Gene expansions contributing to human brain evolution

Daniela C. Soto[1,2*‡], José M. Uribe-Salazar[1,2*], Gulhan Kaya[1,2], Ricardo Valdarrago[3], Aarthi Sekar[1,2], Nicholas K. Haghani[1,2], Keiko Hino[4], Gabriana N. La[1,2], Natasha Ann F. Mariano[1,2,5], Cole Ingamells[1,2], Aidan E. Baraban[1,2], Tychele N. Turner[6], Eric D. Green[7], Sergi Simó[4], Gerald Quon[2,3], Aida M. Andrés[8], Megan Y. Dennis[1,2†]

[1]Department of Biochemistry & Molecular Medicine, MIND Institute, University of California, Davis, CA 95616, USA
[2]Genome Center, University of California, Davis, CA 95616, USA
[3]Department of Molecular and Cellular Biology, University of California, Davis, CA 95616, USA
[4]Department of Cell Biology & Human Anatomy, University of California, Davis, CA 95616, USA
[5]Postbaccalaureate Research Education Program, University of California, Davis, CA 95616, USA
[6]Department of Genetics, Washington University School of Medicine, St Louis, MS, 63110, USA
[7]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20892, USA
[8]UCL Genetics Institute, Department of Genetics, Evolution and Environment, University College, London, WC1E 6BT, UK

[*]These authors contributed equally to this work.

[‡]Current institution: Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, 90095, USA

[†]Corresponding author:
Megan Y. Dennis, Ph.D.
University of California, Davis, School of Medicine
One Shields Avenue
Genome Center, 4303 GBSF
Davis, CA 95616
Email: mydennis@ucdavis.edu

Abstract

Genomic drivers of human-specific neurological traits remain largely undiscovered. Duplicated genes expanded uniquely in the human lineage likely contributed to brain evolution, including the increased complexity of synaptic connections between neurons and the dramatic expansion of the neocortex. Discovering duplicate genes is challenging because the similarity of paralogs makes them prone to sequence-assembly errors. To mitigate this issue, we analyzed a complete telomere-to-telomere human genome sequence (T2T-CHM13) and identified 213 duplicated gene families likely containing human-specific paralogs (>98% identity). Positing that genes important in universal human brain features should exist with at least one copy in all modern humans and exhibit expression in the brain, we narrowed in on 362 paralogs with at least one copy across thousands of ancestrally diverse genomes and present in human brain transcriptomes. Of these, 38 paralogs co-express in gene modules enriched for autism-associated genes and potentially contribute to human language and cognition. We narrowed in on 13 duplicate gene families with human-specific paralogs that are fixed among modern humans and show convincing brain expression patterns. Using long-read DNA sequencing revealed hidden variation across 200 modern humans of diverse ancestries, uncovering signatures of selection not previously identified, including possible balancing selection of *CD8B*. To understand the roles of duplicated genes in brain development, we generated zebrafish CRISPR "knockout" models of nine orthologs and transiently introduced mRNA-encoding paralogs, effectively "humanizing" the larvae. Morphometric, behavioral, and single-cell RNA-seq screening highlighted, for the first time, a possible role for *GPR89B* in dosage-mediated brain expansion and *FRMPD2B* function in altered synaptic signaling, both hallmark features of the human brain. Our holistic approach provides important insights into human brain evolution as well as a resource to the community for studying additional gene expansion drivers of human brain evolution.

Abstract (short)

Duplicated genes expanded in the human lineage likely contributed to brain evolution, yet challenges exist in their discovery due to sequence-assembly errors. We used a complete telomere-to-telomere genome sequence to identify 213 human-specific gene families. From these, 362 paralogs were found in all modern human genomes tested and brain transcriptomes, making them top candidates contributing to human-universal brain features. Choosing a subset of paralogs, we used long-read DNA sequencing of hundreds of modern humans to reveal previously hidden signatures of selection. To understand their roles in brain development, we generated zebrafish CRISPR "knockout" models of nine orthologs and introduced mRNA-encoding paralogs, effectively "humanizing" larvae. Our findings implicate two new genes in possibly contributing to hallmark features of the human brain: *GPR89B* in dosage-mediated brain expansion and *FRMPD2B* in altered synapse signaling. Our holistic approach provides new insights and a comprehensive resource for studying gene expansion drivers of human brain evolution.

Introduction

Significant phenotypic features distinguish modern humans from closely related great apes [1–4]. Arguably, one of the most compelling innovations relates to changes in neuroanatomy, including an expanded neocortex and increased complexity of neuronal connections, which allowed the development of novel cognitive features such as reading and language [5]. While previous work implicated human-specific single-nucleotide variants (SNVs) that impact genes leading to altered brain features, including *FOXP2* [6,7] and human-accelerated regions (HARs) [8], a majority of top gene candidates are the result of segmental duplications (SDs; genomic regions >1 kbp in length that share high sequence identity [>90%]) [9–11]. SDs can give rise to new gene paralogs with the same function, altered functions, or that antagonize conserved, ancestral paralogs and contribute more to genetic divergence across species than SNVs [12]. Previous comparisons of great ape genomes have identified >30 human-specific gene families and hundreds of paralogs enriched for genes important in neurodevelopment and residing at genomic hotspots associated with neuropsychiatric disorders [13–15]. Of these, a handful of genes have been found to function in brain development using model systems, including *SRGAP2* [16,17], *NOTCH2NL* [18–20], *ARHGAP11B* [21–23], *TBC1D3* [24] *CROCCP2* [25], and *LRRC37B* [26]. Most studies have leveraged mice to study gene functions although recent studies have expanded to cortical organoids, ferrets, and primates [27]. Despite their clear importance in contributing to neural features, most duplicate genes remain functionally uncharacterized due to the arduous nature of using such models.

SDs have largely eluded analyses due to difficulties in accurate genome assembly [28] and in discovering variants across nearly identical paralogs [29–33]. As such, many human-duplicated genes are likely left to be discovered. The telomere-to-telomere (T2T) human reference genome T2T-CHM13 [34], representing a gapless sequence of all autosomes and Chromosome X, has enabled a more complete picture of SDs [35] by incorporating 238 Mbp missing from the previous human reference genome (GRCh38). In particular, this new assembly corrects >8 Mbp of collapsed duplications [36], including previously missing paralogs of human-specific duplicated gene families [13] *GPRIN2* [35] and *DUSP22* [36]. Here, using this new T2T genome, we identified thousands of recent gene duplications among hominids. By comparing genomic data between great apes and across thousands of modern humans, we narrowed in on a set of paralogs unique within and fixed across modern humans. Transcriptomic datasets from the human brain identified genes most likely to contribute to brain development and function, providing a catalog of the candidate human-specific gene families contributing to brain evolution for further functional testing in model systems. Finally, we prioritized a set of duplicate gene families to characterize in more detail using long-read sequencing and systematic analysis in zebrafish to connect gene functions to brain development.

Results

Genetic analysis of human-duplicated genes

Identification of human gene duplications in T2T-CHM13

Understanding that highly identical SDs are enriched for human-specific duplications, we narrowed in on 97.8 Mbp of autosomal sequences sharing >98% identity with other genomic regions (or SD98) in the human T2T-CHM13 [35,37] (Figure 1A). These loci represent genes duplicated only in human lineage [13,14] as well as expansions of duplicated gene families present in other great apes. Paralogs in this latter category have experienced recent changes along the *Homo* lineage in expression (e.g., *LRRC37B* [26]) or sequence content (e.g., *NOTCH2NL*, via interlocus gene conversion [18]) resulting in new novel functions. Of the 5,154 SD98 genes (Table S1), we focused on 698 protein-encoding genes and 1,095 unprocessed pseudogenes (that could be mis-annotations of true protein-encoding genes [38]). This list includes well-known duplicated genes important in neurodevelopment (*SRGAP2C*, *ARHGAP11B*), disease (*SMN1* and *SMN2* [39], *KANSL1* [40]), and adaptation (e.g., amylase genes [41–43]). Sequence read depth [35] in modern humans (Simons Genome Diversity Project [SGDP], n=269 [44]) verified that all paralogs had >2 gene-family diploid copy number (famCN; Methods) (Table S2, Figure 1B).

Based on sequence and famCN similarity, we clustered the 1,679 paralogs into 491 multigene families (Figure 1C, Figure S1), with most families having 2–3 members (n=271) (Figure S2). Three extreme high-copy gene families had >50 paralogs, including macrosatellite-associated *DUX4* and *DUB/USP17* as well as primate-specific *FAM90A* [45]. The remaining 114 paralogs were defined as singletons (Table S2), with some failing to cluster due to high and variable copy numbers (CNs) (e.g., *CROCC* and *CROCCP2*) or only a small portion of the gene duplicated (e.g., *AIDA* and *LUZP2*). Within 163 multigene families and 13 singletons, we identified 385 human-specific paralogs within non-syntenic regions present in human but not chimpanzee reference assemblies [35] (Figure 1C, Table S2). Several previously known human-specific genes were notably absent from this list (e.g., *NPY4R2*, *ROCKP1*, and *SERF1B* [13]) because genome alignments across SDs can be imprecise. We next identified human-expanded gene families as those with higher famCN in humans (SDGP, n=269) versus nonhuman great apes (n=4) (97 gene families and 27 singletons; Figure 1C, Table S3), excluding high-copy genes that were difficult to accurately detect CN differences (famCN>10). In total, we conservatively predict 213 gene families and 38 singletons comprising at least one human-specific duplicate paralog (Table S3).

Variation of duplicated genes in modern humans

Positing that all humans should carry a functional version of a gene if important for a species-universal trait, we used *k*-mer-based paralog-specific copy number (parCN) estimates [46] to identify 622 genes (194 duplicate families, 38 singletons) with at least one copy in >98% of humans ("CN constrained", parCN≥0.5; 1000 Genomes Project, 1KGP; n=2,504) (Tables S1 and S4). Of these, 125 paralogs were "fixed" in humans (parCN~2) and likely represent *Homo sapiens*-specific genes. We found 13 CN constrained genes that were largely absent (parCN<0.5) from four archaic human genomes [47–49]. One of these genes, *H3-2/H2BP2*, is a member of a core *H2B* histone family involved in the structure of eukaryotic chromatin [50], homologous with another human-specific *H2BP1* and the ancestral *H2BC18* paralog (Figure 1D). Another *Homo sapiens*-specific gene, *FCGR1CP*, encodes an immunoglobulin gamma Fc Gamma Receptor, a family of proteins vital in regulating immune response [51]. Moving forward, we consider only duplicate gene families comprising CN-constrained genes.

We identified 13 protein-encoding genes as loss-of-function intolerant (pLI ≥0.9 or LOEUF ≤0.35) using SNV data from hundreds of thousands of humans from gnomAD [52] (Table S1, Figure S3), showing that deleterious mutations of these genes are depleted in human populations (e.g., likely not compatible with life). These conserved genes are all ancestral paralogs, including *NOTCH2*, *HERC2*, and *CORO1A*. The gnomAD (v3) metrics rely on variants identified in protein-encoding genes using the human reference genome hg19, which has known errors across SDs [53] and misannotated pseudogenes. As such, all unprocessed pseudogenes and 32% of protein-encoding SD98 genes lacked gnomAD pLI and LOEUF scores. To circumvent these issues, we assessed SNV genetic diversity by Tajima's D [54] using the T2T-

CHM13 reference and the 1KGP cohort [36,55]. Focusing on short-read accessible regions (Figure S4, Note S1), we identified 15 CN-constrained human-duplicated genes with extreme negative D values (<5th percentile of the genome-wide empirical distribution) considered signatures of positive or purifying selection (Figures 1E and S5). These included human-specific paralog *SRGAP2C* previously implicated in cortical neuronal migration and synaptogenesis [16,17] as well as the uncharacterized *LRRC37A3* and the hominid-specific *LRRC37B*, recently found to function in cortical pyramidal neurons by impacting synaptic excitability [26]. We also identified nine genes exhibiting extremely positive D values (>95th percentile) as putative signatures of balancing selection, including T-cell antigen *CD8B*. Collectively, variants discovered using the new T2T-CHM13 genome enabled the identification of new and interesting human-duplicated genes potentially contributing to traits and diseases not previously assayed in genome-wide selection screens.

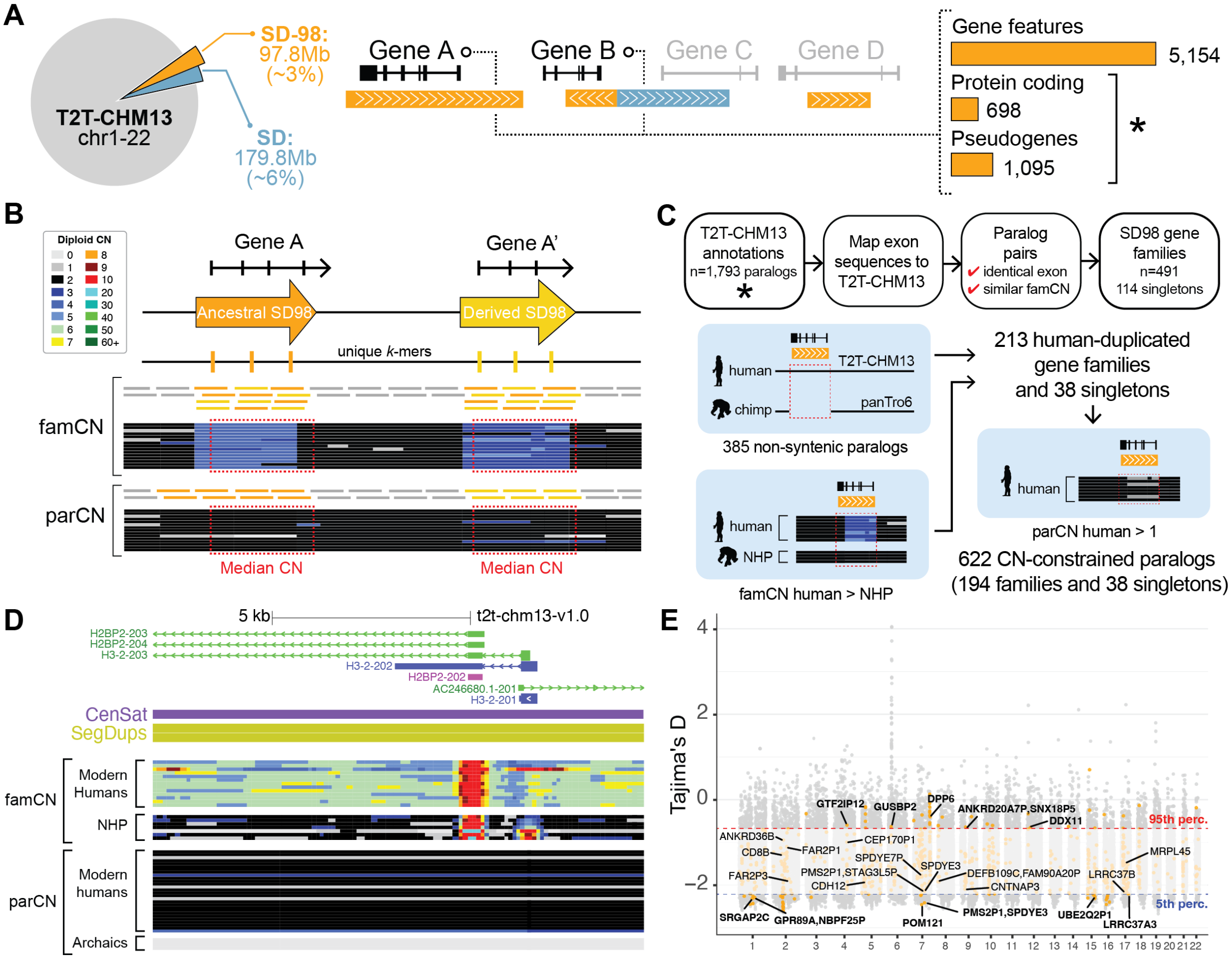

Figure 1. Human gene duplications in T2T-CHM13. (**A**) Diagram of segmental duplications (SDs) with >90% identity (blue) and >98% identity (orange) in T2T-CHM13 and selection of genes within SDs with >98% identity (SD98 genes). Total counts are shown on the right, with protein-encoding genes and pseudogenes used for further analysis indicated with an asterisk. (**B**) Schematic representation of copy number (CN) estimation methods, including gene-family CN (famCN) and paralog-specific CN (parCN). Illustrated horizontal lines represent short-read pileups mapping to unique (gray) and duplicated regions (orange and yellow). Read-depth diploid CN estimates are shown as heatmaps with values explained in the legend (left). The CN-genotyping window is shown as red dashed boxes. (**C**) Pipeline for clustering and stratification of SD98 genes. Gene families were classified as carrying human duplicates based on synteny with the chimpanzee reference genome (panTro6) and famCN comparisons between human and nonhuman primates (NHPs) (left). CN-constrained (fixed or nearly fixed) genes were flagged

based on parCN values across human populations (right). (**D**) UCSC Genome Browser snapshot of the *H3-2*/*H2BP2* locus, including gene models, centromeric satellites (CenSat), SDs (SegDup), and famCN and parCN predictions across modern humans, NHPs, and archaic genomes. (**E**) Distribution of Tajima's D values (y-axis) from individuals of African ancestry from the 1KGP across 25-kbp windows genome wide (gray) and in the SD98 region (orange) across human autosomal chromosomes (x-axis). All human-duplicated gene names with outlier D values in African, European, East Asian, South Asian, and American populations are included.

Human-duplicated genes implicated in brain development

Connecting genetic variation of duplicated genes with neural traits

To narrow in on human-duplicate gene families contributing to neurocognitive features, we identified 187 genes with putative associations with brain-related phenotypes from the genome-wide association study (GWAS) catalog and UK Biobank [56] (Tables S1 and S4). Three variants (rs12725078, rs17537178 and rs4797876) associated with sulcal depth impact *SRGAP2*, *PTPN20*, and *ROCK1P1*, respectively. The ancestral *CORO1A*, implicated in autism [57], is associated with brain morphology. Many implicated genes reside at genomic hotspots (n=58), such as *GPR89* paralogs at chromosome 1q21.1 with recurrent ~2 Mbp deletions/duplications impacting brain size [58]. While interesting, GWAS hits are significantly depleted across SD98 regions (Note S2), in part due to the common use of single-nucleotide polymorphism (SNP) arrays that lack coverage across SDs [59]. As such, we assayed variation in an autism cohort (Simons Simplex Collection [SSC]; n=2,459 quad families). Eighteen genes show significant parCN differences in probands versus unaffected siblings (Wilcoxon signed-rank test, q-value<0.05) (Figure S6), with all but one residing at chromosome 15q25.2 (OMIM: 614294), a region known to undergo recurrent deletions/duplications [60]. The remaining gene, pseudogene AC233280.19, is associated with the chromosome 3q29 genomic disorder (OMIM: 609425). *De novo* copy number variants (CNVs) impact 22 human-duplicated genes in autistic probands (Table S6, Figure S7); this contrasts with six events impacting five paralogs in unaffected siblings (Fisher's exact test, p-value = $4.5\text{x}10^{-4}$). Most impacted genes reside at known autism-associated genomic hotspots (n=15). The other seven, which were not mutated in unaffected siblings, included protein-encoding genes *CD8B2*, *FCGR1B*, *HYDIN* and *LIMS1*, representing possible contributors to autism spectrum disorder (ASD).

