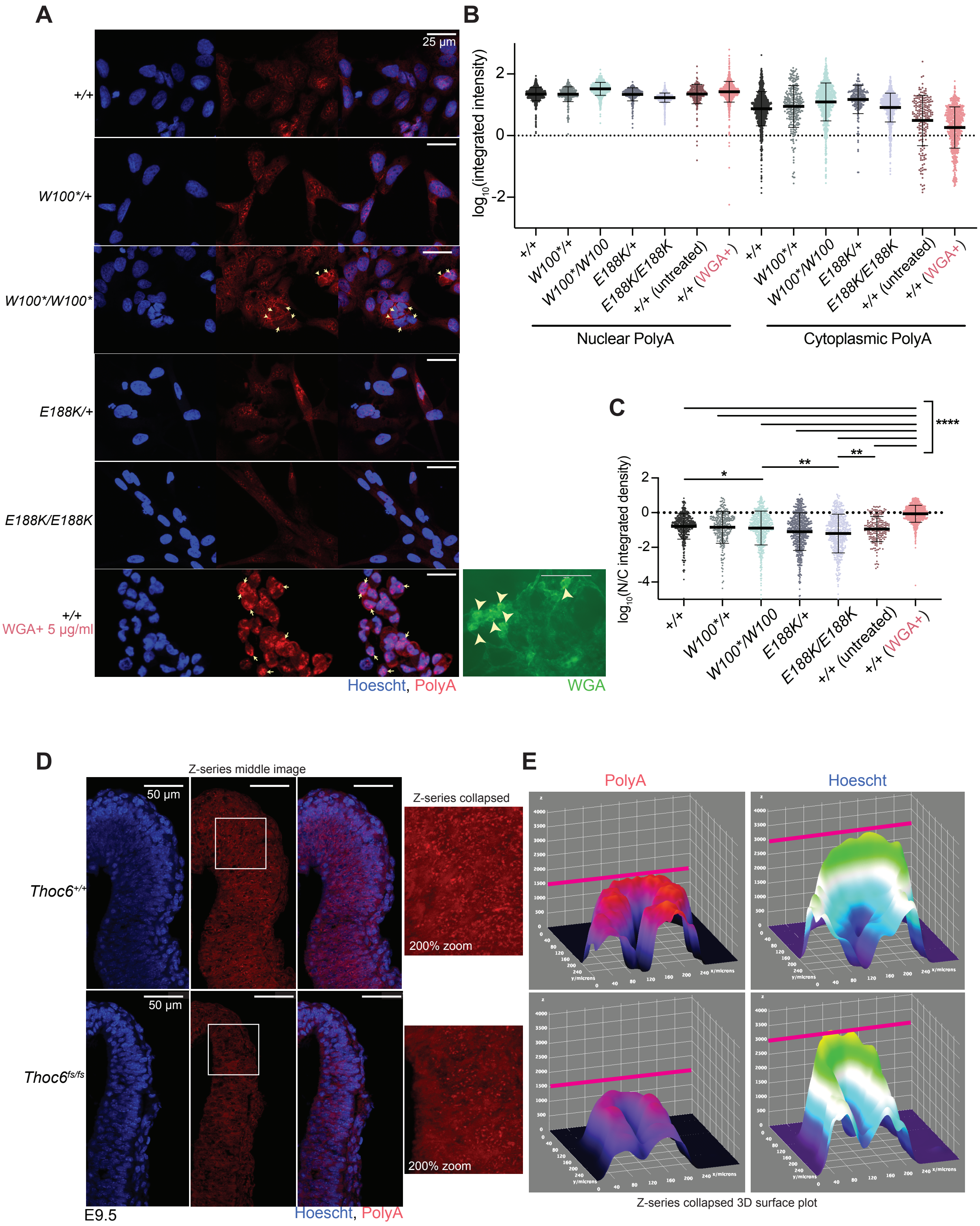

A
25 µm
+/+
W100*/+
W100*/W100*
E188K/+
E188K/E188K
+/+
WGA+ 5 µg/ml
Hoescht, PolyA
WGA
B
log10(integrated intensity)
2
0
-2
+/+
W100*/+
W100*/W100
E188K/+
E188K/E188K
+/+ (untreated)
+/+ (WGA+)
+/+
W100*/+
W100*/W100
E188K/+
E188K/E188K
+/+ (untreated)
+/+ (WGA+)
Nuclear PolyA
Cytoplasmic PolyA
C
****
*
**
**
log10(N/C integrated density)
2
0
-2
-4
+/+
W100*/+
W100*/W100
E188K/+
E188K/E188K
+/+ (untreated)
+/+ (WGA+)
D
Z-series middle image
Z-series collapsed
50 µm
50 µm
Thoc6+/+
Thoc6fs/fs
200% zoom
200% zoom
E9.5
Hoescht, PolyA
E
PolyA
Hoescht
Z-series collapsed 3D surface plot

Figure S5.2. Characterization of bulk mRNA export across genotypes.
Related to Figures 5.3-5.6). (A) Oligo-dT FISH Z-collapsed confocal images (40x magnification) of NPCs differentiated from iPSCs with the following genotypes: *THOC6*$^{+/+}$, *THOC6*$^{W100*/W100*}$, *THOC6*$^{W100*/+}$, *THOC6*$^{E188K/E188K}$, *THOC6*$^{E188K/+}$. Arrows point to cells showing extreme signal differences. As a positive control for impaired nuclear export resulting in aberrant accumulation of transcripts in the nucleus, *THOC6*$^{+/+}$ were treated with WGA at 5 µg/ml which acts to block the nuclear pore complex (bottom). Scale bar: 25 µm. (B) Intensity quantifications from 200-700 cells per genotype from three replicates were performed using an automated CellProfiler (v4.2.1) pipeline that measures polyA signal in nuclear and cytoplasmic fractions. Variability in polyA+ signal intensity observed across genotypes is likely due to slight technical variation in the assay across replicate slides. (C) Ratios of nuclear to cytoplasmic polyA showing minimal differences in bulk export across genotypes relative to WGA+ positive control. (D) Oligo-dT FISH of E9.5 *Thoc6*$^{+/+}$ and *Thoc6*$^{fs/fs}$ mouse neuroepithelium; Z-series middle image (left). Scale bar: 50 µm. White box represents zoomed in area. (Right) 200% zoom of polyA signal in Z-series collapsed image by maximum intensity. (E) 3D surface plot for Z-series collapsed image of polyA intensity (left) and Hoescht intensity (right) in *Thoc6*$^{+/+}$ E9.5 neuroepithelium (top) and *Thoc6*$^{fs/fs}$ E9.5 neuroepithelium (bottom).

