Human-specific gene expansions contribute to 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 La[1,2], Natasha Ann F. Mariano[1,2,5], Cole Ingamells[1,2], Aidan Baraban[1,2], Zoeb Jamal[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: Neuroscience Research Institute, University of California, Santa Barbara, CA 93106

[†]Lead Contact:
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

Summary

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, long-read DNA sequencing of hundreds of modern humans revealed previously hidden signatures of selection, including for T-cell marker *CD8B*. To understand 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 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 insights and a comprehensive resource for studying gene expansion drivers of human brain evolution.

Keywords: segmental duplications, gene duplications, human evolution, brain, neurodevelopment, sequencing, zebrafish, copy-number variation, gene expression

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[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[12] and contribute more to genetic divergence across species than SNVs[13]. Previous comparisons of great ape genomes have identified >30 human-specific gene families and hundreds of paralogs important in neurodevelopment and enriched at genomic hotspots associated with neuropsychiatric disorders[14–16]. Of these, a handful of genes have been found to function in brain development using model systems, including *SRGAP2C*[17,18], *NOTCH2NL*[19–21], *ARHGAP11B*[22–24], *TBC1D3*[25] *CROCCP2*[26], and *LRRC37B*[27]. Most studies have leveraged mice to study gene functions with recent studies expanding to cortical organoids, ferrets, and primates[28]. 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 because of difficulties in accurate genome assembly[29] and discovering variants across nearly identical paralogs[30–34]. As such, many human-duplicated genes are likely left to be discovered. The telomere-to-telomere (T2T) human reference genome T2T-CHM13[35], representing a gapless sequence of all autosomes and chromosome X, has enabled a more complete picture of SDs[36] by incorporating hundreds of megabases missing from the previous human reference genome (GRCh38). This new assembly corrects >8 Mbp of collapsed duplications[37], including previously missing paralogs of human-specific duplicated gene families[14] *GPRIN2*[36] and *DUSP22*[37]. Here, using this new T2T genome, we identified thousands of recent gene duplications among hominids. By comparing great ape genomic data, 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 neurodevelopment 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 elucidate brain functions.

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[36,38] (Figure 1A). These loci represent genes duplicated only in human lineage[14,15] as well as expansions of duplicated gene families present in other great apes. Consistent with the notion that gene duplication can lead to functional innovation, a number of paralogs in the latter category have experienced recent changes along the *Homo* lineage in expression (e.g., *LRRC37B*[27]) or sequence content (e.g., *NOTCH2NL*, via interlocus gene conversion[19]). Focusing our analysis on autosomes, we identified 698 protein-encoding genes and 1,095 unprocessed pseudogenes representing possible mis-annotations of true protein-encoding genes[39] (Table S1A;, 478 paralogs on sex chromosomes in Table S1B). This list includes well-known genes important in neurodevelopment (*SRGAP2C*, *ARHGAP11B*), disease (*SMN1* and *SMN2*[40], *KANSL1*[41]), and adaptation (amylase[42–44]), with 668 (37%) residing in previously missing or erroneous regions in GRCh38. Sequence read depth[36] in modern humans (Simons Genome Diversity Project [SGDP], n=269[45]) verified that all paralogs had >2 gene-family diploid copy number (famCN; STAR Methods, Table S1C, Figure 1B).

Based on sequence and famCN similarity, we clustered 1,679 of the paralogs into 491 multigene families, with many having 2–3 members (n=271 families) (Figures 1C and S1A,B). Three extreme high-copy gene families had >50 paralogs, including macrosatellite-associated *DUX4* and *DUB/USP17* as well as primate-specific *FAM90A*[46]. The remaining 114 paralogs were defined as "singletons" (Table S1C), 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*). We defined 385 of these genes as human-specific, falling within non-syntenic human and chimpanzee reference regions[36] (Table S1C, Figure 1C). Because several known human-specific paralogs were absent from our list (e.g., *NPY4R2*, *ROCKP1*, and *SERF1B*[14]), we also narrowed in on 97 human-expanded gene families and 27 singletons with higher famCN in humans versus nonhuman great apes (see STAR Methods). In total, we conservatively predict 213 gene families and 38 singletons comprising at least one human-specific duplicate paralog (Table S1D). Moving forward, we refer to any of the 1,002 paralogs within one of these gene families as a human-specific duplicated gene.