Duplicated gene expression in the developing human brain

Re-analyzing published RNA-seq datasets [21,61–64] using the new T2T-CHM13 reference, we found nearly half of human-duplicated gene paralogs (455/1,002) are expressed during brain development (TPM≥1) (Table S1, Figures 2A and S8), representing a depletion versus the genome-wide transcriptome (21,513/23,395). This increases to 58% for CN-constrained genes (1.3-fold enrichment, p-value = $2.5\text{x}10^{-24}$, hypergeometric test) and to 84% for CN-constrained protein-encoding genes (1.4-fold enrichment, p-value = $7.8\text{x}10^{-30}$, hypergeometric test) (Figure 2B). These results suggest true functional candidates are more likely to exist in the most CN-constrained protein-encoding genes (Figure 2C). In sum, 147 human-duplicated families carry at least one CN-constrained and brain-expressed gene, including 39 protein-encoding paralogs verified as human specific (non-syntenic). Of these, 21 genes are also expressed in the postnatal brain, including *CD8B2*, which is exclusively expressed after birth.

We next used the longitudinal CORTECON dataset [63], with transcriptomes of different stages of *ex-vivo*-induced neurogenesis from human embryonic stem cells, to infer developmental functions of genes using weighted gene co-expression network analysis (WGCNA) [65] (Figure 2D). Expressed genes (n=15,695) were clustered into 37 co-expression modules, each assigned a random color identifier. Thirty-two modules comprised SD98 genes (n=399), of which 200 paralogs represented human-duplicated families (55 non-syntenic) (Table S7, Figures 2E and S9). Comparing module assignment between paralogs found mostly differential expression patterns, with only six duplicate gene families in complete concordance (i.e., all paralogs in the same module) (Table S8, Figure 2F). This suggests that our approach largely distinguishes transcriptional profiles between similar paralogs, and that expression diverges at relatively short evolutionary time scales (<6 million years), as we have shown for a smaller set of genes [66].

Twenty-two of 35 modules were enriched for functional gene ontology (GO) terms (q-value<0.05, hypergeometric test; Table S9, Figures 2E and S9). To verify module assignments, we searched for duplicated genes with characterized functions. *ARHGAP11B*, which induces cortical neural progenitor amplification by altering glutaminolysis in the mitochondria [23], is a member of turquoise. Genes in this module are expressed highest during pluripotency and are associated with cell proliferation, including DNA replication and chromosome segregation, as well as mitochondrial gene expression. The hominoid-specific gene *TBC1D3*, known to promote basal progenitor amplification in the outer radial glia resulting in cortical folding in mice [24] is a member of purple, a module associated with regulation of neural differentiation. Human-specific *SRGAP2C*, which interacts with F-actin to produce membrane protrusions required for neuronal migration [67], represents blue with co-expressed genes that peak during cortical specification and upper-layer formation. This module is associated with cell motility, including motile cilium organization and assembly and microtubule-based movement.

We also found autism-associated genes [57] significantly enriched in four modules (yellow, black, saddle brown, and cyan), as well as the "unassigned" module (grey) (q-value < 0.05, hypergeometric test), and included 38 paralogs from human-duplicated gene families. Remarkably, three protein-encoding paralogs from the *RGPD* gene family, encoding RANBP2 Like And GRIP Domain Containing proteins, were represented in these modules, including human-specific *RGPD3* (yellow) and *RGPD4* (grey) as well as *RGPD8* (saddle brown). The yellow module, enriched with functions in axon guidance and synaptogenesis, contains the most autism-associated genes (n=20) (Table S7, Figure 2G). Genes in this module exhibit low expression during pluripotency, followed by sustained expression from neural differentiation to deep layer formation, including several markers of glutamatergic neurons (e.g., *SOX5*, *SLC1A6*, *OTX1*, and *TLE4*) [68]. Human-duplicate paralogs in the yellow module include *LRRC37B*, important in synapse function, as well as the causal gene in the chromosome 17q21.31 microdeletion syndrome, *KANSL1* [69]. We also identified compelling candidates residing at autism-associated genomic hotspots (e.g., *GOLGA6L9* and *GOLGA6L10* in chromosome 15q25.2, and *CASTOR2*, *PMS2P6* and *STAG3L1* in chromosome 7q11.23) (Table S1). Collectively, duplicated genes co-expressed with neural and ASD-associated genes representing top candidates contributing to human brain development.

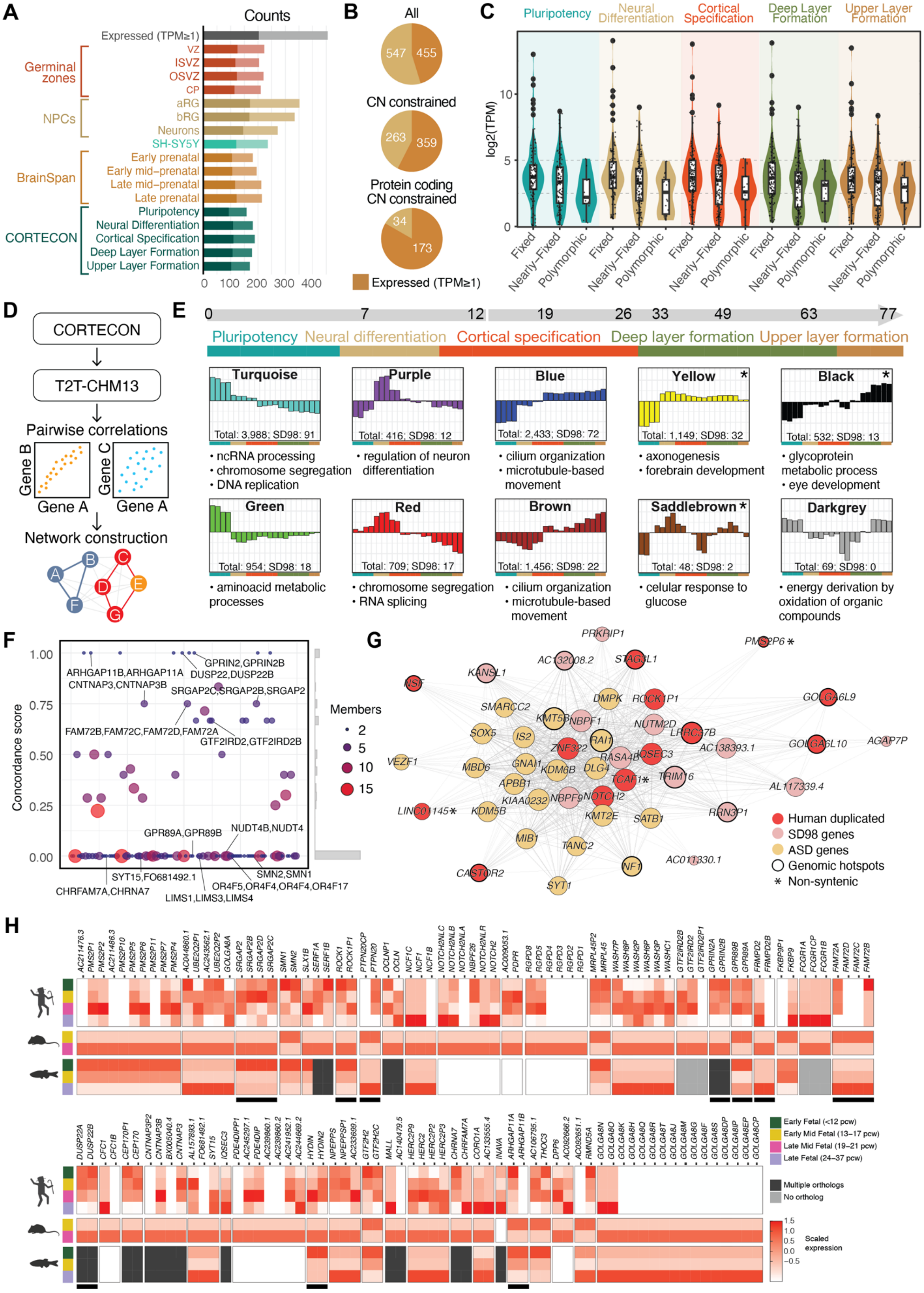
A
Counts
Expressed (TPM≥1)
Germinal zones
VZ
ISVZ
OSVZ
CP
NPCs
aRG
bRG
Neurons
SH-SY5Y
BrainSpan
Early prenatal
Early mid–prenatal
Late mid–prenatal
Late prenatal
CORTECON
Pluripotency
Neural Differentiation
Cortical Specification
Deep Layer Formation
Upper Layer Formation
0 100 200 300 400
B
All
547 455
CN constrained
263 359
Protein coding CN constrained
34 173
Expressed (TPM≥1)
C
Pluripotency
Neural Differentiation
Cortical Specification
Deep Layer Formation
Upper Layer Formation
log2(TPM)
0 5 10
Fixed
Nearly-Fixed
Polymorphic
D
CORTECON
T2T-CHM13
Pairwise correlations
Gene B
Gene A
Gene C
Gene A
Network construction
A B C D E F G
E
0 7 12 19 26 33 49 63 77
Pluripotency Neural differentiation Cortical specification Deep layer formation Upper layer formation
Turquoise
Total: 3,988; SD98: 91
• ncRNA processing
• chromosome segregation
• DNA replication
Purple
Total: 416; SD98: 12
• regulation of neuron differentiation
Blue
Total: 2,433; SD98: 72
• cilium organization
• microtubule-based movement
Yellow *
Total: 1,149; SD98: 32
• axonogenesis
• forebrain development
Black *
Total: 532; SD98: 13
• glycoprotein metabolic process
• eye development
Green
Total: 954; SD98: 18
• aminoacid metabolic processes
Red
Total: 709; SD98: 17
• chromosome segregation
• RNA splicing
Brown
Total: 1,456; SD98: 22
• cilium organization
• microtubule-based movement
Saddlebrown *
Total: 48; SD98: 2
• celular response to glucose
Darkgrey
Total: 69; SD98: 0
• energy derivation by oxidation of organic compounds
F
Concordance score
0.00 0.25 0.50 0.75 1.00
ARHGAP11B,ARHGAP11A
CNTNAP3,CNTNAP3B
GPRIN2,GPRIN2B
DUSP22,DUSP22B
SRGAP2C,SRGAP2B,SRGAP2
FAM72B,FAM72C,FAM72D,FAM72A
GTF2IRD2,GTF2IRD2B
GPR89A,GPR89B
NUDT4B,NUDT4
SYT15,FO681492.1
CHRFAM7A,CHRNA7
SMN2,SMN1
OR4F5,OR4F4,OR4F4,OR4F17
LIMS1,LIMS3,LIMS4
Members
2
5
10
15
G
PRKRIP1
PMS2P6 *
AC132008.2
STAG3L1
KANSL1
NSF
DMPK
ROCK1P1
GOLGA6L9
SMARCC2
KMT5B
NBPF1
NUTM2D
SOX5
IS2
RAI1
LRRC37B
GOLGA6L10
AGAP7P
ZNF322
AC138393.1
VEZF1
MBD6
GNAI1
KDM6B
DLG4
RASA4B
QSEC3
TCAF1 *
TRIM16
APBB1
AL117339.4
KIAA0232
NBPF9
NOTCH2
LINC01145 *
KDM5B
KMT2E
SATB1
RRN3P1
MIB1
TANC2
AC011330.1
NF1
CASTOR2
SYT1
Human duplicated
SD98 genes
ASD genes
Genomic hotspots
* Non-syntenic
H
Early Fetal (<12 pcw)
Early Mid Fetal (13–17 pcw)
Late Mid Fetal (19–21 pcw)
Late Fetal (24–37 pcw)
Multiple orthologs
No ortholog
Scaled expression
1.5 1.0 0.5 0.0 −0.5
AC211476.3 PMS2P1 PMS2P2 AC211486.3 PMS2P10 PMS2P5 PMS2P6 PMS2P11 PMS2P7 PMS2P4 AC044860.1 UBE2Q2P1 AC243562.1 UBE2Q2P2 GOLGA8A SRGAP2 SRGAP2B SRGAP2D SRGAP2C SMN1 SMN2 SLX1B SERF1A SERF1B ROCK1 ROCK1P1 PTPN20CP PTPN20 OCLNP1 OCLN NCF1C NCF1 NCF1B NOTCH2NLC NOTCH2NLB NOTCH2NLA NBPF26 NOTCH2NLR NOTCH2 AC009053.1 PDPR RGPD8 RGPD5 RGPD4 RGPD3 RGPD2 RGPD1 MRPL45P2 MRPL45 WASH7P WASH8P WASH2P WASH6P WASH3P WASHC1 GTF2IRD2B GTF2IRD2 GTF2IRD2P1 GPRIN2A GPRIN2B GPR89B GPR89A FRMPD2 FRMPD2B FKBP9P1 FKBP9 FCGR1A FCGR1CP FCGR1B FAM72A FAM72D FAM72C FAM72B
DUSP22A DUSP22B CFC1 CFC1B CEP170P1 CEP170 CNTNAP3P2 CNTNAP3B BX005040.4 CNTNAP3 AL157893.1 FO681492.1 SYT15 QSEC3 PDE4DIPP1 AC245297.1 PDE4DIP AC239860.1 AC239860.2 AC241952.1 AC244669.2 HYDIN HYDIN2 NPEPPS NPEPPSP1 AC233699.1 GTF2H2 GTF2H2C MALL AC140479.5 HERC2P9 HERC2 HERC2P3 CHRNA7 CHRFAM7A CORO1A AC133555.4 INAVA ARHGAP11A ARHGAP11B AC106795.1 THOC3 DPP6 AC092666.2 AC092651.1 RMND5A GOLGA8N GOLGA8O GOLGA8K GOLGA8H GOLGA8Q GOLGA8R GOLGA8T GOLGA8J GOLGA8M GOLGA8F GOLGA8S GOLGA8DP GOLGA8EP GOLGA8CP

Figure 2. Expression of human-duplicated genes during brain development. (**A**) Counts of human-duplicated genes with transcripts per million (TPM) >1 in fetal brain datasets including germinal zones (VZ: ventricular zone, ISVZ: inner subventricular zone, OSVZ: outer subventricular zone, CP: cortical plate), neuronal progenitor cells (NPCs) (aRGs: apical radial glia, bRGs: basal radial glia), neuroblastoma cell line (SH-SY5Y), the BrainSpan dataset, and the CORTECON dataset. Counts for protein-encoding genes are represented in darker shades. (**B**) counts of expressed (TPM≥1) (dark orange) and non-expressed (light orange) human-duplicated genes across gene categories. (**C**) SD98 gene expression in $\log_2$(TPM) in the CORTECON dataset, spanning pluripotency to upper layer formation and stratified by copy number (CN) category. (**D**) Pipeline used for the weighted gene coexpression analysis (WGCNA) of CORTECON data remapped to the T2T-CHM13 reference. (**E**) Selected WGCNA co-expression modules represented with random colors. Modules were organized based on their temporal expression spanning pluripotency to upper layer formation (day 0 to 77) with overrepresented gene ontology terms shown at the bottom. (**F**) Module assignment concordance scores are shown on the vertical axis for SD98 gene families, with spacing along the horizontal axis for visual separation. The size of each point corresponds to the number of members in the respective gene family. (**G**) Network diagram of yellow module. Only genes within human-duplicated gene families (red) or SD98 (pink) and autism-associated (yellow) categories with high module membership are depicted. Genes with asterisks are non-syntenic with chimpanzee [35] and bold borders are within ±500-kbp of a genomic disorder hotspot [60]. (**H**) Scaled TPMs from post-mortem human fetal brain samples from the BrainSpan dataset, and pseudo-bulk single-cell transcriptomes from whole-brain dissected samples of mouse and zebrafish. Gene families pictured represent a subset of CN-constrained and brain-expressed human-duplicated gene families. Genes with black bars beneath them were prioritized for additional characterization.

Modeling functions of duplicated genes in brain development

The next step in understanding the role of human-duplicated genes in brain development is to test their functions in model systems. Our combined analysis highlights 148 gene families with at least one CN-constrained or brain-expressed human-duplicated paralog, in addition to 30 paralogs not assigned to a family (Table S10). Of these, we found 106 with a homologous gene(s) in either mouse or zebrafish. Using matched brain-expression data from these species corresponding to human developmental stages [64,70,71] (Figure S10, as previously described [72,73]) narrowed in on 76 and 41 single-copy orthologs expressed during neurodevelopment in mice and zebrafish, respectively (Table S11), representing top candidates for functional studies. This leaves 40% of the human duplicate families with no obvious mouse/zebrafish ortholog, including fusion genes, primate-specific genes (e.g., *TBC1D3* paralogs [24,74]), or those associated with great ape ancestral "core" duplicons (e.g., *NBPF* and *NPIP*) [75]. Alternative models are required, such as *in vivo* primate or cell culture organoids, to test the functions of these genes.

Application of the resource: Characterizing candidate duplicated genes

Genetic variation of candidate genes important in neurodevelopment

As a proof of concept, we selected 13 priority human-specific duplicated (pHSD) gene families representing 30 paralogs from our model gene list (Table S12). Since none of the paralogs fully reside within short-read-accessible genomic regions due to their high identity (Table S1), we characterized variation using long-read sequencing. This included published draft assemblies of 47 individuals from the Human Pangenome Reference Consortium (HPRC) [76–78] and nine individuals from the Human Genome Structural Variation Consortium (HGSVC) [79] (112 total haplotypes; Figure S11). We also performed capture high-fidelity (cHiFi) sequencing on 178 individuals of diverse ancestries in the extended 1KGP cohort [55] and 22 individuals from the Human Genome Diversity Project (HGDP) [80] (Table S13, Figure S12, Note S3). Combined, we identified 46,754 variants (33,774 SNVs and 12,980 indels), or 12.7 variants/kbp, across captured regions (Table S14). Levels of variation within gene families were largely different between paralogs (Mann-Whitney U test, $p \leq 0.05$), with the exception of *FRMPD2* and *PTPN20*

(Figure S13). For instance, compared with the ancestral *SRGAP2* paralog, human-specific *SRGAP2B* exhibited the lowest and *SRGAP2C* the highest heterozygosity levels, in line with different mutation rates previously observed at each loci [17].

Functional annotation [81] identified 412 gene-impacting variants (missense = 252, synonymous = 131, likely gene-disruptive [LGD] = 29; Tables S15 and S16, Figure 3A), with eleven paralogs exhibiting no LGD variants suggesting strong selective constraint. To infer purifying selection, an indicator of function, we calculated the Ka/Ks statistic (also known as dN/dS) per gene family (Table S17). Virtually all paralogs had Ka/Ks lower than 1, and seven ancestral and three derived paralogs exhibited Ka/Ks below the genome-wide average (~0.25) [82]. The ancestral paralogs (Table S12) exhibited significantly lower Ka/Ks values than their derived paralogs (Wilcoxon signed-rank test, *p*-value=0.03) (Figure 3B), consistent with stronger purifying selection. To test for more recent selection signatures, we incorporated polymorphic variation to calculate pN/pS and the direction of selection (DoS) statistic [83], which similarly indicated stronger purifying selection in the ancestral versus derived paralogs (Wilcoxon signed-rank test, *p*-value=0.023) (Table S17, Figure 3C). While the tests mostly agree, *NPY4R* shows discordant signatures, being highly conserved according to Ka/Ks but approaching zero in DoS, in line with an excess of observed LGD variants suggesting recent neutral evolution. Most paralogs within gene families were under purifying selection, including *GPR89*, *CD8B*, *DUSP22*, *GPRIN2*, and *ARHGAP11* (also evident from a larger phylogenetic analysis of dN/dS using a maximum likelihood approach [84]; Table S18), although some show conservation in only one paralog, such as *ROCK1*. Human-specific *SRGAP2C* has elevated Ka/Ks and pN/pS, together with low Tajima's D (-2.32) in African individuals from the 1KGP genome-wide screen (Figure 1E), suggesting *SRGAP2C* is evolving under positive selection.