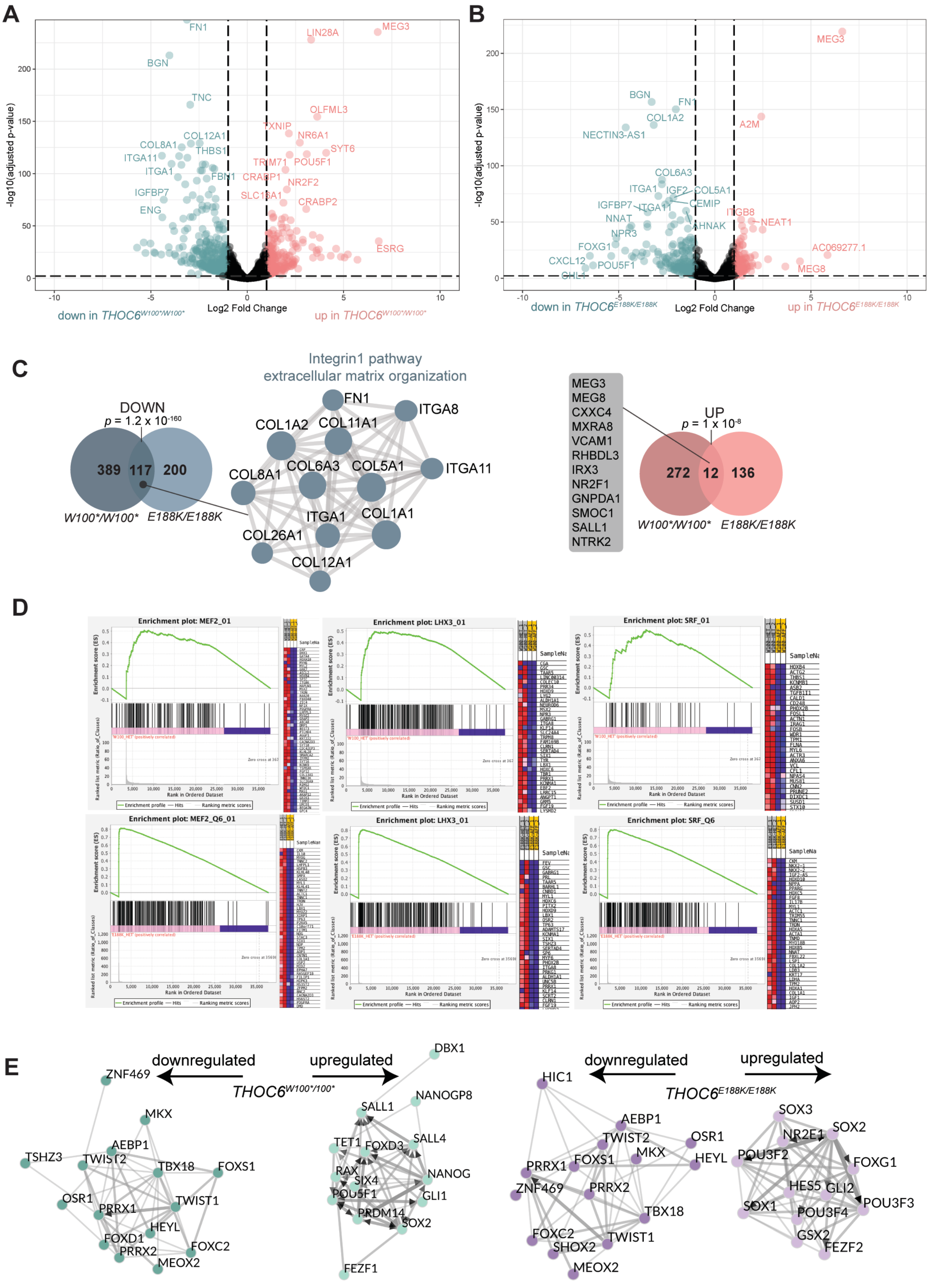

A
-log10(adjusted p-value)
Log2 Fold Change
down in *THOC6*$^{W100*/W100*}$
up in *THOC6*$^{W100*/W100*}$
FN1
BGN
TNC
COL12A1
COL8A1
THBS1
ITGA11
ITGA1
FBN1
IGFBP7
ENG
LIN28A
MEG3
OLFML3
TXNIP
NR6A1
SYT6
TRIM71
POU5F1
CRABP1
NR2F2
SLC18A1
CRABP2
ESRG
B
MEG3
BGN
FN1
COL1A2
A2M
NECTIN3-AS1
COL6A3
ITGA1
IGF2
COL5A1
IGFBP7
ITGA11
CEMIP
ITGB8
NNAT
AHNAK
NEAT1
NPR3
FOXG1
AC069277.1
CXCL12
POU5F1
MEG8
CHL1
down in *THOC6*$^{E188K/E188K}$
up in *THOC6*$^{E188K/E188K}$
C
DOWN
p = 1.2 x 10^{-160}
389
117
200
W100/W100**
E188K/E188K
Integrin1 pathway
extracellular matrix organization
FN1
ITGA8
COL1A2
COL11A1
COL6A3
COL5A1
ITGA11
COL8A1
ITGA1
COL1A1
COL26A1
COL12A1
MEG3
MEG8
CXXC4
MXRA8
VCAM1
RHBDL3
IRX3
NR2F1
GNPDA1
SMOC1
SALL1
NTRK2
UP
p = 1 x 10^{-8}
272
12
136
W100/W100**
E188K/E188K
D
Enrichment plot: MEF2_01
Enrichment plot: LHX3_01
Enrichment plot: SRF_01
Enrichment plot: MEF2_Q6_01
Enrichment plot: LHX3_01
Enrichment plot: SRF_Q6
Enrichment score (ES)
Ranked list metric (Ratio_of_Classes)
Rank in Ordered Dataset
Enrichment profile — Hits — Ranking metric scores
E
downregulated
upregulated
THOC6$^{W100*/100*}$
ZNF469
MKX
AEBP1
TSHZ3
TWIST2
TBX18
FOXS1
OSR1
PRRX1
TWIST1
HEYL
FOXD1
PRRX2
FOXC2
MEOX2
DBX1
NANOGP8
SALL1
SALL4
TET1
FOXD3
RAX
NANOG
SIX4
POU5F1
GLI1
PRDM14
SOX2
FEZF1
downregulated
upregulated
THOC6$^{E188K/E188K}$
HIC1
AEBP1
TWIST2
OSR1
MKX
HEYL
PRRX1
FOXS1
ZNF469
PRRX2
TBX18
FOXC2
SHOX2
TWIST1
MEOX2
SOX3
SOX2
NR2E1
POU3F2
FOXG1
HES5
GLI2
SOX1
POU3F4
POU3F3
GSX2
FEZF2

Figure S5.3. Differential expression analysis in affected hNPCs.
See also Figure 5.5. Volcano plot of differential expression in *THOC6*$^{W100*/W100*}$ (A) and *THOC6*$^{E188K/E188K}$ (B) NPCs relative to *THOC6*$^{W100*/+}$ controls. Data represent analysis of two biological replicates per genotype. The later comparison was chosen following observation of upregulated skeletal muscle genes in *THOC6*$^{E188K/+}$ hNPC replicates, indicating issues during the differentiation of this line. (C) Gene overlap of *THOC6*$^{W100*/W100*}$ and *THOC6*$^{E188K/E188K}$ downregulated (left, blue) and upregulated (right, red) genes. Metascape protein-protein network enrichment analysis identifies integrin1 pathway and extracellular matrix modules enriched among genes downregulated in both genotypes. (D) GSEA enrichment plots for top transcription factors using the transcription factor motif gene dataset (c4.tftv7.5.1.symbols.gmt) in *THOC6*$^{W100*/W100*}$ (top) *THOC6*$^{E188K/E188K}$ (bottom). (E) ChEA3 transcription factor motif enrichment of DEGs in *THOC6*$^{W100*/W100*}$ (left, green) and *THOC6*$^{E188K/E188K}$ (right, purple) NPCs.

Table 7. Clinical descriptions of individuals with TIDS in present study.
Related to Figures 5.1.

Proband	Clinical description	Genetic diagnosis	Variant Interpretation[a]
1:II:1	Proband 1 is from the Netherlands and of European ancestry. He was twelve months old at provided evaluation. The pregnancy was complicated by intrauterine growth restriction at thirty weeks of gestation. Proband 1 was born via a spontaneous vaginal delivery after 38 weeks and 3 days of gestation with a birth weight was 2.252 kg, and a length of 47 cm. Feeding problems were noted after birth. Gastroesophageal reflux disease was diagnosed requiring nasogastric tube feeding intervention. Growth parameters at four months of age included a length of 58 cm (<3rd centile), with a weight of 4.275 kg (<3rd centile) and an occipitofrontal circumference (OFC) of 38 cm (<3rd centile). Dysmorphic features include plagiocephaly, upslanting/narrow palpebral fissures, a straight nose with a broad nasal base, a smooth philtrum with a thin upper lip, and protruding ears with large, upturned earlobes (Figures 5.1B). A renal ultrasound revealed unilateral renal agenesis, though cardiac abnormalities were not detected by ultrasound evaluation. Marked signs of developmental delay at 12 months. Motor delays characterized by deficits in coordination required for prolonged prone positioning on his abdomen and inability to sit unassisted.	Karyotyping results were normal. Trio WES (mother/father/proband) was performed and identified two *THOC6* variants as high priority candidates for the clinical presentation observed for Proband 1. A previously described missense variant was maternally inherited and a novel truncating *THOC6* variant was paternally inherited: c.569G>A, (p.G190E) and c.139C>T, (p.Q47*) (Figures 5.1A). The c.139C>T, (p.Q47*) variant was not found in the gnomAD Browser (v2.1.1) and was submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/; Submission ID:SUB5265728). The parents of Proband 1 were heterozygous for the identified variants, consistent with a recessive mode of inheritance.	p.G190E: pathogenic (ii); p.Q47*: pathogenic (ia)
2:III:2	Proband 2 is the second-born of consanguineous parents of Turkish ancestry, three years old at provided evaluation (Figures 5.1A). Proband 2 was born full-term (41 weeks of gestation) with a birth weight of 2.7 kg (10th centile) and an OFC of 32.5 cm (<3rd centile). She exhibited global developmental delay, feeding problems, and persistent vomiting in the first few postnatal months, consistent with features of gastroesophageal reflux disease. Use of a nasoduodenal feeding tube supported slow, yet persistent weight gain. Physical examination revealed microcephaly and atypical facial features of epicanthus, long nose with low hanging columella, and upslanting palpebral fissures (Figures 5.1B). Cranial MRI revealed corpus callosum hypoplasia (Figures 5.1C). Frequent upper respiratory infections occurred one to three times per month. Echocardiogram demonstrated peripheral pulmonary stenosis and patent foramen ovale. Renal anatomy was unremarkable. At 3 years of age, weight was 13 kg (3rd centile), height was 95 cm (3rd centile) and OFC was 45 cm (<3rd centile). Severe intellectual disability was noted with delays in speech and communication requiring special rehabilitation intervention. Motor delays were observed. At time of evaluation, Proband 2 walked with an unsteady gait. While hearing and vision are normal, strabismus was observed. She has a high number of dental caries, in line with previously described BBIS dental anomalies.	Karyotyping results were normal. WES genetic testing identified a novel biallelic truncating variant c.299G>A, (p.W100*) in exon 4 of *THOC6* in Proband 2. At the time of identification, this variant was not reported and not observed in the gnomAD Browser (v2.1.1) and was submitted to ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/; Submission ID: SUB5265724). Heterozygosity of the *THOC6* variant was confirmed by Sanger sequencing in the mother, as well as in an unaffected sibling.	p.W100*: pathogenic (ia)
3:III:1	Proband 3 is from the United States of European ancestry, eleven years old at provided evaluation. He was evaluated for multiple congenital anomalies	Karyotype and FISH analysis for 22q11.2 were normal (Oxford Gene Technology Syndrome).	p.W100R, p.V234L, &