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[47] to identify 622 "CN

constrained" genes (parCN≥0.5; >98% 1000 Genomes Project, 1KGP; n=2,504) and 125 paralogs "fixed" in humans (parCN~2) (Table S1E). Thirteen genes represent *Homo sapiens*-specific gene duplications largely absent from four archaic human genomes[48–50]; these include *H3-2/H2BP2*, a member of a core *H2B* histone family involved in the structure of eukaryotic chromatin[51], homologous with human-specific *H2BP1* and the ancestral *H2BC18* paralog (Figure 1D), as well as *FCGR1CP*, encoding an immunoglobulin gamma Fc Gamma Receptor implicated in regulating immune response[52].

We identified 13 protein-encoding genes as loss-of-function intolerant using SNV data from hundreds of thousands of humans from gnomAD[53] (Table S1A, Figure S1C), showing that deleterious mutations of these genes are depleted in human populations. The gnomAD (v3) metrics rely on variants identified in protein-encoding genes using the human reference genome hg19, which has known errors across SDs[54] 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 in the 1KGP cohort by Tajima's D[37,55,56] (Figures S1D–G) Genic SD98 loci exhibit significantly reduced Tajima's D compared with non-coding regions, indicating increased functional constraint and matching results in non-duplicated regions (Figure S1H). We identified 15 CN-constrained human-duplicated genes with the most negative D values (<5th percentile; see Methods) considered outliers, suggestive of signatures of positive or strong levels of purifying selection (Table S1F, Figures 1E and S1I). These included human-specific paralog *SRGAP2C* previously implicated in cortical neuronal migration and synaptogenesis[17,18] as well as the uncharacterized *LRRC37A3* and the hominid-specific *LRRC37B*, recently found to function in cortical pyramidal neurons by impacting synaptic excitability[27]. We also identified nine genes exhibiting the highest D values (>95th percentile), suggestive of signatures of balancing selection, including T-cell antigen *CD8B.* We note that duplicated genes exhibiting non-extreme Tajima's D can also be functionally constrained and provide these results across all short-read accessible regions as a resource[57]. Collectively, variants discovered using the new T2T-CHM13 genome enabled the identification of human-duplicated genes potentially contributing to traits and diseases not previously assayed in genome-wide selection screens.

Human-duplicated genes implicated in brain development

Connecting genetic variation of duplicated genes with neural traits

Considering gene ontology (GO) of paralogs from human-duplicated families (n=1,002) and CN constrained (n=622), we did not identify any enrichments of functional features. We note that only 24% (236/1,002) of duplicated genes can be assigned a GO versus 79% of all protein-encoding and unprocessed pseudogenes, highlighting that most paralogs have unknown functions. To narrow in on human-duplicate gene families contributing to neurocognitive features, we identified 341 paralogs (187 CN-constrained) with putative associations with brain-related phenotypes by intersecting with the genome-wide association studies (GWAS) catalog and UK Biobank[58] (Table S1A, Figure S1J, STAR Methods). Many human-duplicated genes reside at genomic hotspots (n=305), such as *GPR89* paralogs at chromosome 1q21.1 with recurrent ~2 Mbp deletions/duplications associated with autism and impacting brain size[59]. To better delineate copy-number variants (CNVs), we assayed parCN in an autism cohort (Simons Simplex Collection [SSC]; n=2,459 quad families [60,61]). Eighteen duplicated genes residing at autism-associated hotspots, including chromosomes 15q25.2 (OMIM:614294) and 3q29 (OMIM:609425),

show significant parCN differences in probands versus unaffected siblings (Wilcoxon signed-rank test, *q*-value<0.05) (Figure S1K). *De novo* CNVs impact 22 human-duplicated genes in autistic probands in contrast to six events impacting five paralogs in unaffected siblings (Fisher's exact test, *p*-value=4.5×10^{-4}) (Table S1G, Figure S1L). While a majority of impacted genes reside at known hotspots, some did not, including *CD8B2*, *FCGR1B*, *HYDIN* and *LIMS1*, representing possible contributors to autism.