We verified selection signatures of pHSDs using high-confidence variants obtained from genome assemblies (n=56, HPRC/HGSVC) using nucleotide diversity π and Tajima's D. *SRGAP2C* again shows negative Tajima's D (-2.14) in AFR, validating genome-wide results (Figure S14). *GPR89* gene family paralogs, with low Ka/Ks, exhibit low nucleotide diversity and negative Tajima's D values across all exons consistent with functional constraints (Figure S15). In contrast, *ROCK1* showed reduced nucleotide diversity and more negative Tajima's D compared to *ROCK1P1*, consistent with their Ka/Ks values (Figure S16). While Ka/Ks was not calculated for *FAM72* paralogs due to a lack of synonymous polymorphisms, Tajima's D values similarly ranged from -2 to -1 indicating conservation of the gene family members (Figure S17).

Revisiting the 1KGP genome-wide signal of balancing selection in individuals of American (Tajima's D=0.12) and European ancestries (D=0.40) centered on *CD8B* (Tables S1 and S5, Figures 1E, 3D and S5), we find positive Tajima's D in American (max 2.66, n=18) but not in African ancestries (max 0.62, n=27) with three major peaks within the gene (Figure 3D). The ancestral *CD8B* paralog, encoding CD8 Subunit Beta, is highly expressed in T cells where the protein dimerizes with itself or CD8A (alpha) to serve as a cell-surface glycoprotein mediating cell-cell interactions and immune response [85,86]. Leveraging the assemblies, we identified two distinct haplotype clusters underlying the Tajima's D peaks, one of them particularly prevalent in individuals of American ancestry (Figure S18). Expanding to the entire long-read dataset (including cHiFi) shows an increase in intermediate-frequency variants, a signature of balancing selection, in *CD8B* among European and American ancestries, compared with those of African ancestry (Kolmogorov-Smirnov, *p*-value=$2.2\text{x}10^{-16}$) (Figure S19); these variants were verified as differentiating the two main haplotypes. Two of the SNPs (rs56063487 and rs6547706) are *CD8B* splice eQTLs in whole blood from GTEx [87] and significantly associated with increased CD8-protein levels on CD8+ T cells within a Sardinian cohort [88]. We note that *CD8B2* paralog-specific variants do not overlap with the SNPs, providing confidence in these short-read-based genotype results. The haplotypes may, thus, play a role in the modulation of the adaptive immune response, a frequent target of balancing selection. Alternatively, the human-specific paralog *CD8B2* exhibits divergent expression in the human postnatal brain rather than in T cells [38] (Figure 3E). These results provide an example of two paralogs with likely divergent functions and contrasting evolutionary pressures over a relatively short evolutionary time span (~5.2 million years ago [mya] [13]). Combined, we demonstrate the efficacy of long-read data to uncover hidden signatures of natural selection.

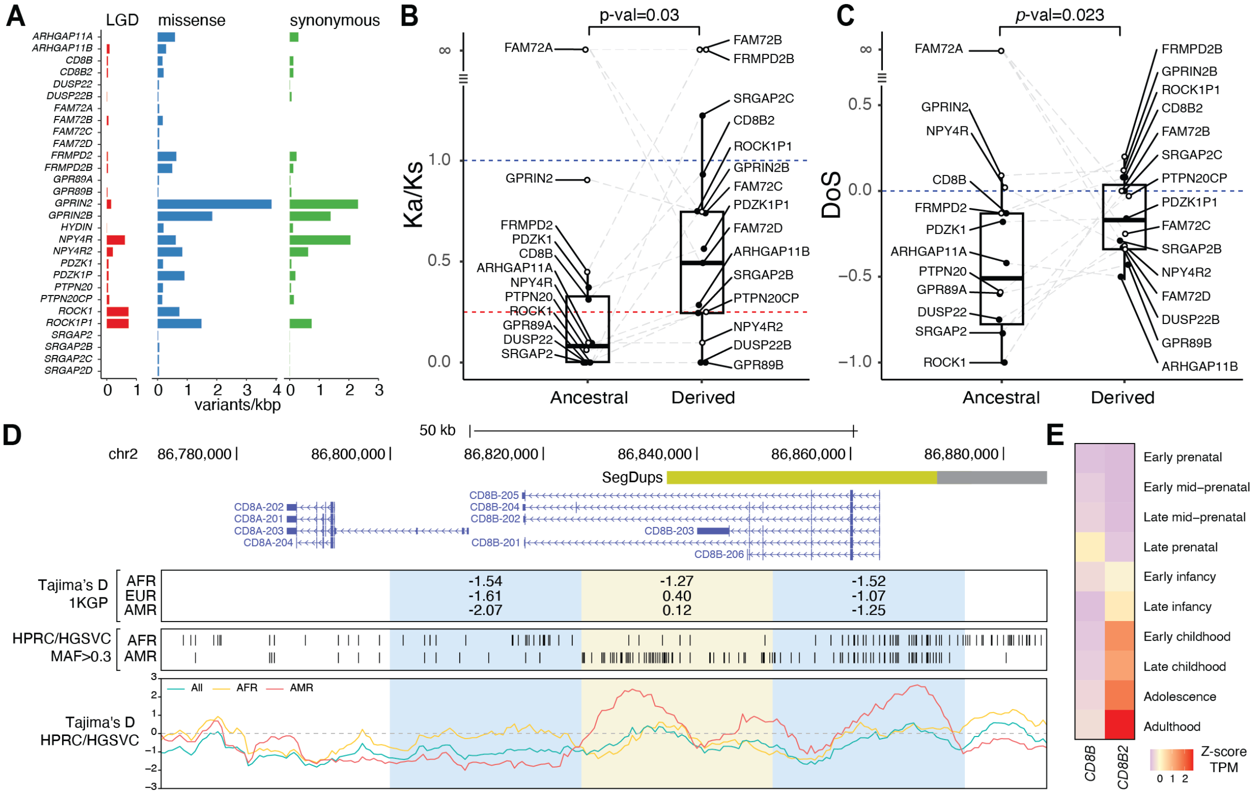

Figure 3. Genetic variation and signatures of selection of priority human-specific duplicated (pHSD) genes. (**A**) Number of likely gene-disruptive (LGD) (red), missense (blue), and synonymous (green) mutations identified in pHSD genes using long-read assemblies (n=56) and PacBio capture high-fidelity (cHiFi) sequencing (n=144). (**B**) Ka/Ks values calculated from human and chimpanzee sequences. Red dashed line indicates the average genome-wide Ka/Ks between humans and chimpanzees. Blue line indicates neutrality in the Ka/Ks test. Differences between the Ka/Ks of the matched ancestral and derived paralogs were tested with the Wilcoxon signed-rank test. (**C**) Direction of selection (DoS) values derived from Ka/Ks and pN/pS estimates. Blue line indicates the threshold for signatures of positive selection (positive values). Significant differences between ancestral and derived paralogs were obtained with the Wilcoxon signed-rank test. Paralogs with infinite values or undetermined ancestral/derived state (hollow dots) were excluded from Ka/Ks and DoS comparisons. (**D**) *CD8B* locus overview, including Tajima's D values derived from 1KGP SNVs in 25-kbp windows, biallelic SNPs with a minor allele frequency greater than 0.3 identified in African (AFR, n=27) and American (AMR, n=18) individuals using continuous assemblies from the Human Pangenome Reference Consortium (HPRC) and the Human Genome Structural Variation Consortium (HGSVC), and Tajima's D values derived from HPRC and HGSVC SNVs using 6-kbp windows and 500-bp steps for AFR, AMR, and all individuals. (**E**) Scaled transcript per million (TPM) expression of *CD8B* and *CD8B2* in postmortem brain tissue from BrainSpan.

Duplicated gene functions modeled using zebrafish

We performed a high-throughput functional screen in zebrafish [89–91] of seven largely uncharacterized pHSD families expressed in both human and zebrafish brain (*GPR89*, *NPY4R*, *PTPN20*, *PDZK1*, *HYDIN*, *FRMPD2,* and *FAM72*; Figure 2H and Table S11). Additionally, we tested two gene families (*SRGAP2* and *ARHGAP11*) previously studied in mammals [16,21–23,67,92–96]. Ancestral gene functions were assessed using loss-of-function knockouts for eight zebrafish orthologs by co-injecting SpCas9 coupled with four guide RNAs (gRNAs) targeting early exons resulting in ~70% ablation of alleles in G_0 lines [97] (termed crispants). The final gene, *arhgap11*, is maternally expressed (Figure S20) prompting us to use a morpholino that impedes translation. We also 'humanized' zebrafish models by introducing transiently *in vitro* transcribed 5'-capped mRNAs encoding human-specific paralogs (Figure 4A) for all genes except *HYDIN2*, due to its large size (4,797 amino acids). There were no significant morbidity differences in mutants compared to controls (log-rank survival tests *p*-values > 0.05, Table S19).

To assay morphology differences in mutant zebrafish, we acquired images [98–100] for 3,146 larvae at 3 and 5 days post-fertilization (dpf) (average of 75±55 larvae per group, Table S20). We first used latent diffusion and convolutional neural networks (CNNs) to test for significant morphological alterations between mutant models and controls without predefining specific features *a priori* (Methods). Both knockout and humanized models of *SRGAP2*, *GPR89*, *FRMPD2*, and *FAM72* exhibited significant differences (F1 scores > 0.2, Figure 4B). Altered features were identified by quantifying body length, head area, and the head-trunk angle, a classic measurement for developmental staging of zebrafish using the same images. This revealed concordant phenotypes for knockout and humanized models of *SRGAP2* (reduced length), and *FRMPD2* (reduced head area), and *FAM72* (both reduced body length and head area) at 3 dpf (Table S21, Figure 4C). Alternatively, *GPR89* models exhibited opposing effects, with head area for *gpr89* knockout larvae ~10% reduced and *GPR89B* 'humanized' larvae ~15% increased. This is also evident in the feature attribution plot indicating that the CNN distinguishes both *gpr89* knockout and *GPR89B* humanized larvae from controls primarily by focusing on the head (Figure 4B). At 5 dpf, the alterations in *FRMPD2* and *SRGAP2* models persisted while no longer observed for *FAM72* and *GPR89* (Table S21, Figure 4C). Knockout models for *gpr89* and *frmpd2* also displayed evidence of developmental delay with subtle yet significant decreases in the head-trunk angle (Table S21).

We next performed single-cell RNA-sequencing (scRNA-seq) [101,102] of dissected heads of 3 dpf larvae to directly characterize impacts on brain development, profiling 95,555 cells (an average of 3,822±3,227 per model) (Figure 4D). Pseudo-bulk differential expression analysis using all cells in each model revealed significant correlations in gene expression changes versus controls between knockout and humanized models (Figure 4E). Positive correlations for *SRGAP2C*, *FAM72B*, *ARHGAP11B*, *FRMPD2B*, and *PDZK1B* humanized larvae with respect to each knockout indicate loss-of-function effects. *GPR89B* gene expression changes are negatively correlated with *gpr89* indicating gene dosage effects, while *PTPN20CP* and *NPY4R2* show low/no relationship between models. These results are in line with our morphometric findings for *SRGAP2*, *FRMPD*2, *FAM72*, and *GPR89* (Figure 4C), as well as from our separate study [103] that verified the human SRGAP2C protein physically interacts with and antagonizes zebrafish Srgap2.

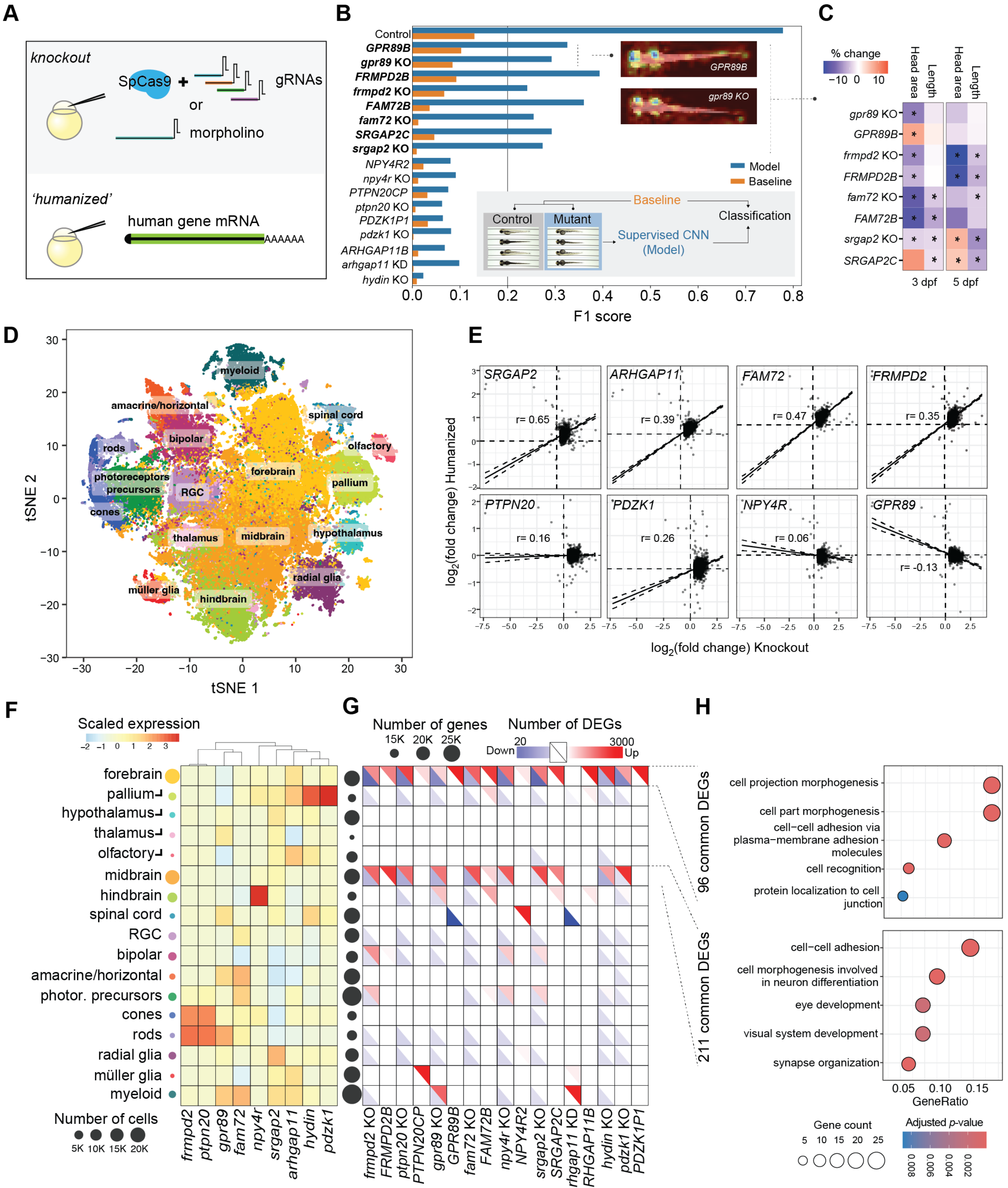

Figure 4. Functional evaluation of selected pHSDs using zebrafish. (A) Functions of each pHSD gene were tested by generating knockout (co-injection of SpCas9 coupled with four gRNAs targeting early exons, or morpholino for *arhgap11*) and 'humanized' models (injection of the human-specific mRNA). **(B)** Morphological assessment using a supervised convolutional neural network (CNN) to distinguish models from matched controls (bottom inset) obtained at 3 dpf and 5 dpf. F1 score indicates the effect size of difference between models and controls and ranges from 0 to 1, where 0 indicates that no sample from that group could be distinguished from the controls. Orange bars indicate the null hypothesis that there is no difference between models and controls. A

threshold F1 score of 0.2 was used to define pHSD groups being robustly classified as different from their control group. Pictured as a top inset are feature attribution plots for two example *GPR89B* and *gpr89* knockout larvae, highlighting the region of the image used by the CNN to correctly classify and distinguish those genotypes from controls. Colors range from red (region is not used for classification; zero gradient), to orange, then blue (region contributes the most to classification; large magnitude gradient). **(C)** Percent change compared to the control group for standard length or head area across selected pHSD models. Asterisks indicate a Benjamini Hochberg-corrected *p*-value below 0.05. **(D)** t-distributed stochastic neighbor embedding (tSNE) plot highlighting the classified 17 cell types from the 95,555 harvested cells across pHSD models at 3 dpf. **(E)** Fold-change comparison between knockout and 'humanized' models for each pHSD across all genes (n= 29,945), versus their controls. Black lines represent the Pearson correlation line and the dotted lines the 95% confidence intervals. **(F)** Endogenous z-score scaled expression of each zebrafish ortholog across defined scRNA-seq cell types. Circle sizes scale with the overall number of cells included in that group. **(G)** Distribution of cell-type-specific differentially expressed genes (DEGs) for each pHSD model. Each square includes the downregulated genes in blue (lower diagonal) and upregulated genes in red (upper diagonal). Circles next to each cell type represent the number of expressed genes. **(H)** Gene ontology results for the common DEGs in forebrain (n= 96) and midbrain (n=211) across pHSD models, with circles representing DEG number in the GO term and color representing the *q*-value.

We classified 17 different neuronal, retinal, and glial cell types using gene markers [71,101,104,105]. While most pHSD orthologs were broadly expressed across cells, a subset showed more narrow expression in specific cell types (e.g., *hydin* and *pdzk1* in the pallium, *npy4r* in the hindbrain; Figure 4F). We repeated pseudo-bulk differential expression analyses across specific cell types revealing gene dysregulation in the forebrain and midbrain across most pHSD models (16 out of 17, Figure 4G). Common differentially expressed genes (DEGs) in the forebrain functioned in cell projection, adhesion, and recognition, while DEGs in the midbrain related to neuronal differentiation and the visual system (Figure 4H). The zebrafish forebrain is the closest related structure to the human cerebral cortex [106], while the midbrain primarily includes the optic tectum [107], the main visual processing center. Some models also highlighted DEGs in specific cell types, including Müller glia in humanized *PTPN20CP*, the spinal cord in humanized *NPY4R2*, and myeloid cells in *gpr89* and *arhgap11* knockout larvae (Figure 4G). Combined, these results indicate that all tested pHSD models impact the developing zebrafish brain, suggesting that they may also play important roles in human brain evolution.

Novel human-specific genes impacting neurodevelopment

<u>*GPR89B* and brain size</u>

Opposite phenotypes were observed for *gpr89* knockout and humanized *GPR89B* zebrafish suggesting gene dosage effects. Considering both *GPR89* human paralogs are impacted by deletions and duplications at the chromosome 1q21.1 genomic hotspot associated with microcephaly and macrocephaly in children with neurocognitive disabilities, respectively [58], we sought to characterize mechanisms underlying larval head-size phenotypes in more detail. We first verified that stable *gpr89* heterozygous and homozygous knockouts exhibited reduced head size at 3 dpf, consistent with crispants. Using a neuronal reporter line Tg(HuC-eGFP) [108], we generated *GPR89* mutant models finding significantly smaller and larger forebrains in knockout and humanized larvae, respectively (Figures 5A). Re-examining scRNA-seq data, we sub-clustered cells from the forebrain and observed endogenous expression of *gpr89* in telencephalon and inner diencephalon (Figure 5B). Focusing on the telencephalon, a brain structure anatomically equivalent to the mammalian forebrain with roles in higher cognitive functions such as social behavior and associative learning [109,110], we performed pseudo-bulk differential expression analysis. DEGs with inverse effects were enriched in negative regulation of the DNA replication and cell cycle (Figure 5C, Tables S22 and S23). Several genes functioning at the G2/M checkpoint were downregulated in the humanized *GPR89B* and upregulated in the knockout *gpr89* pointing to differences in cell proliferation. To test this, we estimated the identity of forebrain cells based on the expression of known markers for neural progenitors (*sox19a*, *sox2*, *rpl5a*, *npm1a*, *s100b*, *dla*) and differentiated neurons (*elavl3*, *elavl4*, *tubb5*). This found humanized *GPR89B* cells more likely to classify as progenitors while *gpr89* knockouts more likely to be differentiated (Figure 5D).