	during the newborn period. Microcephaly, deep set eyes, mild epicanthal folds, upslanting palpebral fissures were described during examination in infancy; prominent antihelices of the ears, broad nasal bridge, mild depression of the right nostril with evidence of right cleft lip repair, nasal columella extending below nares, and short philtrum were noted (Figures 5.1B). Clinical history at age eleven was significant for cleft lip, bifid uvula, ankyloglossia, horseshoe kidney, imperforate anus, developmental delay, autism, and partial complex seizures (onset at 11 years). Developmentally, receptive language skills were noted to be better than expressive language that was restricted to sign language, some words, and gesturing. Assistive technology was used to supplement this deficit. Reading and math skills at age eleven were measured to be equivalent to first grade level. On physical examination, both OFC and weight were less than the 2nd centile, and height was at the 5th centile.	Chromosome microarray (Affymetrix CytoScan Dx) identified a maternally inherited deletion on 5q21.1, the inheritance of which is inconsistent with genetic basis of BBIS clinical features. WES analysis of Proband 3 identified three previously described homozygous missense variants in *cis* in *THOC6* c.[298T>A;700G>C;824G>A], (p.[W100R;V234L;G275D]). This haplotype was previously reported in three other individuals with BBIS from unrelated families (6, 20). In all cases, no consanguinity was reported. The parents of Proband 3 are heterozygous for the identified variants, indicating biparental inheritance of the *THOC6* variants.	p.G275D: pathogenic (ii)
4:IV:1	Proband 4 was evaluated at thirteen years of age and is the first-born of third-degree consanguineous parents from Southern India (Figures 5.1A). She weighed 2.75 kg at birth and had complaints of repeated lower respiratory tract infections since the newborn period. Developmental delay, predominantly cognitive, was noted. She crawled at nine months, sat at one year and walked at two years. At present, she says bisyllables only. She exhibited friendly behavior. Examination revealed an OFC of 44.5 cm (<3rd centile, -7 SD) and a height of 122 cm (normal). Facial dysmorphism included upslanting palpebral fissures, epicanthal folds, microcornea, long nose with overhanging columella, thick vermillion borders, and crowded teeth (Figures 5.1B). Upon neurological examination, she had contractures at the ankle, normal to increased tone, and normal deep tendon reflexes. Vision and hearing were normal. Brain imaging results were also normal.	[see below]	p.G275D is likely pathogenic (ii)
4:IV:2	Individual 5 (4:IV:2) is the affected sibling of Proband 4 (Figures 5.1A). She was eight years old upon examination. She presented with global developmental delay like her elder sibling. She achieved head control at five months of age, sat with support at one year, and stood with and without support at one and two years, respectively. Presently, she follows simple commands. Parents also reported nocturnal enuresis. Her OFC was 44.5 cm (<3rd centile, -6 SD). She was 104 cm tall and weighed 12 kg. Epicanthal folds, microretrognathia, cleft palate, and U-shaped uvula were noted upon examination (Figures 5.1B). Vision and hearing were normal.	Karyotyping results were normal for both siblings. The biallelic missense variant c.824G>A, (p.G275D) in exon 12 of *THOC6* was identified in both Proband 4 and Individual 5 using WES testing. Variant p.G275D has previously been observed in individuals with BBIS and its submission accession is SCV000741884.1. The parents were heterozygous for the identified variants, as confirmed by Sanger sequencing and consistent with the mode of inheritance (Figures 5.1A).	p.G275D is likely pathogenic (ii)

5:II:3	Proband 5 was born to consanguineous parents of Moroccan ancestry. She has a healthy dizygotic twin sister as well as a healthy older sister and younger brother. She was born at 37 weeks of gestation with birth parameters of -2 SD (birth weight of 2.08 kg, birth length of 44 cm, and OFC of 31 cm). She presented short-segment Hirschsprung disease, submucous cleft palate, and unilateral choanal stenosis that was surgically repaired at 18 months. She has delayed psychomotor development and overall growth (parameters of -2 SD). She is shy with nasal speech limited to short sentences. At 10 years of age, she has a long narrow face, arched eyebrows, convergent strabismus, a tubular nose with a high nasal bridge, short columella, cupid bow-shaped mouth, and normal ears. Cutaneous 2-3 syndactyly on her feet and clinodactyly of the 5th fingers were also noted. No ophthalmologic abnormalities were present except for convergent strabismus. Auditory evoked potential (AEP), brain and temporal bones CT-scan, and cardiac ultrasound were normal.	800-bands resolution karyotype showed normal chromosomes on lymphocytes, 46XX, with no 22q11.2 deletion by FISH anailysis at the TUPPLE1 locus. WES analysis identified biallelic variants of c.740G>A, (p.R247Q). Variant was absent in unaffected siblings whereas unaffected parents were heterozygous for the variant. Results confirmed by Sanger sequencing and consistent with recessive mode of inheritance.	p.R247Q is likely pathogenic (ii)
6:IV:1	Proband 6 was 13 months old at last examination. He had an OFC of 42.5 cm (-3/4 SD) that is likely progressive. Proband 6 has dysmorphic facial features. Hypotonicity, no spasticity, and absent tendon reflexes were noted upon neurological examination. He has hypoplastic genitalia. MRI revealed atrophy of cortex and cerebellum as well as ventriculomegaly and a thin corpous callosum. Proband 6 had intermediate delayed psychomotor development.	WES analysis identified biallelic variants of c.562G>A, (p.E188K). Variant was absent in unaffected siblings whereas unaffected parents were heterozygous for the variant. Results confirmed by Sanger sequencing and consistent with recessive mode of inheritance.	p.E188K is likely pathogenic (ii)
7:V:1	Proband 7 had an OFC of -5 SD. He has epilepsy and micropenis.	WES analysis identified biallelic variants of c.299G>A, (p.W100*). Variant was absent in unaffected siblings whereas unaffected parents were heterozygous for the variant. Results confirmed by Sanger sequencing and consistent with recessive mode of inheritance.	p.W100*: pathogenic (ia)

[a] According to American College of Medical Genetics and Genomics (ACMG) guidelines

Table 8. **Clinical summary of all reported cases of TIDS.**
Absence of features may be due to lack of reporting. *Related to Figures 5.1.*

Clinical feature	**Prevalence**	
	current study	all published (excluding prenatal report)
Intellectual disability	8 / 8 (100%)	33 / 33 (100%)
Facial dysmorphisms	8 / 8 (100%)	31 / 33 (93.9%)
Microcephaly	8 / 8 (100%)	27 / 33 (81.8%)
Teeth anomalies	2 / 8 (25%)	15 / 33 (45.5 %)
Short stature	3 / 8 (37.5%)	13 / 33 (39.4%)
Cardiac defects	1 / 8 (12.5%)	12 / 33 (36.4%)
Renal malformations	3 / 8 (37.5%)	11 / 33 (33.3%)
Genitourinary issues	4 / 8 (50%)	21 / 33 (63.6%)
Feeding difficulties	2 / 8 (25%)	6 / 33 (18.2%)
Ventriculomegaly	1 / 8 (12.5%)	7 / 33 (21.2%)
ASD or autistic features	1 / 8 (12.5%)	8 / 33 (24.2%)

REFERENCES

Accogli, A., Scala, M., Calcagno, A., Castello, R., Torella, A., Musacchia, F., Allegri, A.M.E., Mancardi, M.M., Maghnie, M., Severino, M., et al. (2018). Novel CNS malformations and skeletal anomalies in a patient with Beaulieu-boycott-Innes syndrome. Am J Med Genet A *176*, 2835-2840. 10.1002/ajmg.a.40534.

Amos, J.S., Huang, L., Thevenon, J., Kariminedjad, A., Beaulieu, C.L., Masurel-Paulet, A., Najmabadi, H., Fattahi, Z., Beheshtian, M., Tonekaboni, S.H., et al. (2017). Autosomal recessive mutations in THOC6 cause intellectual disability: syndrome delineation requiring forward and reverse phenotyping. Clin Genet *91*, 92-99. 10.1111/cge.12793.

Anazi, S., Alshammari, M., Moneis, D., Abouelhoda, M., Ibrahim, N., and Alkuraya, F.S. (2016). Confirming the candidacy of THOC6 in the etiology of intellectual disability. Am J Med Genet A *170A*, 1367-1369. 10.1002/ajmg.a.37549.

Anazi, S., Maddirevula, S., Faqeih, E., Alsedairy, H., Alzahrani, F., Shamseldin, H.E., Patel, N., Hashem, M., Ibrahim, N., Abdulwahab, F., et al. (2017). Clinical genomics expands the morbid genome of intellectual disability and offers a high diagnostic yield. Mol Psychiatry *22*, 615-624. 10.1038/mp.2016.113.

Andrews, M.G., Subramanian, L., and Kriegstein, A.R. (2020). mTOR signaling regulates the morphology and migration of outer radial glia in developing human cortex. Elife *9*. 10.7554/eLife.58737.

Azam, S., Hou, S., Zhu, B., Wang, W., Hao, T., Bu, X., Khan, M., and Lei, H. (2019). Nuclear retention element recruits U1 snRNP components to restrain spliced lncRNAs in the nucleus. RNA Biol *16*, 1001-1009. 10.1080/15476286.2019.1620061.

Bahar Halpern, K., Caspi, I., Lemze, D., Levy, M., Landen, S., Elinav, E., Ulitsky, I., and Itzkovitz, S. (2015). Nuclear Retention of mRNA in Mammalian Tissues. Cell Rep *13*, 2653-2662. 10.1016/j.celrep.2015.11.036.

Beaulieu, C.L., Huang, L., Innes, A.M., Akimenko, M.A., Puffenberger, E.G., Schwartz, C., Jerry, P., Ober, C., Hegele, R.A., McLeod, D.R., et al. (2013). Intellectual disability associated with a homozygous missense mutation in THOC6. Orphanet J Rare Dis *8*, 62. 10.1186/1750-1172-8-62.