Duplicated gene expression in the developing human brain

Nearly half of human-duplicated gene paralogs (455/1,002) are expressed during brain development (TPM≥1)[22,62–65] (Table S1A, Figures 2A and S2A). While this represents a depletion versus the complete transcriptome (18,918/23,395; hypergeometric test, *p*-value=4.76×10^{-146}), the number of brain-expressed genes increases to 58% for CN-constrained (1.3-fold enrichment over all human-duplicated genes, *p*-value=2.5×10^{-24}) and to 84% for CN-constrained protein-encoding (1.4-fold enrichment, *p*-value=7.8×10^{-30}) (Figures 2B and C). Similar expression increases are also observed in lymphoblastoid cell lines[66] (Figure S2B) showing this is not specific to the brain but, instead, suggests true functional candidate genes are more likely to exist in CN-constrained protein-encoding genes.

We inferred possible functions of human-duplicated genes expressed during brain development using weighted gene co-expression network analysis (WGCNA)[67]. Assessment of post-mortem human prenatal frontal cortex transcriptomes spanning post-conception weeks 8 to 22 from BrainSpan[65] revealed 17 modules with at least one human-duplicated gene (Table S2A, Figures 2D and S2C, Data S1). Of these, only the B-turquoise module, exhibiting low expression in early development and increases during cortical specification and the beginning of deep-layer formation (16 to 20 post-conception weeks), was enriched for human duplicated genes (n=76; *p*-value=4.17×10^{-9}), in addition to autism-associated genes[68] (n=61; *p*-value=1.86×10^{-9}). Co-expressed genes within this module included several markers of GABAergic interneurons (e.g., *DLX1* and *DLX2*[69,70]) and deep-layer excitatory neurons (*SSTR2*[71]), with overrepresentation of GO terms related to neurogenesis, axonogenesis, and dendrite development (Table S2B). Based on these results, we next used transcriptomes from *ex-vivo*-induced neurons modeling human prenatal prefrontal cortex from pluripotency to upper layer formation (CORTECON[64]) to identify 21 WGCNA modules (Table S2C, Figures 2F and S2D, Data S1). Considering co-expression patterns within gene families, paralogs largely belong to different CORTECON modules with only six duplicate gene families in complete concordance (all paralogs in the same module) (Table S2D, Figure 2G). This demonstrates 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[72].

Five CORTECON modules were significantly enriched for paralogs from human-specific gene families (*p*-values<0.05; Figures 2F and S2D)—highlighting neural functions related to stem cell population maintenance, intracellular receptor signaling, microtubule-based movement, and regulation of neuron projections (GO enrichments in Table S2E)—and including *SRGAP2C*, a human-specific gene known to interact with F-actin to produce membrane protrusions required for neuronal migration and synaptogenesis[73] (C-blue module). Three modules were also enriched for autism-associated genes[68] (*q*-value<0.05; Figure 2F), including the C-yellow module (n=20 candidate genes) associated with axon guidance and synaptogenesis (Table S2E). Duplicated genes in this module include *KANSL1*, the causal gene in Koolen-de Vries syndrome[74], and others within autism-associated genomic hotspots (e.g.,

CASTOR2, *PMS2P6* and *STAG3L1* at the Williams-Beuren syndrome region (OMIM:609757)) making them compelling candidates in contributing to neurological features in children carrying CNVs at these loci (Figure S2E).

Collectively, our analysis identifies co-expression modules enriched for human-duplicated genes, such as the B-turquoise and C-blue modules, which both relate to regulation of neuron projection development. Additionally, we provide a complete list of duplicate paralog module assignments using data from post-mortem brain tissue and *in vitro* induced neurons to provide clues of their putative functions in brain development. Paralogs co-expressed in modules enriched for genes with known links to neurodevelopment and autism represent top candidates for follow-up experiments.