GPR89 (G-protein receptor 89 or *GPHR*, Golgi PH regulator) encodes highly conserved transmembrane proteins that participate in intracellular pH regulation in the Golgi apparatus [111]. Loss of function in *Drosophila* leads to global growth deficiencies as a result of defects in the secretory pathway [112]. In humans, a complete duplication of ancestral *GPR89A* ~4.7 mya produced the derived, full-length *GPR89B* [13] (Figure 5E). The two paralogs maintain identical protein similarity but differential and overlapping expression patterns in human brain development, with *GPR89A* highly expressed starting in pluripotency (turquoise module), and *GPR89B* expression turning on slightly later during neural differentiation (red module; Figures 2E and 5E). Both genes are under purifying selection (Figure S14), with *GPR89A* exhibiting extreme negative Tajima's D values in individuals of AFR and AMR ancestries from the 1KGP cohort (<5th percentile; Figure 1E, Table S1). These results provide evidence that both *GPR89* paralogs function in early brain development, possibly with delayed expression of *GPR89B* extending expansion of progenitor cells, a feature observed in human cerebral organoids compared with those of other apes [113,114] (Figure 5E). Together with the increase in forebrain size of "humanized" zebrafish, this suggests a role for *GPR89B* in contributing to the human-lineage expansion of the neocortex.

FRMPD2B and synaptic signaling

While opposing traits were observed in *GPR89* models, similar phenotypes impacting head area and body length suggest that the human FRMPD2B acts as a dominant negative to the endogenous Frmpd2. Validating phenotypes observed in crispants (Figure 4C), we observed reduced head size in stable *frmpd2* homozygous knockout larvae (Figure 5A). Additionally, both the crispant knockout *frmpd2* and humanized *FRMPD2B* larvae exhibit smaller forebrains. We found that shared upregulated DEGs function in cell/axon morphogenesis and growth as well as synaptic signaling in telencephalic cells (Figure 5C, Tables S24 and S25). To better characterize impacts on synaptic signaling, we used motion-tracking [115] to detect seizure susceptibility in mutant zebrafish. Treatment with a low dose of the GABA-antagonizing drug pentylenetetrazol (PTZ) produced a significant increase in high-speed events, indicative of seizures in larvae, in both *FRMPD2* mutant models (4 dpf) versus controls (Figure 5F). These results suggest that Frmpd2 loss of function, through *frmpd2* knockout or antagonism via FRMPD2B, disrupts excitatory synapse transmission which amplifies induced seizures, in line with the known interactions of FRMPD2 with glutamate receptors [116].

FRMPD2 (FERM and PDZ domain containing 2) encodes a scaffold protein that participates in cell-cell junction and polarization [117]. Protein localization has been observed at photoreceptor synapses [118] and the postsynaptic membrane in hippocampal neurons in mice [116]. A partial duplication of the ancestral *FRMPD2* on human chromosome 10q11.23 created the 5'-truncated *FRMPD2B* paralog ~2.3 mya [13]. This shorter *FRMPD2B*-derived paralog encodes 320 amino acids of the C-terminus, versus 1,284 amino acids for the full-length ancestral, maintaining two of three PDZ domains involved in protein binding [119] while lacking the KIND and FERM domains (Figure 5G).

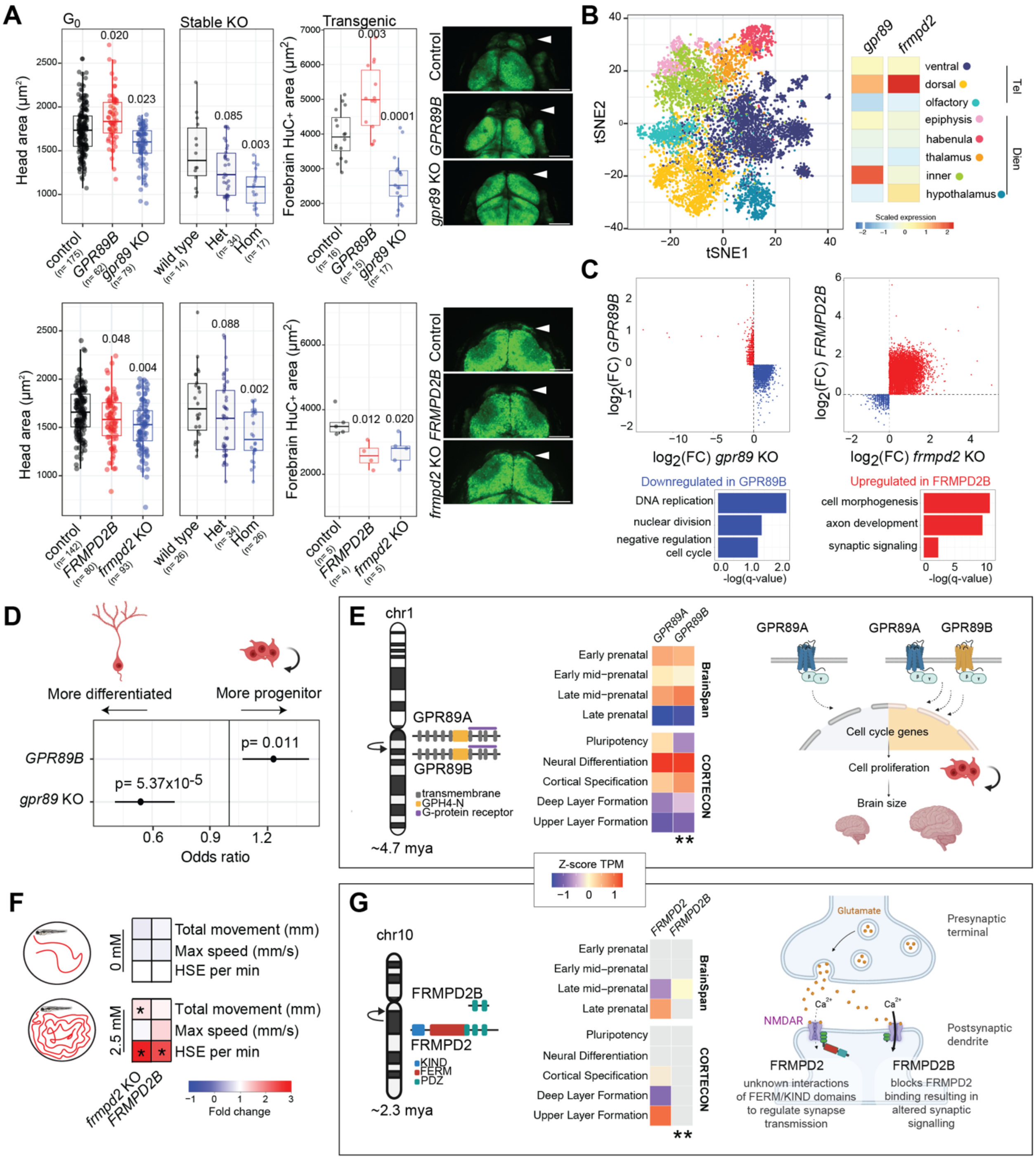

Figure 5. Neurodevelopmental impact of *GPR89* and *FRMPD2*. (A) Head and brain area assessments at 3 dpf for G$_0$ crispants and stable knockout lines for *GPR89* (top) and *FRMPD2* (bottom) models. Results for head area of *GPR89* crispants (ANOVA *p*-values: controls vs. *GPR89B*= 0.020, controls vs. *gpr89* knockouts= 0.023) and stable knockout lines (Wilcoxon signed-rank tests *p*-values: controls vs. Het= 0.085, controls vs. Hom= 0.003), as well as forebrain area of crispants using a transgenic line with fluorescently tagged neurons (ANOVA *p*-values: controls vs. *GPR89B*= 0.003, controls vs. *gpr89* knockouts= 0.0001). Results for head area of *FRMPD2* crispants (ANCOVA *p*-values: controls vs. *FRMPD2B*= 0.048, controls vs. *gpr89* knockouts= 0.004) and stable knockout lines (Wilcoxon signed-rank tests *p*-values: controls vs. Het= 0.088, controls vs. Hom= 0.002), as well as forebrain area of crispants using a transgenic line with fluorescently tagged neurons (ANOVA *p*-values: controls vs. *FRMPD2B*= 0.012, controls vs. *frmpd2* knockouts= 0.020). Representative images of each model in the neuronal transgenic line are

included with scale bars representing 100 µm. **(B)** t-distributed stochastic neighbor embedding (tSNE) plot showing the identified subregions classified from the forebrain (n=10,040 cells) and relative scaled endogenous expression of *gpr89* and *frmpd2* across cell types. **(C)** Log_2 fold change (FC) of gene expression versus controls in cells from the telencephalon between knockout and humanized models in *GPR89* and *FRMPD2*. Red and blue colors correspond to DEGs discordant (*GPR89*) or concordant (*FRMPD2*) between the knockout and humanized models and their top representative gene ontology enrichment analyses results. **(D)** Forest plot with the results from the logistic regression for presence of progenitor versus differentiated states in forebrain cells across *GPR89* models. **(E)** Diagram of the duplication event of *GPR89* giving rise to *GPR89A* and *GPR89B*, encoding two identical proteins with different expression patterns in both neurodevelopmental timing (BrainSpan) and brain regions (CORTECON) (**Wilcoxon signed-rank test, *p*-value < 0.005). A model of GPR89B gain-of-function in neuronal proliferation amplification is depicted on the right. **(F)** Behavioral results from 1 h motion-tracking evaluations in 4 dpf larvae exposed (2.5 mM) or not (0 mM) to pentylenetetrazol (PTZ). Metrics compared included total movement (mm), maximum speed (mm/s), and frequency of high-speed events (≥28 mm/s). Colors represent the fold change relative to the control group and the asterisk indicates a significant Dunn's test (BH-adjusted *p*<0.05). **(G)** Diagram of the duplication event of *FRMPD2* ~2.3 mya that gave rise to the 5'-truncated *FRMPD2B*, which exhibits different temporal (BrainSpan) and spatial (CORTECON) expression patterns (**Wilcoxon signed-rank test, *p*-value < 0.005). A model of FRMPD2B antagonistic functions resulting in altered synaptic signaling is depicted on the right.

Our data shows ancestral *FRMPD2* expressed in the human prenatal cortex during upper layer formation, while *FRMPD2B* is evident only postnatally [64] (Figure 5G). The paralogs also show divergent evolutionary signatures, with the full-length *FRMPD2* strongly conserved and the truncated *FRMPD2B* exhibiting possible positive selection (Figure 3B,C). Results in zebrafish show that loss of Frmpd2 function results in microcephaly and enhanced excitatory synaptic signaling. Combined, we propose a model in which truncated human-specific FRMPD2B counteracts the function of full-length FRMPD2 leading to altered synaptic features in humans, possibly through interactions of its PDZ2 domain with GluN2A of NMDA receptors at the postsynaptic terminal [116]. Its postnatal expression would avoid the detrimental effects of inhibiting FRMPD2 during early fetal development (i.e., microcephaly). We note that recurrent deletions and duplications in chromosome 10q11.21q11.23 impact both paralogs in children with intellectual disability, autism, and epilepsy [120]. Ultimately, *FRMPD2B* could plausibly contribute to the upregulation of glutamate signaling and increased synaptic plasticity observed in human brains compared with other primates that is fundamental to learning and memory [121].

Discussion

Our results provide the scientific community with a prioritized set of hundreds of genes to perform functional analyses with the goal to identify drivers of human brain evolution. Using a complete T2T-CHM13 reference genome, we present the most comprehensive detection of human duplicate genes to date with 213 families and 1,002 total paralogs. Compared to a previous assessment of human-specific duplicated genes [13], this represents an approximately fivefold increase in identified genes, in part because we also included human-expanded gene families and genes with as little as one duplicated exon. We note that these numbers are likely an underestimate, as we excluded 193 high-copy gene families (famCN>10), as well families that have undergone independent gene expansions or incomplete lineage sorting with other great apes. One compelling example is *FOXO3*, encoding the transcription factor forkhead box O-3, implicated in human longevity [122], with all three paralogs CN-constrained and brain expressed (Table S1). Since this gene also exists as duplicated in other great apes at similar CN, we excluded it from our list of human gene expansions. This is, in part, because there is still uncertainty regarding which paralog(s) are human specific for many of the gene families. SDs are often accompanied by secondary structural rearrangements that hamper synteny comparisons across species [57,123]. Moving forward, the availability of nonhuman primate T2T genomes will improve orthology and synteny comparisons between species [124–126] revealing additional human-specific paralogs. As a resource for the community, we have made available the results of our genome-wide analyses across the complete 1,793 SD98 genes (Tables S1-S11).

Collectively, 148 gene families (362 paralogs, 108 annotated as non-syntenic with the chimpanzee reference) represent top candidates for contributing to human-unique neural features based on at least one gene member exhibiting functional constraint across modern humans (1KGP) and brain expression (Table S1). In this study, we chose zebrafish to demonstrate the efficacy of our gene list. Despite notable differences with humans, such as the absence of a neocortex [127], conservation in major brain features make zebrafish well-suited to characterize gene functions in neurological traits, including cranial malformations [128], neuronal imbalances [129], and synaptogenesis [130]. Coupled with CRISPR mutagenesis [89,90], zebrafish have been used as a higher-throughput model for human neurodevelopmental conditions such as epilepsy [115], schizophrenia [131], and autism [73]. While a whole-genome teleost duplication resulted in ~20% of genes with multiple zebrafish paralogs that confounds functional analysis of human gene duplications [132], the nine prioritized gene families tested here were selected in part because each had only one zebrafish ortholog. We characterized gene functions by knocking out the conserved ortholog and introducing the human-specific paralog into developing embryos. Transient availability of the human transgene by injection of *in-vitro*-transcribed mRNA limited our analysis to early developmental traits (up to 5 dpf in zebrafish), approximately equivalent to human mid- to late-fetal stages in brain development (Figure S10). In the future, it will be important to characterize phenotypes in adolescent and adult zebrafish by generating stable transgenic humanized lines.

From our analysis, knockout and humanized models of four genes (*GPR89*, *FRMPD2*, *FAM72*, and *SRGAP2*) resulted in altered morphological features, primarily to head size (often used as a proxy for brain size), and all models exhibited molecular differences in single-cell transcriptomic data, most evident in the fore- and midbrains of larvae (Figure 4G). Two duplicate gene families, *SRGAP2* and *ARHGAP11*, have been extensively studied in diverse model systems (reviewed recently [9]). Our zebrafish model of *SRGAP2*, encoding SLIT-ROBO Rho GTPase-activating protein 2, were consistent with published findings in mouse where the 3'-truncated human-specific SRGAP2C inhibits the function of the endogenous full-length Srgap2 [16]. Further, the shared upregulated genes identified in the forebrains of *SRGAP2* mutant larvae point to alterations in axonogenesis and cell migration, matching studies in mice [11,16,17,67,93,133,134] (Table S26). Alternatively, *ARHGAP11B*, encoding Rho GTPase Activating Protein 11, implicated in the expansion of the neocortex through increased neurogenesis [21,23], exhibited no detectable changes in head/brain size when introduced in zebrafish embryos. Upregulated DEGs were only detected in the forebrains of *ARHGAP11B*-injected mutants and were enriched in cellular biosynthetic processes (mRNA splicing and translation; Table S27). Given that ARHGAP11B impacts the abundance of basal progenitors, a cell type unique to the mammalian neocortex [135], zebrafish may not be suitable to characterize human-specific functions of this gene.

Beyond modeling gene functions, our study also highlighted the considerable amount of genetic variation hiding within SD regions. Even with the resolved gaps and errors across SDs in T2T-CHM13, short-read sequencing is still insufficient to identify variation. Due to high sequence identity, only 10% of SD98 regions are “accessible” to short reads [36] resulting in <10% sensitivity to detect variants (Note S1) and a depletion of GWAS hits (Note S2). Using existing assemblies (HPRC and HGSVC) and cHiFi sequencing of individuals of diverse ancestry uncovered some of this hidden variation within our 13 pHSD gene families. We note that, for some of the most highly identical duplicated genes (*CFC1*), our cHiFi reads (~3 kbp) were still too short to accurately map to respective paralogs (data not included). Nevertheless, long reads revealed that most pHSD paralogs exhibit evolutionary constraints and provided support for balancing selection of *CD8B*, not previously identified in published genome-wide screens [136,137]. Historically, signatures of balancing selection, which include an excess of mid-frequency alleles [138], have been difficult to detect within SDs due to assembly errors [36]. In these cases, paralog-specific variants are mistaken for SNPs when reads from both paralogs map to a single collapsed locus resulting in false mid-frequency alleles. Scientific consortia like *All of Us* are generating long-read datasets at scale [139], ushering in a new era where genomic associations and evolutionary selection may finally be uncovered within human duplications to identify novel drivers of human traits and disease.

Similarly, genome sequencing of patients and their families has discovered hundreds of compelling neuropsychiatric disease candidate genes impacted by rare and *de novo* variants, but the genetic risk underlying conditions such as autism is still not completely elucidated [140]. SD genes may represent a hidden contributor to disease etiology. Our analysis identified 82 SD98 genes (38 human duplicate paralogs) co-expressed in modules enriched for ASD genes (Figure 2E), including several within disease-associated genomic hotspots. Distinct SD mutational mechanisms, including ~60% higher mutation rate compared to unique regions [141] and interlocus gene conversion that can occur between paralogs [142,143], make duplicated genes particularly compelling to screen for *de novo* mutations contributing to idiopathic conditions. For example, nonfunctional paralogs with truncating mutations can “overwrite” conserved functional paralogs leading to detrimental consequences, as is the case of *SMN1* and *SMN2* in spinal muscular atrophy [39]. Human-duplicated gene families include ancestral paralogs *CORO1A*, *TLK2*, and *EIF4E*, with significant genetic associations with ASD [57]. We propose that interlocus gene conversion between their likely nonfunctional duplicate counterparts is an understudied contributor to neurodevelopmental conditions in humans. Our comprehensive list of gene families will enable future work to progress in this research area.

Our study focuses on duplicate genes functioning in brain development, but primates exhibit other prominent differences across musculoskeletal and craniofacial features that have diverged early in human evolution [4]. Since such traits are largely universal across modern humans, our list of CN-constrained genes represent top candidates though re-analysis of transcriptomes from non-brain cells/tissues is required. Meanwhile, duplicate genes, such as those encoding defensins [144–147], mucins [148,149], and amylases [41–43], can also play a role in metabolism and immune response that exhibit population diversification due to the vast variability in diet, environment, and exposures to pathogens across modern humans [27]. Our use of a single complete human T2T-CHM13 haplotype of largely European ancestry [34] could miss some of these CN polymorphic genes. As additional T2T genomes are released [28], it will be important to continue curating our list of duplications. Nevertheless, genes CN stratified by human ancestry can be identified using metrics such as V_{st} [150], as has been highlighted in other studies (reviewed here [9] and most recently in a preprint [151]). Facilitating such analyses for our gene set, we provide a publicly available resource to query parCN median estimates across individuals from 1KGP for our complete set of SD98 paralogs (https://dcsoto.shinyapps.io/shinyc).

One notable limitation of our study is its reliance on existing gene annotations. We attempted for the first time to group human duplicate paralogs into larger multigene families based on shared sequences between annotated genes in SD98 regions. Due to the complexities of SDs, which can result in gene fusions and altered gene structures, some genes were left unassigned to a family (n=114 singletons from SD98 genes). Other noncoding transcripts and lncRNAs were excluded altogether, including a human-specific paralog of *IQSEC3*, a gene implicated in GABAergic synapse maintenance [152]. Additionally, the functional consequences of variants identified in 656 unprocessed pseudogenes are difficult to interpret. Improvements are on the horizon, with ongoing work with long-read transcriptomes that will continue to refine annotations [153] and advancements in protein-prediction models [154] and proteomic approaches [155] that will confirm whether or not these genes encode proteins.