Boycott, K.M., Beaulieu, C., Puffenberger, E.G., McLeod, D.R., Parboosingh, J.S., and Innes, A.M. (2010). A novel autosomal recessive malformation syndrome associated with developmental delay and distinctive facies maps to 16ptel in the Hutterite population. Am J Med Genet A *152A*, 1349-1356. 10.1002/ajmg.a.33379.

Casey, J., Jenkinson, A., Magee, A., Ennis, S., Monavari, A., Green, A., Lynch, S.A., Crushell, E., and Hughes, J. (2016). Beaulieu-Boycott-Innes syndrome: an intellectual disability syndrome with characteristic facies. Clin Dysmorphol *25*, 146-151. 10.1097/MCD.0000000000000134.

Chai, G., Webb, A., Li, C., Antaki, D., Lee, S., Breuss, M.W., Lang, N., Stanley, V., Anzenberg, P., Yang, X., et al. (2021). Mutations in Spliceosomal Genes PPIL1 and PRP17 Cause Neurodegenerative Pontocerebellar Hypoplasia with Microcephaly. Neuron *109*, 241-256.e249. 10.1016/j.neuron.2020.10.035.

Chang, C.T., Hautbergue, G.M., Walsh, M.J., Viphakone, N., van Dijk, T.B., Philipsen, S., and Wilson, S.A. (2013). Chtop is a component of the dynamic TREX mRNA export complex. EMBO J *32*, 473-486. 10.1038/emboj.2012.342.

Chen, F.C., Chen, C.J., Ho, J.Y., and Chuang, T.J. (2006). Identification and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat. BMC Bioinformatics *7*, 136. 10.1186/1471-2105-7-136.

Cheng, H., Dufu, K., Lee, C.S., Hsu, J.L., Dias, A., and Reed, R. (2006). Human mRNA export machinery recruited to the 5' end of mRNA. Cell *127*, 1389-1400. 10.1016/j.cell.2006.10.044.

Chervitz, S.A., Hester, E.T., Ball, C.A., Dolinski, K., Dwight, S.S., Harris, M.A., Juvik, G., Malekian, A., Roberts, S., Roe, T., et al. (1999). Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. Nucleic Acids Res *27*, 74-78. 10.1093/nar/27.1.74.

Chi, B., Wang, Q., Wu, G., Tan, M., Wang, L., Shi, M., Chang, X., and Cheng, H. (2013). Aly and THO are required for assembly of the human TREX complex and association of TREX components with the spliced mRNA. Nucleic Acids Res *41*, 1294-1306. 10.1093/nar/gks1188.

Chávez, S., Beilharz, T., Rondón, A.G., Erdjument-Bromage, H., Tempst, P., Svejstrup, J.Q., Lithgow, T., and Aguilera, A. (2000). A protein complex containing Tho2, Hpr1, Mft1 and a novel protein, Thp2, connects transcription elongation with mitotic recombination in Saccharomyces cerevisiae. EMBO J *19*, 5824-5834. 10.1093/emboj/19.21.5824.

Cui, Y., Yin, Y., Xiao, Z., Zhao, Y., Chen, B., Yang, B., Xu, B., Song, H., Zou, Y., Ma, X., and Dai, J. (2019). LncRNA Neat1 mediates miR-124-induced activation of Wnt/β-catenin signaling in spinal cord neural progenitor cells. Stem Cell Res Ther *10*, 400. 10.1186/s13287-019-1487-3.

Dias, A.P., Dufu, K., Lei, H., and Reed, R. (2010). A role for TREX components in the release of spliced mRNA from nuclear speckle domains. Nat Commun *1*, 97. 10.1038/ncomms1103.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21. 10.1093/bioinformatics/bts635.

Dufu, K., Livingstone, M.J., Seebacher, J., Gygi, S.P., Wilson, S.A., and Reed, R. (2010). ATP is required for interactions between UAP56 and two conserved mRNA export proteins, Aly and CIP29, to assemble the TREX complex. Genes Dev *24*, 2043-2053. 10.1101/gad.1898610.

Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol Cell *46*, 884-892. 10.1016/j.molcel.2012.05.037.

Gieldon, L., Mackenroth, L., Kahlert, A.K., Lemke, J.R., Porrmann, J., Schallner, J., von der Hagen, M., Markus, S., Weidensee, S., Novotna, B., et al. (2018). Diagnostic value of partial exome sequencing in developmental disorders. PLoS One *13*, e0201041. 10.1371/journal.pone.0201041.

Greijer, A.E., and van der Wall, E. (2004). The role of hypoxia inducible factor 1 (HIF-1) in hypoxia induced apoptosis. J Clin Pathol *57*, 1009-1014. 10.1136/jcp.2003.015032.

Gromadzka, A.M., Steckelberg, A.L., Singh, K.K., Hofmann, K., and Gehring, N.H. (2016). A short conserved motif in ALYREF directs cap- and EJC-dependent assembly of export complexes on spliced mRNAs. Nucleic Acids Res *44*, 2348-2361. 10.1093/nar/gkw009.

Guo, C.J., Ma, X.K., Xing, Y.H., Zheng, C.C., Xu, Y.F., Shan, L., Zhang, J., Wang, S., Wang, Y., Carmichael, G.G., et al. (2020). Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. Cell *181*, 621-636.e622. 10.1016/j.cell.2020.03.006.

Gupta, N., Yadav, S., Gurramkonda, V.B., Vl, R., Sg, T., and Kabra, M. (2020). First report of THOC6 related intellectual disability (Beaulieu Boycott Innes syndrome) in two siblings from India. Eur J Med Genet *63*, 103742. 10.1016/j.ejmg.2019.103742.

Guria, A., Tran, D.D., Ramachandran, S., Koch, A., El Bounkari, O., Dutta, P., Hauser, H., and Tamura, T. (2011). Identification of mRNAs that are spliced but not exported to the cytoplasm in the absence of THOC5 in mouse embryo fibroblasts. RNA *17*, 1048-1056. 10.1261/rna.2607011.

Harrison-Uy, S.J., and Pleasure, S.J. (2012). Wnt signaling and forebrain development. Cold Spring Harb Perspect Biol *4*, a008094. 10.1101/cshperspect.a008094.

Hassanvand Amouzadeh, M., Akhavan Sepahi, M., and Abasi, E. (2020). Proteinuria in Two Sisters with Beaulieu-Boycott-Innes Syndrome, A Case Report. Iran J Kidney Dis *14*, 312-314.

Hautbergue, G.M., Hung, M.L., Golovanov, A.P., Lian, L.Y., and Wilson, S.A. (2008). Mutually exclusive interactions drive handover of mRNA from export adaptors to TAP. Proc Natl Acad Sci U S A *105*, 5154-5159. 10.1073/pnas.0709167105.

Hautbergue, G.M., Hung, M.L., Walsh, M.J., Snijders, A.P., Chang, C.T., Jones, R., Ponting, C.P., Dickman, M.J., and Wilson, S.A. (2009). UIF, a New mRNA export adaptor that works together with REF/ALY, requires FACT for recruitment to mRNA. Curr Biol *19*, 1918-1924. 10.1016/j.cub.2009.09.041.

Heath, C.G., Viphakone, N., and Wilson, S.A. (2016). The role of TREX in gene expression and disease. Biochem J *473*, 2911-2935. 10.1042/BCJ20160010.

Izumikawa, K., Ishikawa, H., Simpson, R.J., and Takahashi, N. (2018). Modulating the expression of Chtop, a versatile regulator of gene-specific transcription and mRNA export. RNA Biol *15*, 849-855. 10.1080/15476286.2018.1465795.

Jalali, A., Bassuk, A.G., Kan, L., Israsena, N., Mukhopadhyay, A., McGuire, T., and Kessler, J.A. (2011). HeyL promotes neuronal differentiation of neural progenitor cells. J Neurosci Res *89*, 299-309. 10.1002/jnr.22562.

Jimeno, S., and Aguilera, A. (2010). The THO complex as a key mRNP biogenesis factor in development and cell differentiation. J Biol *9*, 6. 10.1186/jbiol217.

Jimeno, S., Rondón, A.G., Luna, R., and Aguilera, A. (2002). The yeast THO complex and mRNA export factors link RNA metabolism with transcription and genome instability. EMBO J *21*, 3526-3535. 10.1093/emboj/cdf335.

Jin, L., Chen, Y., Crossman, D.K., Datta, A., Vu, T., Mobley, J.A., Basu, M.K., Scarduzio, M., Wang, H., Chang, C., and Datta, P.K. (2020). STRAP regulates alternative splicing fidelity during lineage commitment of mouse embryonic stem cells. Nat Commun *11*, 5941. 10.1038/s41467-020-19698-6.

Juneau, K., Miranda, M., Hillenmeyer, M.E., Nislow, C., and Davis, R.W. (2006). Introns regulate RNA and protein abundance in yeast. Genetics *174*, 511-518. 10.1534/genetics.106.058560.

Katahira, J., Inoue, H., Hurt, E., and Yoneda, Y. (2009). Adaptor Aly and co-adaptor Thoc5 function in the Tap-p15-mediated nuclear export of HSP70 mRNA. EMBO J *28*, 556-567. 10.1038/emboj.2009.5.

Katahira, J., Okuzaki, D., Inoue, H., Yoneda, Y., Maehara, K., and Ohkawa, Y. (2013). Human TREX component Thoc5 affects alternative polyadenylation site choice by recruiting mammalian cleavage factor I. Nucleic Acids Res *41*, 7060-7072. 10.1093/nar/gkt414.