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 and brain-expressed human-duplicated paralog, in addition to 30 paralogs not assigned to a family (Table S1A, Figure 3). Of these, we found 106 with a homologous gene(s) in either mouse or zebrafish (Table S3A). Using matched brain-expression data from these species corresponding to human developmental stages[65,75,76] (Figure S3, as previously described[77,78]) narrowed in on 76 and 41 single-copy orthologs expressed during neurodevelopment in mice and zebrafish (Table S3B), respectively, enabling functional studies. This leaves 40% of the human-duplicated families with no obvious mouse/zebrafish ortholog, including fusion genes, primate-specific genes (e.g., *TBC1D3* paralogs[25,79]), or those associated with great ape ancestral "core" duplicons[80] (e.g., *NBPF* and *NPIP*). 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 S4A, Figure 3). Since none of the paralogs fully reside within short-read-accessible genomic regions due to their high identity, we used published draft assemblies (112 total haplotypes)[81–84] from the Human Pangenome Reference Consortium (HPRC) and Human Genome Structural Variation Consortium (HGSVC; Figure S4A) and performed capture high-fidelity (cHiFi) sequencing of 144 unrelated individuals of diverse ancestries[55,85] (Table S4B–E, Figure S4B–E; see STAR Methods for details) to identify 46,754 variants (33,774 SNVs and 12,980 indels), or 12.7 variants/kbp, across targeted regions. 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 S4F). 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[18].

Functional annotation[86] identified 412 gene-impacting variants (Table S4F,G, Figure 4A), with eleven paralogs exhibiting no likely gene-disruptive (LGD) variants suggesting strong selective constraint. Virtually all paralogs had Ka/Ks lower than one, suggesting purifying selection, with seven ancestral and

three derived paralogs exhibited Ka/Ks below the genome-wide average (~0.25)[87]. The ancestral paralogs exhibited significantly lower Ka/Ks values than their derived paralogs (Wilcoxon signed-rank test, *p*-value=0.03) (Figure 4B), consistent with stronger purifying selection. More recent selection signatures incorporating polymorphic variation (pN/pS and the direction of selection [DoS][88]) similarly indicated stronger purifying selection in the ancestral versus derived paralogs (Wilcoxon signed-rank test, *p*-value=0.023) (Figure 4C, Table S4H). While 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 have patterns expected 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[89]; Table S4I).

Human-specific *SRGAP2C* has elevated Ka/Ks and pN/pS, together with low Tajima's D in African individuals from the 1KGP genome-wide screen (-2.32; Figure 1E) suggesting positive selection, which we validated using high-confidence variants obtained from genome assemblies (-2.14; Figure S4G). *GPR89* gene family paralogs exhibit low nucleotide diversity π and negative Tajima's D values consistent with functional constraints (Figure S4H). *ROCK1* showed reduced π and more negative Tajima's D compared to *ROCK1P1*, consistent with their divergent Ka/Ks values (Figure S4I). 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 S4J). Revisiting the 1KGP genome-wide signal of balancing selection in individuals of American and European ancestries centered on *CD8B* (Figures 1E and S1I), we find positive Tajima's D in American (max 2.66, n=18) but not in African ancestries (max 0.62, n=27) (Figure 4D), evident as distinct haplotype clusters (Figure S4K).

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[90,91]. Considering all variants identified using assemblies and cHiFi sequencing, we observe an increase in intermediate-frequency variants, a signature of balancing selection, in *CD8B* among European and American ancestries, compared with those of African ancestry (Figure S4L, Kolmogorov-Smirnov, *p*-value=2.2×10^{-16}), that differentiate the two main haplotypes. Two of the SNPs (rs56063487 and rs6547706) are *CD8B* splice eQTLs in whole blood from GTEx[92] and significantly associated with increased CD8-protein levels on CD8+ T cells within a Sardinian cohort[93]. 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. The human-specific paralog *CD8B2* exhibits divergent expression in the human postnatal brain rather than in T cells[39] (Figure 4E). 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][14]). Combined, we demonstrate the efficacy of long-read data to uncover hidden signatures of natural selection.