In summary, we identified and featured two genes with strong evidence of contribution to human brain evolution: *GPR89B*, with a possible role in expansion of the neocortex, and *FRMPD2B*, with implications in altered synaptic excitatory signaling. Taking advantage of long-read sequencing in tandem with the new T2T-CHM13 reference genome, we interrogated challenging regions of the genome and demonstrated a method using zebrafish to explore the functions of human-duplicated genes. Among our list of hundreds, we propose that there are additional gene drivers that contribute to unique features of the human brain. In the future, additional genetic analyses across modern and archaic humans and experiments utilizing diverse model systems will reveal hidden roles of these genes in human traits and disease.

Methods

Identification of SD98 genes

Duplicated regions were extracted from previously annotated SDs [156] using T2T-CHM13 (v1.0) coordinates and subsequently merged using BEDTools merge [157]. SD98 regions were defined as an SD with ≥98% sequence identity to another locus in the T2T-CHM13 genome using the fractMatch parameter. Gene coordinates were obtained from T2T-CHM13 (v1.0) CAT/Liftoff annotations (v4) [34]. SD98 genes were defined as gene annotations that contain at least one exon fully contained within an SD98 region, calculated with BEDTools intersect using -f 1 parameter [157]. Overall numbers of distinct gene features overlapping SD98 were counted using the gene ID unique identifiers. We noticed that, in a few cases, two transcript isoforms of the same gene were assigned to different gene IDs. To identify these redundant transcripts, we self-intersected SD98 transcripts, selected those with different gene ID that also shared >90% positional overlap, and performed manual curation of the obtained gene list, removing redundant and read-through fusion transcripts.

Gene family clustering

SD98 genes were grouped into gene families based on shared exons (Figure S1). Starting from T2T-CHM13 (v1.0) annotations, DNA sequences of all SD98 regions were extracted using BEDTools getfasta and mapped back to the reference genome using minimap2 (v2.17) with the following parameters: -c --end-bonus 5 --eqx -N 50 -p 0.5 -t 64. For each SD98 exon, the BEDTools intersect with -f 0.99 parameter was used to select mappings covering >99% of the exon sequence, removing self-mappings. This list was refined using the previously published [35]whole-genome shotgun sequence detection (WSSD) [10] CNs (famCN) of humans from the SGDP (n=269), which provides estimates of the overall CN of a gene family using read depth of multi-mapping reads with nonoverlapping sliding-windows. After comparing the median famCN values of SD98 genes with shared exons, groupings where the mean absolute deviation of the CN was less than one were selected. The list was filtered to focus on gene families containing at least one protein-coding or unprocessed pseudogene. SD98 genes associated with other gene features, including lncRNAs and processed pseudogenes, were also assigned a gene family ID. On the other hand, if a gene was not associated with any other gene feature, they were classified as "unassigned" or "singletons". SD98 gene families were intersected with previously published DupMasker annotations using BEDTools intersect, which indicate ancestral evolutionary units of duplication [35].

Identification of human-duplicated genes families

Human-specific and -expanded gene families were identified using CN comparisons between humans and nonhuman great apes with previously published WSSD [10] (famCN) CNs from humans (SGDP n=269) and four nonhuman great apes, including one representative of chimpanzee (Clint), bonobo (Mhudiblu), gorilla (Kamilah), and orangutan (Susie) [35], mapped to T2T-CHM13 (v1.0). The median famCN per SD98 gene was calculated using a custom Python script. For each SD98 gene, putative gene family duplications and expansion were predicted, excluding genes with median famCN>10 across humans from this analysis. Genes were considered expanded if the median famCN across humans was greater than the maximum famCN across great apes. Human duplications and expansions were distinguished based on whether the maximum famCN value across great apes was less than 2.5 (non-duplicated in great apes) or greater than 2.5 (duplicated in great apes), respectively. Non-syntenic paralogs between humans and chimpanzees were obtained using previously published syntenic data between human (T2T-CHM13v1.0) and chimpanzee (PanTro6) references [35] intersected with SD98 genes using BEDTools intersect. For each paralog, family status was designated as "Human-duplicated gene family" if it was assigned to a gene family containing at least one expanded or duplicated member according to famCN and/or at least one non-syntenic member based on human/chimpanzee synteny. Otherwise, family status was considered "Undetermined".

Paralog-specific copy number genotyping

parCN estimates were obtained using QuicK-mer2 [46] for 1KGP 30× high-coverage Illumina individuals [55] and four archaic genomes (including Altai Neanderthal [PRJEB1265] [47], Vindija Neanderthal [PRJEB21157] [48], Mezmaiskaya Neanderthal [PRJEB1757] [47,48], and Denisova [PRJEB3092] [49]), using T2T-CHM13 (v1.0) as reference [34]. The resulting BED files containing parCN estimates were converted into bed9 format using a custom Python script for visualization in the UCSC Genome Browser. parCN values were genotyped across SD98 regions overlapping protein-encoding and unprocessed pseudogenes by calculating the mean parCN across the region of interest for each sample using a custom Python script. parCN dotplots generated using the R package ggplot2 are available for SD98 genes as an interactive Shiny web application in https://dcsoto.shinyapps.io/shinycn.

Metrics of selective constraint

Loss-of-function intolerance of SD98 genes was assayed using previously published gnomAD (v2.1.1) probability of loss-of-function intolerance scores (pLI) [158] and loss-of-function observed/expected upper fraction (LOEUF) [52]. We considered genes as intolerant to loss of function if either their pLI scores were greater than 0.9 or their LOEUF scores were less than 0.35.

Genome-wide Tajima's D analysis

Additionally, Tajima's D [54] values were calculated using previously published SNPs obtained from high-coverage short-read sequencing data from unrelated 1KGP individuals (n=2,504) [55], remapped to T2T-CHM13 (v1.0) [36]. Windows were defined as SD98 if at least 10% of the bases corresponded to SD98 regions. To define short-read accessible windows, 25-kbp windows were intersected with a published short-read combined accessibility mask [36]. Considering that no SD98 windows were fully accessible (Figure S4), Tajima's D was calculated for each superpopulation using VCFtools [159] across 25-kbp windows of at least 50% accessibility and with five or more SNPs. Because previous studies have highlighted potential discrepancies of evolutionary constraints experienced between duplicated and non-duplicated genomic loci [160], outlier D values were calculated for each continental superpopulation as the 5th and 95th percentiles within SD98 windows only, thereby avoiding comparisons between duplicated and unique regions. Outlier threshold values for each population were defined as follows: AFR, -2.21 and -0.67; EUR, -2.37 and 0.08; EAS, -2.48 and -0.10; SAS, -2.40 and -0.28; and AMR, -2.40 and -0.41.

Association with neural traits

Gene-disease associations were obtained from the GWAS catalog v1.0 [161]. SNPs significantly associated with brain measurements (*p*-value < 0.05) were selected, and the GWAS "mapped genes" were intersected with the SD98 gene list using gene symbols. Similarly, previously published associations between CNVs and neural traits in the UKBB were obtained [56]. Coordinates of CNVs significantly associated with brain measurements (*p*-value<0.05) were lifted over from hg19 to hg38 and from hg38 to T2T-CHM13 (v1.0) using UCSC liftOver tool [162]. Liftover chains were obtained from the UCSC Genome Browser and T2T-CHM13 GitHub page (https://github.com/marbl/CHM13, previous assembly releases of T2T-CHM13), respectively. CNVs were intersected with SD98 gene coordinates using BEDTools intersect[157].

ParCN values from SD98 genes for families with autistic children from the SSC (n = 2,459 families, n = 9,068 individuals) mapped to the T2T-CH13v1.1 reference genome were obtained, following the same steps as described to genotype parCN across 1KGP individuals. Overall, CN differences between autistic probands and unaffected siblings were compared by rounding median CN per individual to the nearest integer, and significance was assessed using the Wilcoxon signed-rank test, correcting for multiple testing with the false discovery rate method. To identify *de novo* deletions or duplications in autistic probands and unaffected siblings, parCN values within ±0.2 of an integer were conservatively selected and rounded to the nearest integer for all family members. Intermediate values, which could potentially confound the analysis, were removed. *De novo* events were classified as cases where both parents exhibited a parCN=2, while the child showed a parCN=3 (duplication) or parCN=1 (deletion).

Previously published genomic hotspots [60] were obtained in hg19 coordinates and lifted over to hg38 and from hg38 to T2T-CHM13 (v1.0) using the UCSC liftOver tool and associated chain files (described above). Three regions failed the liftover process due to differences in reference genome sequences. An extra 500 kbp were added upstream and downstream of each reported genomic hotspot to account for breakpoint errors. SD98 genes, including those exhibiting putative *de novo* events in the SSC dataset, were intersected with expanded genomic hotspots coordinates using BEDTools intersect.

Gene expression and network analysis

Previously published brain transcriptomic datasets, including post-mortem tissue and cell lines, were obtained. These datasets included neocortical germinal zones [61], neural stem and progenitor cells [21], a neuroblastoma cell line SHSY5Y [62], and two longitudinal studies of *in vitro* induced neurogenesis from human embryonic stem cells [63] (CORTECON), and post-mortem brain (BrainSpan) [64]—the latter of which was separated into prenatal and postnatal samples. Raw reads were pseudo-mapped to T2T-CHM13 (v2.0) CAT/Liftoff transcriptome and the CHM13v2.0 assembly as decoy sequence using Salmon v1.8.0 [163] with the flags "--validateMappings --gcBias". Transcripts per million (TPM) values and raw counts were summed to the gene level using tximport [164]. An SD98 gene was considered expressed during development if TPM values were greater than one in at least one of these samples, excluding postnatal BrainSpan data. Conversely, an SD98 gene was considered expressed postnatally if TPM values were greater than one in at least one postnatal stage of BrainSpan.

Gene co-expression analysis was performed using the WGCNA R package [65]. Briefly, samples were analyzed using principal components and hierarchical clustering to assess outliers, removing two samples (SRR1238515 and SRR1238516). Features with consistently low counts across remaining samples (counts <10 in 90% samples) were removed from this analysis. Raw counts for each sample were normalized using variance stabilizing transformation before performing a signed network construction with function blockwiseModules, with parameters soft power = 24, deepSplit = 4, detectCutHeight = 0.995, minModuleSize = 30, and MergeCutHeight = 0.15. GO terms enrichment analysis was performed using the R package clusterProfiler ego function [165]. Enrichment of gene categories were performed using the hypergeometric test in R for autism genes [57], expanded genomic hotspots [60], and cell markers [68].

Visualization of the yellow network was constructed by selecting genes with module membership greater than 0.5, generating an adjacency matrix with remaining genes, and then reconstructing a signed network with soft threshold = 18. Edges with Pearson correlation <0.1 were removed. The network visualization was built with the igraph R package (https://r.igraph.org/), using layout_with_fr for vertex placement. Vertex size was proportional to the degree and edges width was proportional to the Pearson's correlation coefficient. Some vertices were manually adjusted to improve aesthetics of the plot.

Mouse and zebrafish orthologs

Mouse-human orthologs were obtained from the Mouse Genome Informatics (MGI) complete list of human and mouse homologs and ENSEMBL BioMart, intersected with SD98 genes using gene symbols, and manual curation. Zebrafish-human orthologs were obtained from combined ENSEMBL BioMart annotations, MGI complete list of vertebrate homology classes, and manual curation. MGI files were downloaded from their website (https://www.informatics.jax.org/homology.shtml) and BioMart analyses were performed using the R package biomaRt. Comparison of developmental brain expression of SD98 orthologs in model organisms was performed using previously published expression data for mouse (PRJNA637987) [70] and zebrafish (GSE158142) [71], calculating Z-score normalized TPM values. Matching of developmental stages across human, mouse, and zebrafish was done as previously described [72]. In brief, genes with one-to-one orthologs with human genes were identified (mouse n= 19,949; zebrafish n= 16,910) and the principle component analysis rotations of the human BrainSpan data used to predict PC coordinates for the mouse and zebrafish data in human principle component space.

Capture HiFi (cHiFi) sequencing

We performed cHiFi sequencing of 172 individuals from the 1KGP, two trios from Genome in a Bottle [166], and 22 HGDP individuals with available linked-read data via the 10X Genomics platform [167], totaling 200 samples and 18 family trios (Table S13). DNA samples for 1KGP and Genome in a Bottle were obtained from the Coriell Institute (Camden, NJ, USA) and HGDP samples were obtained from the CEPH Biobank at the Fondation Jean Dausset-CEPH (Paris, France). PacBio cHiFi sequencing was performed using the RenSeq protocol [168]. Briefly, genomic DNA (~4 µg) was sheared to approximately 3 kbp with the Covaris E220 sonicator using Covaris blue miniTUBEs, followed by purification and size selection with AMPure XP beads. End repair and adapter ligation were performed using the NEBNext Ultra DNA Library Prep Kit. Barcodes to distinguish each sample were added via PCR using Kapa HiFi Polymerase (Roche, CA, USA). After the first PCR (fewer than 9 cycles), the libraries were purified and size-selected. For target enrichment, 80-mer RNA baits were designed and tiled at 2× coverage across targeted SD regions and unique exonic regions (Table S28). pHSD regions of interest were targeted and enriched for using a custom myBaits kit (Arbor Biosciences, MI, USA) following manufacturer's recommended protocol. Eight pooled barcoded libraries were hybridized overnight to the baits, and the captured DNA was bound to Dynabeads MyOne Streptavidin C1 beads. A second PCR was performed post-hybridization to generate sufficient material for sequencing. A PCR cycle test was conducted prior to the second amplification to limit PCR duplication bias.

The final libraries were size-selected using the Blue Pippin system to enrich for fragments >2 kbp and sequenced on the PacBio Sequel II platform (Maryland Genomics, University of Maryland). Briefly, Sequel II libraries were constructed using SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA) according to manufacturer's instructions. In brief, DNA samples were treated with DNA-damage repair enzymes followed by end-repair enzymes before being ligated to overhang sequencing adaptors. Libraries were then purified with SPRI beads (Beckman Coulter, Indianapolis, IN) and quantified on the Femto Pulse instrument (Agilent Technologies, Santa Clara, CA). Prior to sequencing, libraries were bound to Sequel II polymerase, then sequenced with Sequel II Sequencing kit and SMRT cell 8M on the Sequel II instrument (Pacific Biosciences, Menlo Park, CA). The approach yielded ~3-kbp reads with an average coverage of 27× across regions of interest, considering reads with MAPQ greater than 10 (Table S29).

Long-read genetic variation discovery and analysis

cHiFi reads were processed using the standard PacBio SMRT sequencing software tools available in the Conda repository pbbioconda. Circular consensus was obtained from subreads using CCS command with the following parameters --minPasses 3 and --minPredictedAccuracy 0.9. PacBio adapters and sample barcodes were removed using lima software and duplicates were removed with pbmarkdup. Resulting cHiFi reads were aligned to T2T-CHM13v1.0 reference using pbmm2 align, a wrapper of minimap2, with the CCS preset and default parameters. Read groups were added with Picard AddOrReplaceReadGroups and variants were called on each sample using GATK HaplotypeCaller [169], using ploidy = 2 and minimum mapping quality thresholds for genotyping of 0, 2, 5, 10 and 20, resulting in gVCF files per sample for joint genotyping. Joint genotyping was performed with GATK CombineGVCFs and GenotypeGVCFs tools using the pedigree file for accurate calculation of inbreeding coefficients. Genotyping was performed using minimum confidence threshold of 0, 10, 20 and 30. As the technical profile of variants in SDs differs from Variant Quality Score Recalibration training sets, a hard-filtering approach was utilized, including genotyping quality threshold of 0, 20, 50, 70, and depth of 0, 4, 8, 12, and 16. Based on a comprehensive benchmark, including comparison of cHiFi with HPRC/HGSCV and trio mendelian concordance (Note S3), the following minimum thresholds were selected: mapping quality of 20, confidence of 30, genotype quality of 20, and depth of 8. Only unrelated samples from the 1KGP were selected for downstream population analyses (n=144).

Fully phased haplotypes from 47 individuals from the HPRC Year 1 freeze (https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0) and 15 from the HGSVC [79] were downloaded. Each haplotype was mapped to T2T-CHM13v1.0 reference genome using minimap2 with parameters -a --eqx --cs -x asm5 --secondary=no -s 25000 -K 8G, and unmapped contigs and non-primary alignments were discarded. For each region of interest, the longest alignment spanning the locus was selected and additional alignments were removed. This process ensured that one single contiguous contig was used for variant detection. Variants were called with htsbox pileup with parameters -q 0 -evcf and converted into diploid calls using dipcall-aux.js vcfpair. For each region of interest, individual sample calls were merged into a multi-sample VCF file using BCFtools merge, only including individuals whose two haplotypes fully spanned the region of interest. Redundant samples between the HPRC and HGSVC (HG00733, HG02818, HG03486, NA19240, NA24385) were removed, prioritizing HPRC assemblies. Finally, the HPRC/HGSVC dataset was merged with cHiFi variants from 144 unrelated samples into a combined dataset for downstream analyses using BCFtools merge. Functional consequences of the combined dataset were assessed using the ENSEMBL Variant Effect Prediction (VEP) tool.

Haplotype networks for *CD8B* were constructed using HPRC/HGSVC continuous haplotypes extracted with BEDtools getfasta and aligned with Muscle using Mega Software [170]. Networks were generated using a minimum spanning tree with the software PopArt [171].

Tests for signatures of natural selection

Ka/Ks ratios (also known as dN/dS) were calculated for pHSD paralogs, performing pairwise comparison between human and chimpanzee sequences, based on T2T-CHM13v1.0 and panTro6 reference genomes, respectively. Alignments between human and chimpanzee canonical transcripts sequences were manually curated and used as input for seqinr package for Ka/Ks estimation. pN/pS ratios were calculated using as input variant sites estimated by seqinr package as well as polymorphic variation from the combined cHiFi and HPRC/HGSCV dataset, considering only biallelic SNPs from unrelated samples (n=144). Synonymous and nonsynonymous mutations were defined based on previously calculated VEP consequences. Ka/Ks and pN/pS values were jointly analyzed using the Direction of Selection (DoS) statistic, a derivation of McDonald–Kreitman's neutrality index, defined as DoS = Dn/(Dn + Ds) - Pn/(Pn + Ps) [83]. Significant differences in Ka/Ks or DoS between ancestral and derived paralogs were assessed using Wilcoxon signed-rank test, pairing each derived paralog to its ancestral counterpart. dN/dS was determined, in parallel, across gene families using codeml as part of the Phylogenetic Analysis by Maximum Likelihood (PAML [84]) from generated multiple-species alignments for each gene family (MAFFT [172]), using T2T-CHM13 for human paralog sequences and orthologous sequences from respective genomes for chimpanzee (panTro6), gorilla (gorGor6), orangutan (ponAbe3), rhesus (rheMac10), mouse (mm39), and rat (rn7). Ancestral and derived states for pHSD genes were assigned based on previously published predicted states [13]. Conservatively, the evolutionary status of four gene families was considered as "unknown" and excluded from calculations of statistical differences

(*FRMPD2*/*FRMPD2B*, *PTPN20*/*PTPN20CP*, *GPRIN2*/*GPRIN2B*, and *NPY4R*/*NPY4R2*). Paralogs with infinite values were also excluded from the analysis.

Nucleotide diversity (π) and Tajima's D statistics were calculated across selected pHSD loci using biallelic SNPs derived from continuous haplotypes from HPRC and HGSVC assemblies, utilizing the PopGenome R package and its functions F_ST.stats and neutrality.stats, respectively. For the gene bodies of *GPR89*, *ROCK1*, *FAM72*, and *CD8B*, π and Tajima's D values were calculated using 15-kbp windows with 1-kbp steps. For GPR89 paralogs, π was calculated across extended surrounding duplicated regions using 20-kbp windows and 1-kbp steps. For *CD8B* paralogs, Tajima's D was calculated in surrounding regions using 6-kbp windows and 500-bp steps.