Keenan, A.B., Torre, D., Lachmann, A., Leong, A.K., Wojciechowicz, M.L., Utti, V., Jagodnik, K.M., Kropiwnicki, E., Wang, Z., and Ma'ayan, A. (2019). ChEA3: transcription factor enrichment analysis by orthogonal omics integration. Nucleic Acids Res *47*, W212-W224. 10.1093/nar/gkz446.

Kiraz, A., Tubaş, F., and Seber, T. (2022). A truncating variant in the THOC6 gene with new findings in a patient with Beaulieu-Boycott-Innes syndrome. Am J Med Genet A *188*, 1568-1571. 10.1002/ajmg.a.62667.

Kochinke, K., Zweier, C., Nijhof, B., Fenckova, M., Cizek, P., Honti, F., Keerthikumar, S., Oortveld, M.A., Kleefstra, T., Kramer, J.M., et al. (2016). Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. Am J Hum Genet *98*, 149-164. 10.1016/j.ajhg.2015.11.024.

Köhler, A., and Hurt, E. (2007). Exporting RNA from the nucleus to the cytoplasm. Nat Rev Mol Cell Biol *8*, 761-773. 10.1038/nrm2255.

Lancaster, M.A., Renner, M., Martin, C.A., Wenzel, D., Bicknell, L.S., Hurles, M.E., Homfray, T., Penninger, J.M., Jackson, A.P., and Knoblich, J.A. (2013). Cerebral organoids model human brain development and microcephaly. Nature *501*, 373-379. 10.1038/nature12517.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921. 10.1038/35057062.

Lemire, G., Innes, A.M., and Boycott, K.M. (2020). THOC6 intellectual disability syndrome. *GeneReviews®[Internet]*.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323. 10.1186/1471-2105-12-323.

Li, Y., Muffat, J., Omer, A., Bosch, I., Lancaster, M.A., Sur, M., Gehrke, L., Knoblich, J.A., and Jaenisch, R. (2017). Induction of Expansion and Folding in Human Cerebral Organoids. Cell Stem Cell *20*, 385-396.e383. 10.1016/j.stem.2016.11.017.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464-469. 10.1038/nature07488.

Llorian, M., Schwartz, S., Clark, T.A., Hollander, D., Tan, L.Y., Spellman, R., Gordon, A., Schweitzer, A.C., de la Grange, P., Ast, G., and Smith, C.W. (2010). Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. Nat Struct Mol Biol *17*, 1114-1123. 10.1038/nsmb.1881.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550. 10.1186/s13059-014-0550-8.

Luna, R., Rondón, A.G., and Aguilera, A. (2012). New clues to understand the role of THO and other functionally related factors in mRNP biogenesis. Biochim Biophys Acta *1819*, 514-520. 10.1016/j.bbagrm.2011.11.012.

Luna, R., Rondón, A.G., Pérez-Calero, C., Salas-Armenteros, I., and Aguilera, A. (2019). The THO Complex as a Paradigm for the Prevention of Cotranscriptional R-Loops. Cold Spring Harb Symp Quant Biol *84*, 105-114. 10.1101/sqb.2019.84.039594.

Luo, M.L., Zhou, Z., Magni, K., Christoforides, C., Rappsilber, J., Mann, M., and Reed, R. (2001). Pre-mRNA splicing and mRNA export linked by direct interactions between UAP56 and Aly. Nature *413*, 644-647. 10.1038/35098106.

Maeder, C.I., Kim, J.I., Liang, X., Kaganovsky, K., Shen, A., Li, Q., Li, Z., Wang, S., Xu, X.Z.S., Li, J.B., et al. (2018). The THO Complex Coordinates Transcripts for Synapse Development and Dopamine Neuron Survival. Cell *174*, 1436-1449.e1420. 10.1016/j.cell.2018.07.046.

Mancini, A., Niemann-Seyde, S.C., Pankow, R., El Bounkari, O., Klebba-Färber, S., Koch, A., Jaworska, E., Spooncer, E., Gruber, A.D., Whetton, A.D., and Tamura, T. (2010). THOC5/FMIP, an mRNA export TREX complex protein, is essential for hematopoietic primitive cell survival in vivo. BMC Biol *8*, 1. 10.1186/1741-7007-8-1.

Masuda, S., Das, R., Cheng, H., Hurt, E., Dorman, N., and Reed, R. (2005). Recruitment of the human TREX complex to mRNA during splicing. Genes Dev *19*, 1512-1517. 10.1101/gad.1302205.

Mattioli, F., Isidor, B., Abdul-Rahman, O., Gunter, A., Huang, L., Kumar, R., Beaulieu, C., Gecz, J., Innes, M., Mandel, J.L., and Piton, A. (2018). Clinical and functional characterization of recurrent missense variants implicated in THOC6-related intellectual disability. Hum Mol Genet. 10.1093/hmg/ddy391.

Mattioli, F., Isidor, B., Abdul-Rahman, O., Gunter, A., Huang, L., Kumar, R., Beaulieu, C., Gecz, J., Innes, M., Mandel, J.L., and Piton, A. (2019). Clinical and functional characterization of recurrent missense variants implicated in THOC6-related intellectual disability. Hum Mol Genet *28*, 952-960. 10.1093/hmg/ddy391.

Mauger, O., Lemoine, F., and Scheiffele, P. (2016). Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity. Neuron *92*, 1266-1278. 10.1016/j.neuron.2016.11.032.

Mazin, P.V., Jiang, X., Fu, N., Han, D., Guo, M., Gelfand, M.S., and Khaitovich, P. (2018). Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques. RNA *24*, 585-596. 10.1261/rna.064931.117.

Mazin, P.V., Khaitovich, P., Cardoso-Moreira, M., and Kaessmann, H. (2021). Alternative splicing during mammalian organ development. Nat Genet *53*, 925-934. 10.1038/s41588-021-00851-w.

Meinel, D.M., Burkert-Kautzsch, C., Kieser, A., O'Duibhir, E., Siebert, M., Mayer, A., Cramer, P., Söding, J., Holstege, F.C., and Sträßer, K. (2013). Recruitment of TREX to the transcription machinery by its direct binding to the phospho-CTD of RNA polymerase II. PLoS Genet *9*, e1003914. 10.1371/journal.pgen.1003914.

Merz, C., Urlaub, H., Will, C.L., and Lührmann, R. (2007). Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. RNA *13*, 116-128. 10.1261/rna.336807.

Meyers, E.A., and Kessler, J.A. (2017). TGF-β Family Signaling in Neural and Neuronal Differentiation, Development, and Function. Cold Spring Harb Perspect Biol *9*. 10.1101/cshperspect.a022244.

Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra, S., Mohammed, A., James, A.R., Hoberg, E., et al. (2015). MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. Nat Commun *6*, 7743. 10.1038/ncomms8743.

Moon, S.L., Anderson, J.R., Kumagai, Y., Wilusz, C.J., Akira, S., Khromykh, A.A., and Wilusz, J. (2012). A noncoding RNA produced by arthropod-borne flaviviruses inhibits the cellular exoribonuclease XRN1 and alters host mRNA stability. RNA *18*, 2029-2040. 10.1261/rna.034330.112.

Mor, A., Suliman, S., Ben-Yishay, R., Yunger, S., Brody, Y., and Shav-Tal, Y. (2010). Dynamics of single mRNP nucleocytoplasmic transport and export through the nuclear pore in living cells. Nat Cell Biol *12*, 543-552. 10.1038/ncb2056.

Müller, H., Jimenez-Heredia, R., Krolo, A., Hirschmugl, T., Dmytrus, J., Boztug, K., and Bock, C. (2017). VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data. Nucleic Acids Res *45*, W567-W572. 10.1093/nar/gkx425.

Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M., and Gotoh, O. (2005). Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. Gene *364*, 53-62. 10.1016/j.gene.2005.07.027.

Park, J.W., Tokheim, C., Shen, S., and Xing, Y. (2013). Identifying differential alternative splicing events from RNA sequencing data using RNASeq-MATS. Methods Mol Biol *1038*, 171-179. 10.1007/978-1-62703-514-9_10.

Peña, A., Gewartowski, K., Mroczek, S., Cuéllar, J., Szykowska, A., Prokop, A., Czarnocki-Cieciura, M., Piwowarski, J., Tous, C., Aguilera, A., et al. (2012). Architecture and nucleic acids recognition mechanism of the THO complex, an mRNP assembly factor. EMBO J *31*, 1605-1616. 10.1038/emboj.2012.10.

Pérez-Calero, C., Bayona-Feliu, A., Xue, X., Barroso, S.I., Muñoz, S., González-Basallote, V.M., Sung, P., and Aguilera, A. (2020). UAP56/DDX39B is a major cotranscriptional RNA-DNA helicase that unwinds harmful R loops genome-wide. Genes Dev *34*, 898-912. 10.1101/gad.336024.119.

Pühringer, T., Hohmann, U., Fin, L., Pacheco-Fiallos, B., Schellhaas, U., Brennecke, J., and Plaschka, C. (2020). Structure of the human core transcription-export complex reveals a hub for multivalent interactions. Elife *9*. 10.7554/eLife.61503.

Qu, Q., Sun, G., Murai, K., Ye, P., Li, W., Asuelime, G., Cheung, Y.T., and Shi, Y. (2013). Wnt7a regulates multiple steps of neurogenesis. Mol Cell Biol *33*, 2551-2559. 10.1128/MCB.00325-13.

Retterer, K., Juusola, J., Cho, M.T., Vitazka, P., Millan, F., Gibellini, F., Vertino-Bell, A., Smaoui, N., Neidich, J., Monaghan, K.G., et al. (2016). Clinical application of whole-exome sequencing across clinical indications. Genet Med *18*, 696-704. 10.1038/gim.2015.148.

Rondón, A.G., Jimeno, S., and Aguilera, A. (2010). The interface between transcription and mRNP export: from THO to THSC/TREX-2. Biochim Biophys Acta *1799*, 533-538. 10.1016/j.bbagrm.2010.06.002.

Roy, B., Friesen, W.J., Tomizawa, Y., Leszyk, J.D., Zhuo, J., Johnson, B., Dakka, J., Trotta, C.R., Xue, X., Mutyam, V., et al. (2016). Ataluren stimulates ribosomal selection of near-cognate tRNAs to promote nonsense suppression. Proc Natl Acad Sci U S A *113*, 12508-12513. 10.1073/pnas.1605336113.

Ruaud, L., Roux, N., Boutaud, L., Bessières, B., Ageorges, F., Achaiaa, A., Bole, C., Nitschke, P., Masson, C., Vekemans, M., et al. (2022). Biallelic THOC6 pathogenic variants: Prenatal phenotype and review of the literature. Birth Defects Res *114*, 499-504. 10.1002/bdr2.2011.

Schalock, R.L., Luckasson, R., and Tassé, M.J. (2021). An Overview of Intellectual Disability: Definition, Diagnosis, Classification, and Systems of Supports (12th ed.). Am J Intellect Dev Disabil *126*, 439-442. 10.1352/1944-7558-126.6.439.

Shen, J., Zhang, L., and Zhao, R. (2007). Biochemical characterization of the ATPase and helicase activity of UAP56, an essential pre-mRNA splicing and mRNA export factor. J Biol Chem *282*, 22544-22550. 10.1074/jbc.M702304200.

Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A *111*, E5593-5601. 10.1073/pnas.1419161111.

Shi, M., Zhang, H., Wu, X., He, Z., Wang, L., Yin, S., Tian, B., Li, G., and Cheng, H. (2017). ALYREF mainly binds to the 5' and the 3' regions of the mRNA in vivo. Nucleic Acids Res *45*, 9640-9653. 10.1093/nar/gkx597.

Sobreira, N., Schiettecatte, F., Valle, D., and Hamosh, A. (2015). GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. Hum Mutat *36*, 928-930. 10.1002/humu.22844.

Srivastava, A., Srivastava, K.R., Hebbar, M., Galada, C., Kadavigrere, R., Su, F., Cao, X., Chinnaiyan, A.M., Girisha, K.M., Shukla, A., and Bielas, S.L. (2018). Genetic diversity of NDUFV1-dependent mitochondrial complex I deficiency. Eur J Hum Genet *26*, 1582-1587. 10.1038/s41431-018-0209-0.

Strässer, K., and Hurt, E. (2001). Splicing factor Sub2p is required for nuclear mRNA export through its interaction with Yra1p. Nature *413*, 648-652. 10.1038/35098113.

Strässer, K., Masuda, S., Mason, P., Pfannstiel, J., Oppizzi, M., Rodriguez-Navarro, S., Rondón, A.G., Aguilera, A., Struhl, K., Reed, R., and Hurt, E. (2002). TREX is a conserved complex coupling transcription with messenger RNA export. Nature *417*, 304-308. 10.1038/nature746.

Taniguchi, I., and Ohno, M. (2008). ATP-dependent recruitment of export factor Aly/REF onto intronless mRNAs by RNA helicase UAP56. Mol Cell Biol *28*, 601-608. 10.1128/MCB.01341-07.

Tran, D.D., Saran, S., Williamson, A.J., Pierce, A., Dittrich-Breiholz, O., Wiehlmann, L., Koch, A., Whetton, A.D., and Tamura, T. (2014). THOC5 controls 3'end-processing of immediate early genes via interaction with polyadenylation specific factor 100 (CPSF100). Nucleic Acids Res *42*, 12249-12260. 10.1093/nar/gku911.

Truett, G.E., Heeger, P., Mynatt, R.L., Truett, A.A., Walker, J.A., and Warman, M.L. (2000). Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). Biotechniques *29*, 52, 54. 10.2144/00291bm09.

Tsui, D., Vessey, J.P., Tomita, H., Kaplan, D.R., and Miller, F.D. (2013). FoxP2 regulates neurogenesis during embryonic cortical development. J Neurosci *33*, 244-258. 10.1523/JNEUROSCI.1665-12.2013.

Vasudevan, P., and Suri, M. (2017). A clinical approach to developmental delay and intellectual disability. Clin Med (Lond) *17*, 558-561. 10.7861/clinmedicine.17-6-558.

Veiga, D.F.T. (2022). *maser: Mapping Alternative Splicing Events to pRoteins* .

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304-1351. 10.1126/science.1058040.

Viphakone, N., Sudbery, I., Griffith, L., Heath, C.G., Sims, D., and Wilson, S.A. (2019). Co-transcriptional Loading of RNA Export Factors Shapes the Human Transcriptome. Mol Cell *75*, 310-323.e318. 10.1016/j.molcel.2019.04.034.

Vogel, T., Ahrens, S., Büttner, N., and Krieglstein, K. (2010). Transforming growth factor beta promotes neuronal cell fate of mouse cortical and hippocampal progenitors in vitro and in vivo: identification of Nedd9 as an essential signaling component. Cereb Cortex *20*, 661-671. 10.1093/cercor/bhp134.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476. 10.1038/nature07509.

Wang, H., Ge, G., Uchida, Y., Luu, B., and Ahn, S. (2011). Gli3 is required for maintenance and fate specification of cortical progenitors. J Neurosci *31*, 6440-6448. 10.1523/JNEUROSCI.4892-10.2011.

Wang, L., Miao, Y.L., Zheng, X., Lackford, B., Zhou, B., Han, L., Yao, C., Ward, J.M., Burkholder, A., Lipchina, I., et al. (2013). The THO complex regulates pluripotency gene mRNA export and controls embryonic stem cell self-renewal and somatic cell reprogramming. Cell Stem Cell *13*, 676-690. 10.1016/j.stem.2013.10.008.

Wang, Y., Liu, J., Huang, B.O., Xu, Y.M., Li, J., Huang, L.F., Lin, J., Zhang, J., Min, Q.H., Yang, W.M., and Wang, X.Z. (2015). Mechanism of alternative splicing and its regulation. Biomed Rep *3*, 152-158. 10.3892/br.2014.407.

Weyn-Vanhentenryck, S.M., Feng, H., Ustianenko, D., Duffié, R., Yan, Q., Jacko, M., Martinez, J.C., Goodwin, M., Zhang, X., Hengst, U., et al. (2018). Precise temporal regulation of alternative splicing during neural development. Nat Commun *9*, 2189. 10.1038/s41467-018-04559-0.

Wickramasinghe, V.O., and Laskey, R.A. (2015). Control of mammalian gene expression by selective mRNA export. Nat Rev Mol Cell Biol *16*, 431-442. 10.1038/nrm4010.

Xie, Y., and Ren, Y. (2019). Mechanisms of nuclear mRNA export: A structural perspective. Traffic *20*, 829-840. 10.1111/tra.12691.

Yang, Y., Muzny, D.M., Xia, F., Niu, Z., Person, R., Ding, Y., Ward, P., Braxton, A., Wang, M., Buhay, C., et al. (2014). Molecular findings among patients referred for clinical whole-exome sequencing. JAMA *312*, 1870-1879. 10.1001/jama.2014.14601.

Yap, K., Lim, Z.Q., Khandelia, P., Friedman, B., and Makeyev, E.V. (2012). Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev *26*, 1209-1223. 10.1101/gad.188037.112.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377-394. 10.1089/1066527041410418.

Zhang, Q., Chen, S., Qin, Z., Zheng, H., and Fan, X. (2020). The first reported case of Beaulieu-Boycott-Innes syndrome caused by two novel mutations in THOC6 gene in a Chinese infant. Medicine (Baltimore) *99*, e19751. 10.1097/MD.0000000000019751.

Zhao, J., Zhang, X., Zhou, Y., Ansell, P.J., and Klibanski, A. (2006). Cyclic AMP stimulates MEG3 gene expression in cells through a cAMP-response element (CRE) in the MEG3 proximal promoter region. Int J Biochem Cell Biol *38*, 1808-1820. 10.1016/j.biocel.2006.05.004.

Zhou, S., Zhong, Z., Huang, P., Xiang, B., Li, X., Dong, H., Zhang, G., Wu, Y., and Li, P. (2021). IL-6/STAT3 Induced Neuron Apoptosis in Hypoxia by Downregulating ATF6 Expression. Front Physiol *12*, 729925. 10.3389/fphys.2021.729925.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun *10*, 1523. 10.1038/s41467-019-09234-6.

Zuckerman, B., Ron, M., Mikl, M., Segal, E., and Ulitsky, I. (2020). Gene Architecture and Sequence Composition Underpin Selective Dependency of Nuclear Export of Long RNAs on NXF1 and the TREX Complex. Mol Cell *79*, 251-267.e256. 10.1016/j.molcel.2020.05.013.

Zylka, M.J., Simon, J.M., and Philpot, B.D. (2015). Gene length matters in neurons. Neuron *86*, 353-355. 10.1016/j.neuron.2015.03.059.