Generation of zebrafish lines

Wild-type NHGRI-1[173] and Tg[HuC-GFP] [108] adult zebrafish were kept in a temperature (28±0.5°C) and light (10h dark / 14h light) controlled environment following standard protocols [174] with flowing water filtered via UV (Aquaneering, San Diego, CA). As described previously [97,100], feeding included brine shrimp (Artemia Brine Shrimp 90% hatch, Aquaneering, San Diego, CA) and flakes (Select Diet, Aquaneering, San Diego, CA). To obtain embryos for the different assays, males and females were placed in a 1L breeding tank in a 1:1 or 1:2 ratio and eggs from at least five crosses collected and kept in Petri dishes with E3 media (0.03% Instant Ocean salt in deionized water) in an incubator at 28°C until used. All animal use was approved by the Institutional Animal Care and Use Committee from the Office of Animal Welfare Assurance, University of California, Davis.

Creation of CRISPR lines to knockout genes of interest was done as previously described [97,100,175]. Briefly, crRNAs were annealed with tracrRNA (Alt-R system, Integrated DNA Technologies, Newark, NJ) in a 100 µM final concentration to make the sgRNA duplex, which was then coupled with SpCas9 (20 µM, New England BioLabs, Ipswich, MA) to prepare injection mixes. All oligonucleotide sequences can be found in Table S30. Microinjection of one-cell stage zebrafish embryos was performed using an air injector (Pneumatic MPP1-2 Pressure Injector) to release ~1 nl of injection mix into each embryo. Injection mixes to knockout-specific genes included ribonucleoproteins with four different sgRNAs targeting early exons in equimolar concentrations. In parallel, stable CRISPR knockout lines were made using a single sgRNA (Table S30) and adults carrying candidate knockout alleles for each gene of interest were further outcrossed to remove potential off-target edits and then incrossed to generate a batch of wild-type, heterozygous, and homozygous larvae, following standard protocols [176]. Knockout alleles in stable lines corresponded to a 5 bp deletion in *frmpd2* and an 8 bp deletion in *gpr89* (allele sequences can be found in Figure S21). For *arhgap11* knockdown, morpholinos blocking translation (GeneTools, Philomath, OR) were reconstituted to 2 mM and ~1 nl of a 2 ng/nl mix was microinjected into one-cel-stage embryos. Assessments of potential off-target sites for all sgRNAs used in this study were performed with the CIRCLE-seq protocol [177,178] and top potential off-target sites were evaluated via Sanger sequencing as previously described [97]. No editing was observed in potential off-target sites for any sgRNA used in this study, suggesting that phenotypes observed are due to the targeted knockout.

"Humanized" zebrafish larvae were generated by temporal expression of transcribed mRNAs. Expression vectors containing human transcripts were used to generate mRNA, including pEF-DEST51 (*SRGAP2C* and *ARHGAP11B*), pGCS1 (*GPR89B*, *PDZK1P1*, and *PTPN20CP*), pCR4 (*NPY4R*), and pCMV-

SPORT6 (*FAM72B* and *FRMPD2B*). The cDNA inserts of two genes were synthesized (Twist Biosciences, San Francisco, CA) based on transcript evidence from IsoSeq data from the ENCODE [179] project (*PDZK1P1*: ENCFF158KCA, ENCFF939EUU; *PTPN20CP*: ENCFF305AFY). All plasmids were sequenced through either Azenta or Plasmidsaurus and are included as Data S1. Following plasmid linearization using restriction enzyme digest and DNA purification, 5'-capped *in vitro* mRNA was generated using the MEGAshortscript transcription kit (Thermo Fisher, Waltham, MA) following the manufacturer's protocol with a 3.5 h 56°C incubation with T7 or SP6 RNA polymerase, depending on the plasmid. The resulting transcripts were purified with the MEGAclear transcription clean-up kit (Thermo Fisher, Waltham, MA), measured quantity with the Qubit, and visualized on a 2% agarose gel to ensure intact transcript. All mRNA injection mixes included mRNA at a 100 ng/ul concentration and ~1 nl of the mix microinjected into one-cell stage embryos, as described above.

Morphometric assessments

High-throughput imaging of the zebrafish larvae was performed using the VAST BioImager system (Union Biometrica, Holliston, MA) as previously described [96,97]. Mutant and control larvae at 3 or 5 dpf were placed into 96-well plates where they were then acquired by a robotic arm, placing the larvae in a rotating 600 µm capillary coupled with a camera, allowing for the automatic acquisition of images from four sides. Images were then processed and analyzed using the TableCreator tool in FishInspector v1.7 [95] to measure the head area and body length of 3,146 larvae—discarding images with general issues (e.g., dead or truncated larvae). To validate changes in head area, a neuronal reporter transgenic zebrafish line Tg[HuC-GFP][108] was used to create CRISPR-knockouts or humanized larvae that were then kept in an incubator at 28°C until imaged at 3 dpf using tricaine as anesthesia (0.0125%) and low-melting agarose. Imaging was performed in the Dragonfly spinning disk confocal microscope system with an iXon camera (Andor Technology, Belfast, United Kingdom). Z-stacks of 10 µm slices for each larva were collected and processed using Fiji [180] to generate hyperstacks with maximum intensity projections. Forebrain areas were measured in a blinded manner by a different trained investigator by manually delimiting the forebrain region. Any image with tilted larvae or unclear definition of the different brain regions were not included.

Supervised classification of mutants and controls

As an alternative to performing statistical tests to identify changes in predefined morphological measurements between mutants and controls, we employed a CNN to broadly identify differences between mutants and controls without the need to measure predefined features. Due to the use of multiple 96-well plates for each mutant, we observed significant batch effects in the resulting images, where larvae images from the same plate were significantly more similar to each other than to genotypically matched larvae from different plates. Therefore, before training our CNN-based classifier, we trained a latent diffusion model (LDM) to minimize the plate batch effect before input into the CNN. The broad goal of the LDM is to use the larvae with control genotypes present on each plate to learn the plate-specific batch effects, or 'style'. We then select a single plate as a reference and use the LDM to transform all images to the reference plate style, therefore making them comparable. The LDM removes batch effects by computing four transformations of the original image x. First, the original image x is transformed into a latent representation z_0 through the use of an variational autoencoder function $\mathcal{E}$ that summarizes the input image x but does not remove any batch effect:

$$z_0 = \mathcal{E}(x)$$

We then pass the encoded image z_0 through an LDM forward process, in which Gaussian noise is gradually added to the latent representation over T time steps z_t, calculated as:

$$z_t = \sqrt{\bar{a}_t} z_0 + \sqrt{1 - \bar{a}_t} \varepsilon_t \ \varepsilon, \ldots, \varepsilon_{t-2}, \varepsilon_{t-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$$

Where $\bar{a}_t$ is the cumulative product of the noise scheduler, and ε_t is the Gaussian noise sampled from a standard normal distribution $N(0,I)$ at time step t. This ultimately transforms the initial image embedding z_0 into the embedding z_T. This embedding z_T represents the larvae as an embedded image, free of association with any batch effect.

In the third step, we apply a reverse process of the LDM, we successively transform the image z_T into a new z_0, but 'add back in' the effect of a reference plate batch by introducing a condition variable c which comprises the desired batch and mutant ID. This conditional reverse process can be expressed as:

$$p_\theta(z_t - 1 | z_t, c) = \mathcal{N}(z_t - 1; \mu_\theta(z_t, t, c), \Sigma_\theta(z_t, t, c))$$

Where $\mu_\theta(z_t,t,c)$ represents the predicted mean of our denoised latent image through the weighted autoencoder with weights θ. Likewise, on top of predicting the mean noise we can predict the variance presented by $\Sigma_\theta(z_t,t,c)$. Finally, we pass our model through the decoder of a variational autoencoder to reconstruct the original image x into a new image, x', that represents the original image x but in the new reference plate style, suitable for input into the CNN classifier.

Our LDM is trained by minimizing the mean squared error between the true ε and predict noise $\varepsilon_\theta(z_t,t)$ expressed as:

$$L_{LDM} := E_{\varepsilon, t \sim N(0,1), t} \left[|| \varepsilon - \varepsilon_\theta(z_t, t) ||^2 \right]$$

This ensures that the model is accurately predicting the noise applied during the forward process, and by conditioning on the batch and mutant ID, we ensure we can reconstruct into our desired batch with respect to the original mutant.

Furthermore, we trained the model using 350 diffusion steps with a linear noise scheduler [181,182]. After training the model, we applied the model to transform all images to one reference plate, which is selected as the one with the highest number of controls. This transformation process minimizes the batch effect by generating images that appear as if they were collected from the same plate.

Having minimized batch effects on the larvae images, we then trained a CNN image classifier to determine the extent to which each mutant genotype differs from matched controls on the basis of the raw morphometric images alone. Higher classification accuracy, as measured by F1 score, indicates a larger effect size of mutant genotype on morphology. Our CNN framework involved fine-tuning a pretrained Alexnet classifier on the transformed larvae images [183]. More specifically, we trained 17 different Alexnet classifiers, one per mutant genotype, to perform binary classification to distinguish one specific mutant genotype from controls. The models were trained and evaluated in an x-fold cross validation framework, with F1 scores averaged over all folds. To generate feature attribution heat maps highlighting the morphological regions used to distinguish each mutant genotype from controls (Figure 4B), we used the GradCAM (Gradient-weighted Class Activation Mapping) approach [165]. We selected a *GPR89B* and *gpr89KO* sample representative of the pattern exhibited across mutants from this family.

sciRNA-seq

We performed cellular assessments using the single-cell combinatorial indexing RNA sequencing (sciRNA-seq) protocol [102]. Zebrafish larvae from CRISPR knockout or mRNA-injected lines were generated as described above and kept in an incubator at 28°C until 3 dpf when they were euthanized in cold tricaine (0.025%) and their heads immediately dissected, pooling 15 heads together per sample. Dissociation of the dissected heads was performed following two washes in 1 ml of cold 1x PBS on ice with a 15 min incubation in dissociation mix (480 µl of 0.25% trypsin-EDTA and 20 µl of collagenase P at 100 mg/ml), gently pipetting each sample every 5 min with a cut-open P1000 tip for complete dissociation. Once all tissue was visibly dissociated, 800 µl of DMEM with 10% FBS was added to each sample and centrifuged for 5 min at 700g at 4°C, resuspended in cold 1x PBS and centrifuged again at 700g for 5 min at 4°C. Cells were then resuspended in 800 µl of DMEM with 10% FBS and filtered through a 40 µm cell strainer (Flowmi, Sigma Aldrich, St. Louis, MO) using low-bind DNA tubes (Eppendorf, Hamburg, Germany). Cells were counted using a Countess II (Thermo Fisher, Waltham, MA) and all samples with viability >65% used further. Immediately after viability confirmation, cells were fixed as previously described [184] with a 10 min incubation in 1.33% formaldehyde in 1x PBS on ice followed by permeabilization with 5% Triton-X for 3 min on ice, and neutralization with 10% Tris-HCl (1M, pH 8). Cells were then filtered through a 40 µm cell strainer again, 15 µl of DMSO added to each sample, and then slowly freezed in a Mr Frosty (Thermo Fisher, Waltham, MA) freezing container filled with isopropanol at -80°C overnight.

Library preparation was performed following the sciRNA-seq protocol as described [102], including three rounds of combinatorial indexing of the cells (all primer sequences correspond to Plate 1 of the original protocol and can be found in www.github.com/JunyueC/sci-RNA-seq3_pipeline). The first round involved reverse transcription with barcoded oligo-dT primers to introduce the initial index. Cells were then pooled and redistributed into new wells for the second round, where a second index was added via ligation. The third round included second-strand synthesis, tagmentation with Tn5 transposase, and PCR amplification to incorporate the final index. Libraries were evaluated for quality control in a BioAnalyzer and Qubit to check integrity and concentration, and then sequenced in three NovaSeq 6000 lanes (Novogene, Sacramento, CA). Raw fastq files were processed following the available sci-RNA-seq3 pipeline [185] (www.github.com/JunyueC/sci-RNA-seq3_pipeline). This pipeline includes attachment of the unique molecular identifier (UMI) sequence to each read2 based on the identified RT and ligation barcodes from read1 (edit distance ≤1), and trimming with TrimGalore v0.4.1 (https://zenodo.org/records/7598955), using cutadapt [186] and fastqc [187]. Reads were then mapped to the improved zebrafish transcriptome [188] with STAR [189] using the --outSAMstrandField intronMotif option. Duplicates (reads with the same UMI) were removed with the available custom-made python scripts found in the Cao lab GitHub repository. Lastly, filtered SAM files were splitted by their UMI sequences (corresponding to individual cells) and gene-cell count matrices constructed by mapping reads to the zebrafish v4 GTF file [188].

Gene-cell count matrices were loaded into R to generate Seurat v4 [190] objects and cells with transcript counts below 150 or above two standard deviations over the mean, mitochondrial or ribosomal gene counts >5%, or potential doublets (with a ~4% doublet expectation based on previous reports [185,191] and estimated using DoubletFinder [192]) were removed (Figure S22). Cells from different libraries were normalized using SCTransform [193] with the glmGamPoi method and regressing by the percentage of mitochondrial and ribosomal counts. Then, normalized counts across sequencing libraries were integrated with Harmony [194] with a PCA reduction using batch as a grouping variable. Hierarchical clustering was performed by calculating the euclidean distances across all cells using the Harmony cell embeddings and clustering with the hclust function using the ward.D2 method. The hierarchical tree was cut at a K of 50, gene markers for each cluster estimated using the FindAllMarkers function (logfc.threshold=0.10, test.use="MAST", min.pct=0.15, min.diff.pct=0.10), and classification into cell types using available zebrafish brain scRNA-seq atlases [71,105] and the Zebrafish Information Network (ZFIN [195]) website. Focusing on neuronal, glial, and eye-related clusters left a total of 95,555 cells for further analysis (Tables S31 and S32). General correlations across samples (knockout vs. "humanized" models for each gene of interest) were done with a balanced number of cells for each pair and pseudo-bulking gene counts by sample and cluster, so counts across cells were summed together for each sample, allowing for biological replicates to be maintained. Then, pseudo-counts were processed with DESeq2 [196] with the Wald test option to obtain fold-change values for each gene compared to their respective control (SpCas9-scrambled gRNA injected for crispants, GFP-mRNA-injected for "humanized", and control-morpholino-injected for *arhgap11*-knockdown). Then, cell-type-specific differential gene expression tests were performed similarly but with previous subsetting of the matrix for each cell type. For *FRMPD2* and *GPR89* models, forebrain cells were further re-clustered to obtain more detailed cell types; gene expression across samples correlated as described above using a pseudo-bulk approach with the telencephalic cells. Progenitor and differentiated cell classification was performed using known neural progenitor (*sox19a*, *sox2*, *rpl5a*, *npm1a*, *s100b*, *dla*) or mature neuron (*elavl3*, *elavl4*, *tubb5*) markers and the PercentageFeatureSet function to estimate the weight of these genes per cell. Enrichment of DEGs in gene ontology terms was estimated with clusterProfiler [165] using only the expressed genes as the background list for the tests.

Seizure susceptibility

To assess changes to chemically induced seizure susceptibility, we employed an optimized published protocol [115]. Briefly, larvae were collected and kept in an incubator at 28°C until 4 dpf, when they were distributed in a 96-well plate and placed in a Zebrabox system chamber (ViewPoint, Montreal, Canada) that has a camera with an acquisition speed of 30 frames per second. Treatments included 0 or 2.5 mM of pentylenetetrazol (PTZ, #P6500, Sigma-Aldrich, St. Louis, MO) in a total volume of 200 µl per well. Once placed in the Zebrabox chamber, larvae were left for 10 min unbothered before starting a 15 min recording (acquisition in 1 s bins) to then extract the frequency of high-speed events (>28 mm/s) using a published MATLAB script [115] to compare against batch-sibling controls.

Data Availability

NCBI GenBank numbers of deposited data pending: cHiFi sequencing, scRNA-seq for zebrafish. Code and data: https://github.com/mydennislab/public_data/ (zenodo pending).

Acknowledgements

We would like to thank the 1KGP, HGDP, T2T, SSC, gnomAD, HPRC, and HGSVC for access to their data and biospecimens. Evolutionary analysis (Ka/Ks) of pHSDs was performed by Dr. Alfie Gleeson of University College London (UCL). We thank Tonia Brown for providing assistance with editing the manuscript. The Cell Biology and Human Anatomy Department at UC Davis School of Medicine provided significant support in using their imaging equipment. This work was supported, in part, by U.S.

National Institutes of Health (NIH) grants from the Office of the Director and National Institute of Mental Health (DP2MH119424 and R01MH132818 to M.Y.D., DP2MH129987 to G.Q.), National Institute of Neurological Disorders and Stroke grant (R01NS109176 to S.S.), National Institute of Child Health and Human Development (P50HD103526 to G.Q.). A.S. was supported through the UC Davis Molecular and Cellular Biology T32 (T32GM007377). N.K.H. is supported by an NIGMS UC Davis eMCDB T32 (T32GM153586). N.A.F.M. was supported by the NIH as a UC Davis Postbaccalaureate Research Education Program fellow (R25GM116690). Support for A.M.A. came from UCL Wellcome Trust ISSF3 award (204841/Z/16/Z) and Biotechnology and Biological Sciences Research Council, BBSRC (BB/WW007703/1). Some figures were made with Biorender.

References

1. Carroll, S.B. (2003). Genetics and the making of Homo sapiens. Nature *422*, 849–857.

2. Varki, A., and Altheide, T.K. (2005). Comparing the human and chimpanzee genomes: Searching for needles in a haystack. Preprint, https://doi.org/10.1101/gr.3737405 https://doi.org/10.1101/gr.3737405.

3. Pääbo, S. (2014). The Human Condition—A Molecular Approach. Preprint, https://doi.org/10.1016/j.cell.2013.12.036 https://doi.org/10.1016/j.cell.2013.12.036.

4. Pollen, A.A., Kilik, U., Lowe, C.B., and Camp, J.G. (2023). Human-specific genetics: new tools to explore the molecular and cellular basis of human evolution. Nat. Rev. Genet. *24*, 687–711.

5. Sousa, A.M.M., Meyer, K.A., Santpere, G., Gulden, F.O., and Sestan, N. (2017). Evolution of the Human Nervous System Function, Structure, and Development. Cell *170*, 226–247.

6. Enard, W., Gehre, S., Hammerschmidt, K., Holter, S.M., Blass, T., Somel, M., Bruckner, M.K., Schreiweis, C., Winter, C., Sohr, R., et al. (2009). A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. Cell *137*, 961–971.

7. Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S., Wiebe, V., Kitano, T., Monaco, A.P., and Paabo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. Nature *418*, 869–872.

8. Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

9. Soto, D.C., Uribe-Salazar, J.M., Shew, C.J., Sekar, A., McGinty, S.P., and Dennis, M.Y. (2023). Genomic structural variation: A complex but important driver of human evolution. Am J Biol Anthropol. https://doi.org/10.1002/ajpa.24713.

10. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., and Eichler, E.E. (2002). Recent segmental duplications in the human genome. Science *297*, 1003–1007.

11. Dennis, M.Y., and Eichler, E.E. (2016). Human adaptation and evolution by segmental duplication. Curr. Opin. Genet. Dev. *41*, 44–52.

12. Porubsky, D., and Eichler, E.E. (2024). A 25-year odyssey of genomic technology advances and

structural variant discovery. Cell *187*, 1024–1037.