Chapter 6
Conclusion

With advancements in sequencing technology and clinical genomics testing, we are gradually uncovering the genetic basis and molecular features of human-specific traits and disorders. By application of evolutionary medicine approaches, we are beginning to investigate how evolutionary biology influences modern disorders with the goal of prevention and treatment. Innovations in stem cell technology and 3D organoid modeling made over the last few decades provide unprecedented opportunities for researchers to investigate hypotheses of developmental biology. The combination of genetic testing, molecular biology techniques, and organoid modeling provide a powerful approach for evolutionary medicine research focused on human developmental biology. This thesis employs an evolutionary medicine framework to investigate the complex interplay between evolutionary cortical expansion and neurodevelopmental disorders, utilizing genetics and organoid models as tools to unveil this biology.

When, in evolutionary time, the genetic variation that underlies a particular phenotype first arose varies. The most apparent genetic candidates for novel traits are recent genetic variation that is derived in, or specific to, a lineage following speciation. However, ancestral variation, which is present prior to speciation and conserved across several lineages of organisms, can also contribute to the expression of species-specific physiology. This dissertation leverages derived and ancestral genetic variation to investigate the overarching question: *what is the genetic basis of human neural development?* To address this question, one approach taken in **Chapter 3** is to investigate genetic variation derived in the human lineage, or specific to the human genome, that has been implicated in neural development. The second approach, taken in **Chapters 4 and 5**, is to investigate genetic variation, or genes, known to cause NND when disrupted. In this case, genes can be either derived or ancestral. If ancestral, or conserved across species, their investigation can still inform our understanding of how evolution and disease interact and how conserved functions within a human developmental context manifest into human-specific physiology. Human-specific expression of disease can be uncovered using this approach, especially when phenotypes have not been recapitulated in model organisms.

Chapters 3 and 4 converge by demonstrating how evolution has favored gene duplicates in neural development. Both projects explore the functional fates of gene duplications that are characterized by repetitive protein domains and long coding sequence with evidence for retained ancestral functions. In **Chapter 3**, we show evidence for NBPF *cis*-regulatory divergence in human evolution resulting in enriched expression in NPCs with putative retained ancestral functions in microtubule dynamics that regulate human-specific corticogenesis. We can have confidence of *NBPF* gene duplications with intragenic domain expansion as a genetic basis of cortical size and neurodevelopmental disorders based on the investigation of other evolved duplicate genes with similar sequence fates that determine cortical size, such as the Cub and Sushi Multiple Domain (*CSMD*) gene family explored in **Chapter 4.** Although *CSMD* genes evolved further back in evolutionary time, we also identified evidence for conserved functions with other *CSMD* family members in neural development. In this chapter, we identified *CSMD1* as a novel genetic basis of human neurodevelopmental disorders with functions in NPC proliferation, differentiation, and neuronal migration. Lastly, **Chapter 5** explored the functional evolution of the THOC6 protein that has co-evolved to support changing mRNA processing requirements in human neural development. While the evolution of *THOC6* within metazoans was not evaluated, it is derived in metazoans, a disease gene in humans, and conveys human-specific biology. Therefore, this project utilized the principles of evolutionary medicine to uncover novel functions of a conserved protein complex required for cortical size and function. We provide evidence that THOC6-mediated TREX tetramers have evolved to meet the increased mRNA processing demands of long genes in the developing brain requiring elevate splicing, in parallel with the increased susceptibility to transcription-associated instability due to the accumulation of highly repetitive structural variation in the human genome. Overall, our work demonstrates THOC6 functions are required for human neural development and are necessary for mouse embryogenesis. Together, these projects provide an evolutionary medicine framework to study human developmental biology and identify conserved and derived genes implicated in different aspects of human corticogenesis (Figure 6.1). Here, I summarize the novel findings of each research project that implicate important functions in NPC activity and enhance our understanding of the genetics of human cortical evolution and NND. I also discuss major limitations of each project and postulate future research directions.

6.1 APPLICATIONS OF EVOLUTIONARY MEDICINE FOR STUDYING CORTICOGENESIS AND NEURODEVELOPMENTAL DISORDERS

6.1.1 Functional fates of closely related duplicates in brain development: *NBPF*

Our investigation of NBPF protein expression, localization, and protein interactions using an interspecific comparative experimental approach in human and chimpanzee NPCs and cerebral organoid model systems supports a model of novel expression in hNPCs in early cortical development, with a striking enrichment in mitotically active NPCs. Shared subcellular localizations and interactions with proteins implicated in the interactome of PDE4DIP, which harbors the ancestral DUF1220 domain sequence, supports a model of ancestral functional retention via DUF1220 domains in human corticogenesis. However, we demonstrate a critical difference between human and chimpanzee NBPF overexpression models that have implications for a dosage threshold of NBPF function in human cells required for regulation of NPC proliferative dynamics. We also identify novel NBPF-NBPF protein interactions that have implications for NBPF-mediated biology in a human background. This work represents the first attempt at functional investigation of human-specific NBPF proteins, and our findings serve to implicate *NBPF*s in the etiology of NND.

While chapter 3 represents the first functional study of NBPF proteins in human cerebral organoid models, there are several challenges with this work that need to be addressed. First, short read sequencing data to assess NBPF paralog-specific expression is suggestive at best, but to confidently characterize expression profiles, high quality long-read RNA-sequencing is necessary, likely with targeted probe enrichment. High coverage long-read RNA-sequencing is also required to annotate *NBPF* isoform variability that is likely contributing to NBPF-mediated biology. Second, the inconclusive genotyping of edited lines warrants further investigation of the long-read HiFi genome sequencing data with more sensitive parameters to better annotate introduced edits as well as a re-evaluation of the most effective targeted-editing and validation strategy. The project would benefit from simultaneously taking a multi-pronged approach. Editing strategies that have not been tried yet but may prove fruitful include utilizing small molecules during CRISPR/Cas9 editing identified to enhance homologous recombination, such as provided with the Alt-R system (Integrated DNA Technologies), and/or repeated electroporations in edited cells to increase the likelihood of homozygous editing events. Technology continues to increase the success of zygosity in editing which may improve this project moving forward.

To investigate functional divergence among NBPF duplicate proteins, future work needs to be conducted to test for paralog-specific functional differences in cortical organoid and NPC models (both in humans and between humans and chimpanzees), preferentially by direct investigation of hs*NBPF14* or hs*NBPF10*. However, given futile efforts in our research, how to successfully isolate *hsNBPF* cDNA remains unsolved. Given that mice lack NBPF sequence, generation of a mouse model, or differentiation of mouse cortical organoids, with ectopic (or novel) expression of primate-specific (*NBPF15*) and human-specific *NBPF* would be an informative comparative model for NBPF dosage in NPC proliferation in combination with chimpanzee and human cortical organoids. This approach would be especially informative if NBPFs function to enhance proliferation of oRG in the OSVZ, given that mice lack this zone.

In addition, future work investigating functional divergence between DUF1220 protein domain subtypes is pertinent to our understanding of DUF1220-expanded NBPF function in human neural development and 1q21DDS pathology. One way to investigate domain functions is isolation and cloning of sequence encoding the different DUF1220 domains representing each clade into an expression vector followed by expression in NPCs to assess localization and overexpression effects on proliferation. Given that increased domain-specific copy number has been associated with autism severity, specifically of CON1 and HLS1, and hsNBPF proteins are primarily distinguished by their extreme HLS-triplet amplification, comparison of the more conserved CON1 versus the HLS-triplet is a promising experimental design that could make a compelling argument for NBPF functional divergence.

Despite challenges, we feel that this thesis work represents a major stride in understanding NBPF biological functions and raises an array of questions for the field regarding *NBPF* dosage in neural development and disease.

6.1.2 Functional fates of closely related duplicates in brain development: *CSMD*

Findings from human genetics in conjunction with functional studies of CSMD1-depleted human cortical organoids strongly implicate CSMD1 function in the regulation of proper neural development. Specifically, we identify novel functions in the proliferative dynamics of cortical NPCs, the timing of differentiation, and putative facilitation of neuronal migration. This work represents the first paper to make an argument for CSMD1 as a genetic etiology for MCD.

CSMD family genes, including *CSMD1*, *CSMD2*, and *CSMD3*, are comprised of repetitive sequence encoding several interspersed CUB and Sushi domains—motifs that are conserved in regulators of the complement pathway. *CSMD1* and *CSMD2* have documented associations with schizophrenia susceptibility, and *CSMD2* and *CSMD3* have been functionally investigated in neural development. Like *CSMD2* and *CSMD3*, *CSMD1* is well-situated to regulate brain-specific functions given its high expression in the brain. Considering emerging functions of the complement pathway in neural development, namely stimulation of NPC proliferation, neuronal migration, and synaptic pruning, our findings of over-proliferation in *CSMD1*$^{fs/fs}$ organoids support a model of CSMD1 as an inhibitor of complement in hNPCs, as has been proposed for neurons. One way to test this is to assess complement protein deposition in *CSMD1*$^{fs/fs}$ organoids with the prediction of increased accumulation at the ventricular zone of developing *CSMD1*$^{fs/fs}$ NRs. Another option is to perform a rescue experiment by use of known drug inhibitors of the complement pathway by addition to *CSMD1*$^{fs/fs}$ organoid media, with the prediction of improved apicobasal polarity, pseudostratified organization, and differentiation timing in treated *CSMD1*$^{fs/fs}$ NRs. An additional research direction is to generate conditional knockout organoids of different members of the complement pathway to compare phenotypic overlap with *CSMD1*$^{fs/fs}$ organoids.