13. Dennis, M.Y., Harshman, L., Nelson, B.J., Penn, O., Cantsilieris, S., Huddleston, J., Antonacci, F., Penewit, K., Denman, L., Raja, A., et al. (2017). The evolution and population diversity of human-specific segmental duplications. Nat Ecol Evol *1*, 69.

14. Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C., Malig, M., Tsalenko, A., Sampas, N., Bruhn, L., Shendure, J., 1000 Genomes Project, et al. (2010). Diversity of human copy number variation and multicopy genes. Science *330*, 641–646.

15. Sudmant, P.H., Huddleston, J., Catacchio, C.R., Malig, M., Hillier, L.W., Baker, C., Mohajeri, K., Kondova, I., Bontrop, R.E., Persengiev, S., et al. (2013). Evolution and diversity of copy number variation in the great ape lineage. Genome Res. *23*, 1373–1382.

16. Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.-E., Lambert, N., de Marchena, J., Jin, W.-L., Vanderhaeghen, P., Ghosh, A., Sassa, T., et al. (2012). Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell *149*, 923–935.

17. Dennis, M.Y., Nuttle, X., Sudmant, P.H., Antonacci, F., Graves, T.A., Nefedov, M., Rosenfeld, J.A., Sajjadian, S., Malig, M., Kotkiewicz, H., et al. (2012). Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. Cell *149*, 912–922.

18. Fiddes, I.T., Lodewijk, G.A., Mooring, M., Bosworth, C.M., Ewing, A.D., Mantalas, G.L., Novak, A.M., van den Bout, A., Bishara, A., Rosenkrantz, J.L., et al. (2018). Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell *173*, 1356–1369.e22.

19. Suzuki, I.K., Gacquer, D., Van Heurck, R., Kumar, D., Wojno, M., Bilheu, A., Herpoel, A., Lambert, N., Cheron, J., Polleux, F., et al. (2018). Human-Specific NOTCH2NL Genes Expand Cortical Neurogenesis through Delta/Notch Regulation. Cell *173*, 1370–1384.e16.

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

21. Florio, M., Albert, M., Taverna, E., Namba, T., Brandl, H., Lewitus, E., Haffner, C., Sykes, A., Wong, F.K., Peters, J., et al. (2015). Human-specific gene *ARHGAP11B* promotes basal progenitor amplification and neocortex expansion. Science *347*, 1465–1470.

22. Florio, M., Namba, T., Pääbo, S., Hiller, M., and Huttner, W.B. (2016). A single splice site mutation in human-specific ARHGAP11B causes basal progenitor amplification. Sci Adv *2*, e1601941.

23. Namba, T., Dóczi, J., Pinson, A., Xing, L., Kalebic, N., Wilsch-Bräuninger, M., Long, K.R., Vaid, S., Lauer, J., Bogdanova, A., et al. (2020). Human-Specific ARHGAP11B Acts in Mitochondria to Expand Neocortical Progenitors by Glutaminolysis. Neuron *105*, 867–881.e9.

24. Ju, X.-C., Hou, Q.-Q., Sheng, A.-L., Wu, K.-Y., Zhou, Y., Jin, Y., Wen, T., Yang, Z., Wang, X., and Luo, Z.-G. (2016). The hominoid-specific gene TBC1D3 promotes generation of basal neural progenitors and induces cortical folding in mice. Elife *5*. https://doi.org/10.7554/eLife.18197.

25. Van Heurck, R., Bonnefont, J., Wojno, M., Suzuki, I.K., Velez-Bravo, F.D., Erkol, E., Nguyen, D.T., Herpoel, A., Bilheu, A., Beckers, S., et al. (2023). CROCCP2 acts as a human-specific

modifier of cilia dynamics and mTOR signaling to promote expansion of cortical progenitors. Neuron *111*, 65–80.e6.

26. Libé-Philippot, B., Lejeune, A., Wierda, K., Louros, N., Erkol, E., Vlaeminck, I., Beckers, S., Gaspariunaite, V., Bilheu, A., Konstantoulea, K., et al. (2023). LRRC37B is a human modifier of voltage-gated sodium channels and axon excitability in cortical neurons. Cell *186*, 5766–5783.e25.

27. Karageorgiou, C., Gokcumen, O., and Dennis, M.Y. (2024). Deciphering the role of structural variation in human evolution: a functional perspective. Curr. Opin. Genet. Dev. *88*, 102240.

28. Taylor, D.J., Eizenga, J.M., Li, Q., Das, A., Jenike, K.M., Kenny, E.E., Miga, K.H., Monlong, J., McCoy, R.C., Paten, B., et al. (2024). Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. Annu. Rev. Genomics Hum. Genet. https://doi.org/10.1146/annurev-genom-021623-081639.

29. Ebbert, M.T.W., Jensen, T.D., Jansen-West, K., Sens, J.P., Reddy, J.S., Ridge, P.G., Kauwe, J.S.K., Belzil, V., Pregent, L., Carrasquillo, M.M., et al. (2019). Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol. *20*, 97.

30. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354.

31. Cabanski, C.R., Wilkerson, M.D., Soloway, M., Parker, J.S., Liu, J., Prins, J.F., Marron, J.S., Perou, C.M., and Hayes, D.N. (2013). BlackOPs: increasing confidence in variant detection through mappability filtering. Nucleic Acids Res. *41*, e178.

32. Derrien, T., Estellé, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigó, R., and Ribeca, P. (2012). Fast computation and applications of genome mappability. PLoS One *7*, e30377.

33. Lee, H., and Schatz, M.C. (2012). Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics *28*, 2097–2105.

34. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53.

35. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., et al. (2022). Segmental duplications and their variation in a complete human genome. Science *376*, eabj6965.

36. Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., et al. (2022). A complete reference genome improves analysis of human genetic variation. Science *376*, eabl3533.

37. Numanagic, I., Gökkaya, A.S., Zhang, L., Berger, B., Alkan, C., and Hach, F. (2018). Fast characterization of segmental duplications in genome assemblies. Bioinformatics *34*, i706–i714.

38. Dougherty, M.L., Underwood, J.G., Nelson, B.J., Tseng, E., Munson, K.M., Penn, O., Nowakowski, T.J., Pollen, A.A., and Eichler, E.E. (2018). Transcriptional fates of human-specific segmental duplications in brain. Genome Res. *28*, 1566–1576.

39. Larson, J.L., Silver, A.J., Chan, D., Borroto, C., Spurrier, B., and Silver, L.M. (2015). Validation of a high resolution NGS method for detecting spinal muscular atrophy carriers among phase 3

participants in the 1000 Genomes Project. BMC Med. Genet. *16*, 100.

40. Moreno-Igoa, M., Hernández-Charro, B., Bengoa-Alonso, A., Pérez-Juana-del-Casal, A., Romero-Ibarra, C., Nieva-Echebarria, B., and Ramos-Arroyo, M.A. (2015). KANSL1 gene disruption associated with the full clinical spectrum of 17q21.31 microdeletion syndrome. BMC Med. Genet. *16*, 68.

41. Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., et al. (2007). Diet and the evolution of human amylase gene copy number variation. Nat. Genet. *39*, 1256–1260.

42. Bolognini, D., Halgren, A., Lou, R.N., Raveane, A., Rocha, J.L., Guarracino, A., Soranzo, N., Chin, C.-S., Garrison, E., and Sudmant, P.H. (2024). Recurrent evolution and selection shape structural diversity at the amylase locus. Nature, 1–9.

43. Yilmaz, F., Karageorgiou, C., Kim, K., Pajic, P., Scheer, K., Human Genome Structural Variation Consortium, Beck, C.R., Torregrossa, A.-M., Lee, C., and Gokcumen, O. (2024). Paleolithic Gene Duplications Primed Adaptive Evolution of Human Amylase Locus Upon Agriculture. bioRxiv. https://doi.org/10.1101/2023.11.27.568916.

44. Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature *538*, 201–206.

45. Bosch, N., Cáceres, M., Cardone, M.F., Carreras, A., Ballana, E., Rocchi, M., Armengol, L., and Estivill, X. (2007). Characterization and evolution of the novel gene family FAM90A in primates originated by multiple duplication and rearrangement events. Hum. Mol. Genet. *16*, 2572–2582.

46. Shen, F., and Kidd, J.M. (2020). Rapid, Paralog-Sensitive CNV Analysis of 2457 Human Genomes Using QuicK-mer2. Genes *11*, 141.

47. Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P.H., de Filippo, C., et al. (2014). The complete genome sequence of a Neanderthal from the Altai Mountains. Nature *505*, 43–49.

48. Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., et al. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. Science *358*, 655–658.

49. Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prüfer, K., de Filippo, C., et al. (2012). A high-coverage genome sequence from an archaic Denisovan individual. Science *338*, 222–226.

50. Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41–45.

51. Nimmerjahn, F., and Ravetch, J.V. (2008). Fcγ receptors as regulators of immune responses. Nat. Rev. Immunol. *8*, 34–47.

52. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.

53. Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. *27*, 849–864.

54. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

55. Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K., et al. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell *185*, 3426–3440.e19.

56. Aguirre, M., Rivas, M.A., and Priest, J. (2019). Phenome-wide Burden of Copy-Number Variation in the UK Biobank. Am. J. Hum. Genet. *105*, 373–383.

57. Trost, B., Thiruvahindrapuram, B., Chan, A.J.S., Engchuan, W., Higginbotham, E.J., Howe, J.L., Loureiro, L.O., Reuter, M.S., Roshandel, D., Whitney, J., et al. (2022). Genomic architecture of autism from comprehensive whole-genome sequence annotation. Cell *185*, 4409–4427.e18.

58. Brunetti-Pierri, N., Berg, J.S., Scaglia, F., Belmont, J., Bacino, C.A., Sahoo, T., Lalani, S.R., Graham, B., Lee, B., Shinawi, M., et al. (2008). Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. Nat. Genet. *40*, 1466–1471.

59. Winchester, L., Yau, C., and Ragoussis, J. (2009). Comparing CNV detection methods for SNP arrays. Brief. Funct. Genomics *8*, 353–366.

60. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568–584.e23.

61. Fietz, S.A., Lachmann, R., Brandl, H., Kircher, M., Samusik, N., Schröder, R., Lakshmanaperumal, N., Henry, I., Vogt, J., Riehn, A., et al. (2012). Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal. Proc. Natl. Acad. Sci. U. S. A. *109*, 11836–11841.

62. Pezzini, F., Bettinetti, L., Di Leva, F., Bianchi, M., Zoratti, E., Carrozzo, R., Santorelli, F.M., Delledonne, M., Lalowski, M., and Simonati, A. (2017). Transcriptomic Profiling Discloses Molecular and Cellular Events Related to Neuronal Differentiation in SH-SY5Y Neuroblastoma Cells. Cell. Mol. Neurobiol. *37*, 665–682.

63. van de Leemput, J., Boles, N.C., Kiehl, T.R., Corneo, B., Lederman, P., Menon, V., Lee, C., Martinez, R.A., Levi, B.P., Thompson, C.L., et al. (2014). CORTECON: a temporal transcriptome analysis of in vitro human cerebral cortex development from human embryonic stem cells. Neuron *83*, 51–68.

64. Miller, J.A., Ding, S.-L., Sunkin, S.M., Smith, K.A., Ng, L., Szafer, A., Ebbert, A., Riley, Z.L., Royall, J.J., Aiona, K., et al. (2014). Transcriptional landscape of the prenatal human brain. Nature *508*, 199–206.

65. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

66. Shew, C.J., Carmona-Mora, P., Soto, D.C., Mastoras, M., Roberts, E., Rosas, J., Jagannathan, D., Kaya, G., O'Geen, H., and Dennis, M.Y. (2021). Diverse Molecular Mechanisms Contribute to Differential Expression of Human Duplicated Genes. Mol. Biol. Evol. *38*, 3060–3077.

67. Guerrier, S., Coutinho-Budd, J., Sassa, T., Gresset, A., Jordan, N.V., Chen, K., Jin, W.-L., Frost, A., and Polleux, F. (2009). The F-BAR domain of srGAP2 induces membrane protrusions required for neuronal migration and morphogenesis. Cell *138*, 990–1004.

68. Parikshak, N.N., Luo, R., Zhang, A., Won, H., Lowe, J.K., Chandran, V., Horvath, S., and Geschwind, D.H. (2013). Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. Cell *155*, 1008–1021.

69. Koolen, D.A., Kramer, J.M., Neveling, K., Nillesen, W.M., Moore-Barton, H.L., Elmslie, F.V., Toutain, A., Amiel, J., Malan, V., Tsai, A.C., et al. (2012). Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. Nat. Genet. *44*, 639–641.

70. La Manno, G., Siletti, K., Furlan, A., Gyllborg, D., Vinsland, E., Mossi Albiach, A., Mattsson Langseth, C., Khven, I., Lederer, A.R., Dratva, L.M., et al. (2021). Molecular architecture of the developing mouse brain. Nature *596*, 92–96.

71. Raj, B., Farrell, J.A., Liu, J., El Kholtei, J., Carte, A.N., Navajas, A.J., Du, L.Y., McKenna, A., Relić, Đ., Leslie, J.M., et al. (2020). Emergence of Neuronal Diversity during Vertebrate Brain Development. Neuron *108*. https://doi.org/10.1016/j.neuron.2020.09.023.

72. Willsey, H.R., Exner, C.R.T., Xu, Y., Everitt, A., Sun, N., Wang, B., Dea, J., Schmunk, G., Zaltsman, Y., Teerikorpi, N., et al. (2021). Parallel in vivo analysis of large-effect autism genes implicates cortical neurogenesis and estrogen in risk and resilience. Neuron *109*, 1409.

73. Weinschutz Mendes, H., Neelakantan, U., Liu, Y., Fitzpatrick, S.E., Chen, T., Wu, W., Pruitt, A., Jin, D.S., Jamadagni, P., Carlson, M., et al. (2023). High-throughput functional analysis of autism genes in zebrafish identifies convergence in dopaminergic and neuroimmune pathways. Cell Rep. *42*, 112243.

74. Guitart, X., Porubsky, D., Yoo, D., Dougherty, M.L., Dishuck, P., Munson, K.M., Lewis, A.P., Hoekzema, K., Knuth, J., Chang, S., et al. (2024). Independent expansion, selection and hypervariability of the gene family in humans. Genome Res. https://doi.org/10.1101/gr.279299.124.

75. Jiang, Z., Tang, H., Ventura, M., Cardone, M.F., Marques-Bonet, T., She, X., Pevzner, P.A., and Eichler, E.E. (2007). Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. Nat. Genet. *39*, 1361–1368.

76. Jarvis, E.D., Formenti, G., Rhie, A., Guarracino, A., Yang, C., Wood, J., Tracey, A., Thibaud-Nissen, F., Vollger, M.R., Porubsky, D., et al. (2022). Automated assembly of high-quality diploid human reference genomes. bioRxiv, 2022.03.06.483034. https://doi.org/10.1101/2022.03.06.483034.

77. Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2022). A Draft Human Pangenome Reference. bioRxiv, 2022.07.09.499321. https://doi.org/10.1101/2022.07.09.499321.

78. Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H.A., Lucas, J.K., Phillippy, A.M., Popejoy, A.B., Asri, M., Carson, C., Chaisson, M.J.P., et al. (2022). The Human Pangenome Project: a global resource to map genomic diversity. Nature *604*, 437–446.

79. Ebler, J., Ebert, P., Clarke, W.E., Rausch, T., Audano, P.A., Houwaart, T., Mao, Y., Korbel, J.O., Eichler, E.E., Zody, M.C., et al. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat. Genet. *54*, 518–525.

80. Cavalli-Sforza, L.L. (2005). The Human Genome Diversity Project: past, present and future. Nat. Rev. Genet. *6*, 333–340.

81. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122.

82. Ellegren, H. (2005). Evolution: natural selection in the evolution of humans and chimps. Curr. Biol. *15*, R919–R922.

83. Stoletzki, N., and Eyre-Walker, A. (2011). Estimation of the neutrality index. Mol. Biol. Evol. *28*, 63–70.

84. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

85. Connolly, J.M., Hansen, T.H., Ingold, A.L., and Potter, T.A. (1990). Recognition by CD8 on cytotoxic T lymphocytes is ablated by several substitutions in the class I alpha 3 domain: CD8 and the T-cell receptor recognize the same class I molecule. Proc. Natl. Acad. Sci. U. S. A. *87*, 2137–2141.

86. Salter, R.D., Benjamin, R.J., Wesley, P.K., Buxton, S.E., Garrett, T.P., Clayberger, C., Krensky, A.M., Norment, A.M., Littman, D.R., and Parham, P. (1990). A binding site for the T-cell co-receptor CD8 on the alpha 3 domain of HLA-A2. Nature *345*, 41–46.

87. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F., and Guigó, R. (2021). Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. Nat. Commun. *12*, 727.

88. Orrù, V., Steri, M., Sidore, C., Marongiu, M., Serra, V., Olla, S., Sole, G., Lai, S., Dei, M., Mulas, A., et al. (2020). Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. Nat. Genet. *52*, 1036–1045.

89. Holtzman, N.G., Iovine, M.K., Liang, J.O., and Morris, J. (2016). Learning to Fish with Genetics: A Primer on the Vertebrate Model Danio rerio. Genetics *203*, 1069–1089.

90. Meyers, J.R. (2018). Zebrafish: Development of a Vertebrate Model Organism: Zebrafish : Development of a Vertebrate Model Organism. Current Protocols Essential Laboratory Techniques *16*, e19.

91. Sakai, C., Ijaz, S., and Hoffman, E.J. (2018). Zebrafish Models of Neurodevelopmental Disorders: Past, Present, and Future. Front. Mol. Neurosci. *11*, 294.

92. Schmidt, E.R.E., Kupferman, J.V., Stackmann, M., and Polleux, F. (2019). The human-specific paralogs SRGAP2B and SRGAP2C differentially modulate SRGAP2A-dependent synaptic development. Sci. Rep. *9*, 18692.

93. Schmidt, E.R.E., Zhao, H.T., Park, J.M., Dipoppa, M., Monsalve-Mercado, M.M., Dahan, J.B., Rodgers, C.C., Lejeune, A., Hillman, E.M.C., Miller, K.D., et al. (2021). A human-specific modifier of cortical connectivity and circuit function. Nature *599*, 640.

94. Heide, M., Haffner, C., Murayama, A., Kurotaki, Y., Shinohara, H., Okano, H., Sasaki, E., and Huttner, W.B. (2020). Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. Science *369*. https://doi.org/10.1126/science.abb2401.

95. Kalebic, N., Gilardi, C., Albert, M., Namba, T., Long, K.R., Kostic, M., Langen, B., and Huttner, W.B. (2018). Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex. Elife *7*. https://doi.org/10.7554/eLife.41241.

96. Meng, X., Lin, Q., Zeng, X., Jiang, J., Li, M., Luo, X., Chen, K., Wu, H., Hu, Y., Liu, C., et al. (2023). Brain developmental and cortical connectivity changes in transgenic monkeys carrying the human-specific duplicated gene SRGAP2C. Natl Sci Rev *10*, nwad281.

97. Uribe-Salazar, J.M., Kaya, G., Sekar, A., Weyenberg, K., Ingamells, C., and Dennis, M.Y. (2022). Evaluation of CRISPR gene-editing tools in zebrafish. BMC Genomics *23*, 12.

98. Teixidó, E., Kießling, T.R., Krupp, E., Quevedo, C., Muriana, A., and Scholz, S. (2019). Automated Morphological Feature Assessment for Zebrafish Embryo Developmental Toxicity Screens. Toxicol. Sci. *167*. https://doi.org/10.1093/toxsci/kfy250.

99. Pulak, R. (2016). Tools for automating the imaging of zebrafish larvae. Methods *96*. https://doi.org/10.1016/j.ymeth.2015.11.021.

100. Colón-Rodríguez, A., Uribe-Salazar, J.M., Weyenberg, K.B., Sriram, A., Quezada, A., Kaya, G., Jao, E., Radke, B., Lein, P.J., and Dennis, M.Y. (2020). Assessment of Autism Zebrafish Mutant Models Using a High-Throughput Larval Phenotyping Platform. Front. Cell Dev. Biol. *8*, 586296.

101. Saunders, L.M., Srivatsan, S.R., Duran, M., Dorrity, M.W., Ewing, B., Linbo, T.H., Shendure, J., Raible, D.W., Moens, C.B., Kimelman, D., et al. (2023). Embryo-scale reverse genetics at single-cell resolution. Nature *623*, 782–791.

102. Martin, B.K., Qiu, C., Nichols, E., Phung, M., Green-Gladden, R., Srivatsan, S., Blecher-Gonen, R., Beliveau, B.J., Trapnell, C., Cao, J., et al. (2023). Optimized single-nucleus transcriptional profiling by combinatorial indexing. Nat. Protoc. *18*, 188–207.

103. Uribe-Salazar, J.M., Kaya, G., Weyenberg, K.B., Radke, B., Hino, K.K., Soto, D.C., Shiu, J.-L., Zhang, W., Ingamells, C., Haghani, N.K., et al. (2024). Zebrafish models of human-duplicated gene SRGAP2 reveal novel functions in microglia and visual system development. bioRxiv. https://doi.org/10.1101/2024.09.11.612570.

104. d'Amora, M., and Giordani, S. (2018). The Utility of Zebrafish as a Model for Screening Developmental Neurotoxicity. Front. Neurosci. *12*. https://doi.org/10.3389/fnins.2018.00976.

105. Zhang, H., Wang, H., Shen, X., Jia, X., Yu, S., Qiu, X., Wang, Y., Du, J., Yan, J., and He, J. (2021). The landscape of regulatory genes in brain-wide neuronal phenotypes of a vertebrate brain. Elife *10*. https://doi.org/10.7554/eLife.68224.

106. Kozol, R.A., Abrams, A.J., James, D.M., Buglo, E., Yan, Q., and Dallman, J.E. (2016). Function Over Form: Modeling Groups of Inherited Neurological Conditions in Zebrafish. Front. Mol. Neurosci. *9*, 55.

107. Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., and Schilling, T.F. (1995). Stages of embryonic development of the zebrafish. Dev. Dyn. *203*, 253–310.

108. Park, H.C., Kim, C.H., Bae, Y.K., Yeo, S.Y., Kim, S.H., Hong, S.K., Shin, J., Yoo, K.W., Hibi, M., Hirano, T., et al. (2000). Analysis of upstream elements in the HuC promoter leads to the establishment of transgenic zebrafish with fluorescent neurons. Dev. Biol. *227*, 279–293.

109. Porter, B.A., and Mueller, T. (2020). The Zebrafish Amygdaloid Complex - Functional Ground Plan, Molecular Delineation, and Everted Topology. Front. Neurosci. *14*, 608.

110. Anneser, L., Satou, C., Hotz, H.-R., and Friedrich, R.W. (2024). Molecular organization of neuronal cell types and neuromodulatory systems in the zebrafish telencephalon. Curr. Biol. *34*, 298–312.e4.

111. Deckstein, J., van Appeldorn, J., Tsangarides, M., Yiannakou, K., Müller, R., Stumpf, M., Sukumaran, S.K., Eichinger, L., Noegel, A.A., and Riyahi, T.Y. (2015). The Dictyostelium discoideum GPHR ortholog is an endoplasmic reticulum and Golgi protein with roles during development. Eukaryot. Cell *14*, 41–54.

112. Charroux, B., and Royet, J. (2014). Mutations in the Drosophila ortholog of the vertebrate Golgi pH regulator (GPHR) protein disturb endoplasmic reticulum and Golgi organization and affect systemic growth. Biol. Open *3*, 72–80.

113. Otani, T., Marchetto, M.C., Gage, F.H., Simons, B.D., and Livesey, F.J. (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*, 467–480.

114. Benito-Kwiecinski, S., Giandomenico, S.L., Sutcliffe, M., Riis, E.S., Freire-Pritchett, P., Kelava, I., Wunderlich, S., Martin, U., Wray, G.A., McDole, K., et al. (2021). An early cell shape transition drives evolutionary expansion of the human forebrain. Cell *184*, 2084–2102.e19.

115. Griffin, A., Carpenter, C., Liu, J., Paterno, R., Grone, B., Hamling, K., Moog, M., Dinday, M.T., Figueroa, F., Anvar, M., et al. (2021). Phenotypic analysis of catastrophic childhood epilepsy genes. Commun Biol *4*, 680.

116. Lu, X., Zhang, Q., and Wang, T. (2019). The second PDZ domain of scaffold protein Frmpd2 binds to GluN2A of NMDA receptors. Biochem. Biophys. Res. Commun. *516*, 63–67.

117. Stenzel, N., Fetzer, C.P., Heumann, R., and Erdmann, K.S. (2009). PDZ-domain-directed basolateral targeting of the peripheral membrane protein FRMPD2 in epithelial cells. J. Cell Sci. *122*, 3374–3384.

118. Ueno, A., Omori, Y., Sugita, Y., Watanabe, S., Chaya, T., Kozuka, T., Kon, T., Yoshida, S., Matsushita, K., Kuwahara, R., et al. (2018). Lrit1, a Retinal Transmembrane Protein, Regulates Selective Synapse Formation in Cone Photoreceptor Cells and Visual Acuity. Cell Rep. *22*, 3548–3561.

119. Lee, H.-J., and Zheng, J.J. (2010). PDZ domains and their binding partners: structure, specificity, and modification. Cell Commun. Signal. *8*, 8.

120. Stankiewicz, P., Kulkarni, S., Dharmadhikari, A.V., Sampath, S., Bhatt, S.S., Shaikh, T.H., Xia, Z., Pursley, A.N., Cooper, M.L., Shinawi, M., et al. (2012). Recurrent deletions and reciprocal duplications of 10q11.21q11.23 including CHAT and SLC18A3 are likely mediated by complex low-copy repeats. Hum. Mutat. *33*, 165–179.

121. Muntané, G., Horvath, J.E., Hof, P.R., Ely, J.J., Hopkins, W.D., Raghanti, M.A., Lewandowski,

A.H., Wray, G.A., and Sherwood, C.C. (2015). Analysis of synaptic gene expression in the neocortex of primates reveals evolutionary changes in glutamatergic neurotransmission. Cereb. Cortex *25*, 1596–1607.

122. Willcox, B.J., Donlon, T.A., He, Q., Chen, R., Grove, J.S., Yano, K., Masaki, K.H., Willcox, D.C., Rodriguez, B., and Curb, J.D. (2008). FOXO3A genotype is strongly associated with human longevity. Proc. Natl. Acad. Sci. U. S. A. *105*, 13987–13992.

123. Antonacci, F., Dennis, M.Y., Huddleston, J., Sudmant, P.H., Steinberg, K.M., Rosenfeld, J.A., Miroballo, M., Graves, T.A., Vives, L., Malig, M., et al. (2014). Palindromic GOLGA8 core duplicons promote chromosome 15q13.3 microdeletion and evolutionary instability. Nat. Genet. *46*, 1293–1302.

124. Yoo, D., Rhie, A., Hebbar, P., Antonacci, F., Logsdon, G.A., Solar, S.J., Antipov, D., Pickett, B.D., Safonova, Y., Montinaro, F., et al. (2024). Complete sequencing of ape genomes. bioRxiv. https://doi.org/10.1101/2024.07.31.605654.

125. Makova, K.D., Pickett, B.D., Harris, R.S., Hartley, G.A., Cechova, M., Pal, K., Nurk, S., Yoo, D., Li, Q., Hebbar, P., et al. (2024). The complete sequence and comparative analysis of ape sex chromosomes. Nature *630*, 401–411.

126. L Rocha, J., Lou, R.N., and Sudmant, P.H. (2024). Structural variation in humans and our primate kin in the era of telomere-to-telomere genomes and pangenomics. Curr. Opin. Genet. Dev. *87*, 102233.

127. Tropepe, V., and Sive, H.L. (2003). Can zebrafish be used as a model to study the neurodevelopmental causes of autism? Genes Brain Behav. *2*, 268–281.

128. Golzio, C., Willer, J., Talkowski, M.E., Oh, E.C., Taniguchi, Y., Jacquemont, S., Reymond, A., Sun, M., Sawa, A., Gusella, J.F., et al. (2012). KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature *485*, 363–367.

129. Hoffman, E.J., Turner, K.J., Fernandez, J.M., Cifuentes, D., Ghosh, M., Ijaz, S., Jain, R.A., Kubo, F., Bill, B.R., Baier, H., et al. (2016). Estrogens Suppress a Behavioral Phenotype in Zebrafish Mutants of the Autism Risk Gene, CNTNAP2. Neuron *89*, 725–733.

130. Shah, A.N., Davey, C.F., Whitebirch, A.C., Miller, A.C., and Moens, C.B. (2015). Rapid reverse genetic screening using CRISPR in zebrafish. Nat. Methods *12*, 535–540.

131. Thyme, S.B., Pieper, L.M., Li, E.H., Pandey, S., Wang, Y., Morris, N.S., Sha, C., Choi, J.W., Herrera, K.J., Soucy, E.R., et al. (2019). Phenotypic Landscape of Schizophrenia-Associated Genes Defines Candidates and Their Shared Functions. Cell *177*, 478–491.e20.

132. Glasauer, S.M.K., and Neuhauss, S.C.F. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. Mol. Genet. Genomics *289*, 1045–1060.

133. Fossati, M., Pizzarelli, R., Schmidt, E.R., Kupferman, J.V., Stroebel, D., Polleux, F., and Charrier, C. (2016). SRGAP2 and Its Human-Specific Paralog Co-Regulate the Development of Excitatory and Inhibitory Synapses. Neuron *91*, 356–369.

134. Schmidt, E.R.E., Kupferman, J.V., and Stackmann, M. (2019). The human-specific paralogs SRGAP2B and SRGAP2C differentially modulate SRGAP2A-dependent synaptic development.

Scientific Reports *9*. https://doi.org/10.1101/596940.

135. Kalebic, N., and Huttner, W.B. (2020). Basal Progenitor Morphology and Neocortex Evolution. Trends Neurosci. *43*, 843–853.

136. Bitarello, B.D., de Filippo, C., Teixeira, J.C., Schmidt, J.M., Kleinert, P., Meyer, D., and Andrés, A.M. (2018). Signatures of Long-Term Balancing Selection in Human Genomes. Genome Biol. Evol. *10*, 939–955.

137. Andrés, A.M., Hubisz, M.J., Indap, A., Torgerson, D.G., Degenhardt, J.D., Boyko, A.R., Gutenkunst, R.N., White, T.J., Green, E.D., Bustamante, C.D., et al. (2009). Targets of balancing selection in the human genome. Mol. Biol. Evol. *26*, 2755–2764.

138. Bitarello, B.D., Brandt, D.Y.C., Meyer, D., and Andrés, A.M. (2023). Inferring Balancing Selection From Genome-Scale Data. Genome Biol. Evol. *15*. https://doi.org/10.1093/gbe/evad032.

139. Mahmoud, M., Huang, Y., Garimella, K., Audano, P.A., Wan, W., Prasad, N., Handsaker, R.E., Hall, S., Pionzio, A., Schatz, M.C., et al. (2024). Utility of long-read sequencing for All of Us. Nat. Commun. *15*, 837.

140. Searles Quick, V.B., Wang, B., and State, M.W. (2021). Leveraging large genomic datasets to illuminate the pathobiology of autism spectrum disorders. Neuropsychopharmacology *46*, 55–69.

141. Vollger, M.R., DeWitt, W.S., Dishuck, P.C., Harvey, W.T., Guitart, X., Goldberg, M.E., Rozanski, A.N., Lucas, J., Asri, M., The Human Pangenome Reference Consortium, et al. (2022). Increased mutation rate and interlocus gene conversion within human segmental duplications. bioRxiv, 2022.07.06.498021. https://doi.org/10.1101/2022.07.06.498021.

142. Dumont, B.L. (2015). Interlocus gene conversion explains at least 2.7% of single nucleotide variants in human segmental duplications. BMC Genomics *16*, 456.

143. Dumont, B.L., and Eichler, E.E. (2013). Signals of historical interlocus gene conversion in human segmental duplications. PLoS One *8*, e75949.

144. Hardwick, R.J., Machado, L.R., Zuccherato, L.W., Antolinos, S., Xue, Y., Shawa, N., Gilman, R.H., Cabrera, L., Berg, D.E., Tyler-Smith, C., et al. (2011). A worldwide analysis of beta-defensin copy number variation suggests recent selection of a high-expressing DEFB103 gene copy in East Asia. Hum. Mutat. *32*, 743–750.

145. Hughes, T., Hansson, L., Akkouh, I., Hajdarevic, R., Bringsli, J.S., Torsvik, A., Inderhaug, E., Steen, V.M., and Djurovic, S. (2020). Runaway multi-allelic copy number variation at the α-defensin locus in African and Asian populations. Sci. Rep. *10*, 9101.

146. Linzmeier, R.M., and Ganz, T. (2005). Human defensin gene copy number polymorphisms: comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. Genomics *86*, 423–430.

147. Mohajeri, K., Cantsilieris, S., Huddleston, J., Nelson, B.J., Coe, B.P., Campbell, C.D., Baker, C., Harshman, L., Munson, K.M., Kronenberg, Z.N., et al. (2016). Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the Chromosome 8p23.1 region. Genome Res. *26*, 1453–1467.

148. Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P.,

Chavan, S., Vergara, C., Ortega, V.E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat. Genet. *51*, 30–35.

149. Plender, E.G., Prodanov, T., Hsieh, P., Nizamis, E., Harvey, W.T., Sulovari, A., Munson, K.M., Kaufman, E.J., O'Neal, W.K., Valdmanis, P.N., et al. (2024). Structural and genetic diversity in the secreted mucins MUC5AC and MUC5B. Am. J. Hum. Genet. *111*, 1700–1716.

150. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.

151. Jeong, H., Dishuck, P.C., Yoo, D., Harvey, W.T., Munson, K.M., Lewis, A.P., Kordosky, J., Garcia, G.H., Human Genome Structural Variation Consortium (HGSVC), Yilmaz, F., et al. (2024). Structural polymorphism and diversity of human segmental duplications. bioRxiv. https://doi.org/10.1101/2024.06.04.597452.

152. Kim, S., Kim, H., Park, D., Kim, J., Hong, J., Kim, J.S., Jung, H., Kim, D., Cheong, E., Ko, J., et al. (2024). Loss of IQSEC3 Disrupts GABAergic Synapse Maintenance and Decreases Somatostatin Expression in the Hippocampus. Cell Rep. *43*, 114254.

153. Pardo-Palacios, F.J., Wang, D., Reese, F., Diekhans, M., Carbonell-Sala, S., Williams, B., Loveland, J.E., De María, M., Adams, M.S., Balderrama-Gutierrez, G., et al. (2023). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. bioRxiv. https://doi.org/10.1101/2023.07.25.550582.

154. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

155. Timp, W., and Timp, G. (2020). Beyond mass spectrometry, the next step in proteomics. Science Advances. https://doi.org/10.1126/sciadv.aax8978.

156. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A., Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.M., et al. (2021). Segmental duplications and their variation in a complete human genome. bioRxiv, 2021.05.26.445678. https://doi.org/10.1101/2021.05.26.445678.

157. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

158. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

159. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

160. Hartasánchez, D.A., Brasó-Vives, M., Heredia-Genestar, J.M., Pybus, M., and Navarro, A. (2018). Effect of Collapsed Duplications on Diversity Estimates: What to Expect. Genome Biol. Evol. *10*, 2899–2905.

161. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012.

162. Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. (2006). The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. *34*, D590–D598.

163. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

164. Soneson, C., Love, M.I., and Robinson, M.D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. *4*, 1521.

165. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) *2*, 100141.

166. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. Scientific Data *3*, 160025.

167. Bergström, A., McCarthy, S.A., Hui, R., Almarri, M.A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science *367*. https://doi.org/10.1126/science.aay5012.

168. Witek, K., Jupe, F., Witek, A.I., Baker, D., Clark, M.D., and Jones, J.D.G. (2016). Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. Nat. Biotechnol. *34*, 656–660.

169. Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Van der Auwera, G.A., Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., et al. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv, 201178. https://doi.org/10.1101/201178.

170. Stecher, G., Tamura, K., and Kumar, S. (2020). Molecular Evolutionary Genetics Analysis (MEGA) for macOS. Mol. Biol. Evol. *37*, 1237–1239.

171. Leigh, J.W., and Bryant, D. (2015). Popart: Full-feature software for haplotype network construction. Methods Ecol. Evol. *6*, 1110–1116.

172. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

173. LaFave, M.C., Varshney, G.K., Vemulapalli, M., Mullikin, J.C., and Burgess, S.M. (2014). A defined zebrafish line for high-throughput genetics and genomics: NHGRI-1. Genetics *198*, 167–170.

174. Westerfield, M. (1995). The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio Rerio).

175. Jao, L.-E., Wente, S.R., and Chen, W. (2013). Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. Proc. Natl. Acad. Sci. U. S. A. *110*, 13904–13909.

176. Varshney, G.K., Carrington, B., Pei, W., Bishop, K., Chen, Z., Fan, C., Xu, L., Jones, M., LaFave, M.C., Ledin, J., et al. (2016). A high-throughput functional genomics workflow based on CRISPR/Cas9-mediated targeted mutagenesis in zebrafish. Nat. Protoc. *11*, 2357–2375.

177. Tsai, S.Q., Nguyen, N.T., Malagon-Lopez, J., Topkar, V.V., Aryee, M.J., and Joung, J.K. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. Nat. Methods *14*, 607–614.

178. Lazzarotto, C.R., Nguyen, N.T., Tang, X., Malagon-Lopez, J., Guo, J.A., Aryee, M.J., Joung, J.K., and Tsai, S.Q. (2018). Defining CRISPR-Cas9 genome-wide nuclease activities with CIRCLE-seq. Nat. Protoc. *13*, 2615–2642.

179. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

180. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). Fiji: an open-source platform for biological-image analysis. Nat. Methods *9*, 676–682.

181. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models.

182. Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM *60*, 84–90.

183. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In 2017 IEEE International Conference on Computer Vision (ICCV) (IEEE). https://doi.org/10.1109/iccv.2017.74.

184. Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science *360*, 176–182.

185. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.

186. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

187. Andrews, S., and Others (2010). FastQC: a quality control tool for high throughput sequence data. Preprint at Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

188. Lawson, N.D., Li, R., Shin, M., Grosse, A., Yukselen, O., Stone, O.A., Kucukural, A., and Zhu, L. (2020). An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes. Elife *9*. https://doi.org/10.7554/eLife.55792.

189. 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.

190. Chen, G., Liu, Z., and Peng, C. (2021). Multimodal and Integrative Analysis of Single-Cell or Bulk Sequencing Data (Frontiers Media SA).

191. Tran, V., Papalexi, E., Schroeder, S., Kim, G., Sapre, A., Pangallo, J., Sova, A., Matulich, P., Kenyon, L., Sayar, Z., et al. (2022). High sensitivity single cell RNA sequencing with split pool barcoding. bioRxiv, 2022.08.27.505512. https://doi.org/10.1101/2022.08.27.505512.

192. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst *8*, 329–337.e4.

193. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

194. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

195. Bradford, Y.M., Van Slyke, C.E., Ruzicka, L., Singer, A., Eagle, A., Fashena, D., Howe, D.G., Frazer, K., Martin, R., Paddock, H., et al. (2022). Zebrafish information network, the knowledgebase for Danio rerio research. Genetics *220*. https://doi.org/10.1093/genetics/iyac016.

196. 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.