The lack of neurodevelopmental defects in *Csmd1* knockout mouse suggest species differences in CSMD1 function in corticogenesis between mouse and human. This raises the possibility that complement proteins or complement regulators have evolved different functions in mouse and human cortical development. Given that we see early developmental defects in day 28ND organoids, this suggests different evolved functions of CSMD1 at the ventricular zone, a feature which is shared in human and mouse developing brain. This discrepancy could also be attributed to compensatory effects in mouse but not human cells. Further investigation of *Csmd1* knockout mouse at early embryonic timepoints coupled with interspecific characterization of complement pathway expression profiles will be needed to help elucidate these species differences.

Disentangling the evolution and molecular functions of *CSMD* genes in human corticogenesis is an important future direction. Evidence of shared functions of *CSMD1*, *CSMD2*, and *CSMD3* in neural development could reflect subfunctionalization following *CSMD* duplicative evolution. It will be interesting from an evolutionary medicine perspective to assess if *CSMD* paralogs contribute redundant functions or have divvied up functions throughout neural development. There is the possibility that subfunctionalization of *CSMD* genes has occurred during mammalian evolution outside of the rodent lineage, contributing to the phenotypic findings in humans but not mouse. Phylogenetic analyses of *CSMD* sequence (including coding, non-coding, and regulatory) and transcriptomic profiling of *CSMD* genes may help pinpoint species differences and inform evolutionary models. One way to test for compensatory effects is rescue of CSMD1 loss in human cortical organoids by either *CSMD2* or *CSMD3* overexpression. Conversely, one could generate a double, or triple, conditional *Csmd* knockout mouse to assess potential neurodevelopmental defects due to a loss of compensatory effects.

6.1.3 Novel functions in mRNA processing during brain development: *THOC6*

The THO subcomplex within TREX has been predominantly linked to nuclear export with less attention given to its contribution to other features of mRNP processing. Use of models with pathogenic variants in *THOC6*, which mediates formation of a TREX tetramer in metazoans, allows us to dissect THO functions within a TREX tetramer for the coordination of mRNP processing. We identify aberrant splicing and processing defects due to loss of THOC6-dependent TREX tetramer functions in human neural cells, yet with intact global nuclear RNA export that we predict is afforded by stable dimers. Further, we identify expression dynamics of major signaling pathways correlated to lncRNA dysregulation in affected human cells not identified in mouse. Together, we generate the first mouse model and human cortical organoid models of TIDS, discovering novel evidence for elevated apoptosis, retention of multipotency, and delayed differentiation as pathogenic mechanisms of TIDS.

Although our research offers novel insights into the functions of human TREX in neural development, our data is unable to distinguish between direct and indirect effects of loss of THOC6 on mRNA processing. To assess direct functions, future experiments can be performed with targeted splicing reporter assays using candidate unspliced transcripts identified by our current pipeline and model systems as well as RNA-immunoprecipitation (RIP) experiments followed by bulk RNAseq or targeted transcript sequencing. RIP experiments would help identify evidence of THOC6 interaction enriched at specific locations on unspliced transcripts, which would implicate direct mechanisms of THOC6 in mediating splicing decisions.

Our data is unable to determine if specific transcripts are affected by nuclear export impairment. RNA sequencing of cellular fractions in affected cells may help to address these outstanding questions. In addition, variation in transcriptional and processing requirements throughout neuronal differentiation to facilitate proper neuronal function and activity warrant consideration of single-cell RNAseq profiling of cerebral organoids to capture cellular heterogeneity and hone cell type dependencies on TREX tetramers that are altered in TIDS pathology. In line with this, short read RNAseq is limited in its ability to annotate isoform diversity, especially of long transcripts with highly repetitive sequence, of which THO preferentially regulates in human cells. Future research utilizing full-length transcript RNAseq technologies, such as PacBio IsoSeq, potentially on a single cell level (scIsoSeq), will allow for improved annotation of splicing patterns and isoform diversity in TIDS.

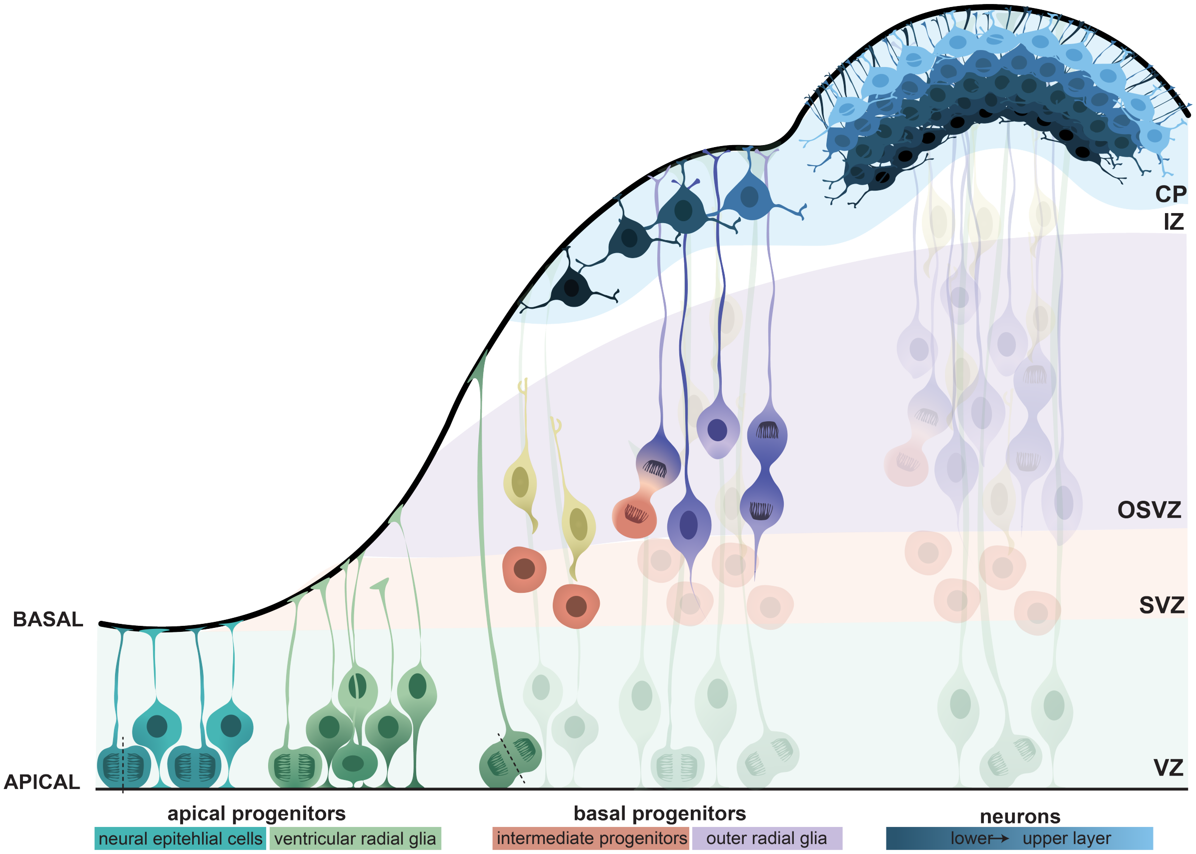
The phenotypic discrepancy between *Thoc6*$^{fs/fs}$ mouse embryonic lethality and human TIDS raises the possibility that differences in downstream genetic mechanisms between variants positioned upstream versus within WD40 repeat domains could underlie species differences. One way to better discriminate between species differences in THOC6 loss-of-function models is to introduce human biallelic variants into mouse *Thoc6* and assess the phenotypic and molecular consequences.

Lastly, due to the difficulty of capturing the dynamic and massive protein-binding platform of TREX under native protein conditions, we were unable to confirm TREX tetramer disruption in *THOC6*-affected hNPCs. An alternative strategy is necessary to assess impact on TREX tetramer assembly. One option is to express *THOC6*-variants together with all other THO members and core TREX factors followed by characterization by cryo-electron microscopy. Given our findings, we predict that THOC6-variant proteins will inhibit formation of a stable TREX tetramer *in vitro*. Data confirming TREX tetramer disruption by biallelic *THOC6* loss-of-function variants is pertinent for follow-up investigations of the biology affected in TIDS pathology.

6.2 EVOLUTIONARY MEDICINE INFORMS MODELS OF HUMAN CORTICAL EVOLUTION

Use of cerebral organoids allows us to investigate human and non-human development separate from mouse development and identify the molecular features of biology specific to the human lineage. Our understanding of human cortical development is largely derived from literature on mouse models. Here, we demonstrate how organoid models can lead to discoveries of human-specific neural development. In this thesis, we identified human-specific traits associated with genes that are implicated in NND (Figure 6.1). Human-specific copy number expansions have played a prominent role in affecting both the activity of both apical and basal NPCs, which is predicted to have a profound effect on cortical size. *CSMD1* impacts the pseudostratified organization of the neuroepithelium, which has implications for all steps of corticogenesis, preferentially impacting NPCs and neurons. *THOC6* has evolved to accommodate the processing demands of long, highly spliced genes expressed during mammalian brain development, and our studies have uncovered an intriguing phenotype that is not shared between mouse and humans that will be important for disentangling species differences while highlighting human-specific features.

Collectively, this research integrates interdisciplinary concepts and methods to advance knowledge of the genetic and molecular origins of neural developmental traits that are unique to the human lineage. This work offers an experimental framework for applying evolutionary medicine principles to the study of human developmental biology.



CHAPTER 3		***NBPF***	***NBPF***	***NBPF***	
CHAPTER 4		***CSMD1***			***CSMD1***
CHAPTER 5		***THOC6***			***THOC6?***

Figure 6.1. Schematic demonstrating proposed cellular biology during human corticogenesis of genes investigated in this dissertation.