Duplicated gene functions modeled using zebrafish

We performed a high-throughput functional screen in zebrafish[94–96] of seven largely uncharacterized pHSD families expressed in both human and zebrafish brain[97,98] (*GPR89*, *NPY4R*, *PTPN20*, *PDZK1*,

HYDIN, *FRMPD2,* and *FAM72*; Table S3B, Figures 3 and S5A–C). Additionally, we tested two gene families (*SRGAP2* and *ARHGAP11*) previously studied in mammals[17,22–24,73,99–103]. While a whole-genome teleost duplication resulted in ~20% of genes with multiple zebrafish paralogs that might confound functional analysis of human gene duplications[104], the nine prioritized gene families tested here were selected in part because each had only one zebrafish ortholog (Table S3A). Ancestral gene functions were assessed using loss-of-function knockouts resulting in ~70% ablation of alleles in G_0 lines[105] (termed crispants; Table S6 and STAR Methods) except for *arhgap11*, which is maternally deposited[106], prompting us to use a morpholino that impedes translation. We also 'humanized' zebrafish models by introducing mRNAs encoding human-specific paralogs (Figure 5A) into wild-type embryos for all genes except *HYDIN2*, due to its large size (4,797 amino acids [aa]). This produced transient and ectopic presence of the transcript, detectable by RT-PCR at 3 dpf (Figure S5D). There were no significant morbidity differences in any tested models compared to controls (log-rank survival tests *p*-values>0.05, Table S5B).

Significant morphological differences were first identified without predefining specific features *a priori* by imaging[107–109] at 3 and 5 days post-fertilization (dpf) and using latent diffusion and convolutional neural networks (CNNs)[110] (Table S5C, STAR Methods). This flagged knockout and humanized models of *SRGAP2*, *GPR89*, *FRMPD2*, and *FAM72* (F1 scores>0.2, Figure 5B, STAR Methods). Quantifying specific features using the same images (Table S5D,E) revealed concordant phenotypes for knockout and humanized models of *SRGAP2* (reduced length), *FRMPD2* (reduced head area), and *FAM72* (both reduced body length and head area) at 3 dpf (Figure 5C). 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 5B). At 5 dpf, the alterations in *FRMPD2* and *SRGAP2* models persisted while no longer observed for *FAM72* and *GPR89* (Figure 5C). Knockout models for *gpr89* and *frmpd2* also displayed evidence of developmental delay with subtle yet significant decreases in the head-trunk angle[111].

To directly characterize impacts on brain development, we profiled 95,555 cells (an average of 3,822±3,227 per model; Figure S5E) using single-cell RNA-sequencing (scRNA-seq)[112,113] from dissected heads of 3 dpf larvae (Figure 5D). Pseudo-bulk differential expression analysis of all cells revealed significant positive correlations for *SRGAP2C*, *FAM72B*, *ARHGAP11B*, *FRMPD2B*, and *PDZK1B* humanized larvae with respect to each knockout indicating loss-of-function effects (Figure 5E). *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 5C), as well as from our separate study[114] that verified the human SRGAP2C protein physically interacts with and antagonizes zebrafish Srgap2. The concordant phenotypes for gain- and loss-of-function for over half of the tested genes fits the gene-balance hypothesis[115,116] and recent results in humans showing that large increases and decreases in gene expression via CNVs can impact certain complex traits in the same direction[117].

Classifying 17 different neuronal, retinal, and glial cell types[76,112,118,119] shows pHSD orthologs are broadly expressed, with a subset showing more narrow expression patterns (e.g., *hydin* and *pdzk1* in the pallium, *npy4r* in the hindbrain; Table S5F,G, Figure 5F). Pseudo-bulk analyses revealed gene

dysregulation in most cell types (Figure 5G; DEGs available here[57]). GO enrichment of DEGs in the forebrain (the closest related structure to the human cerebral cortex[120]) and midbrain (the main visual processing center primarily consisting of the optic tectum), where we had the greatest power to detect differences due to the largest abundances of cells, suggests alterations in cell-cell adhesion, chemotaxis, and altered synaptic signaling (Figure 5H, Table S5H). Several humanized models exhibited unique effects in DEGs, such as in the midbrain of *GPR89B* models and regulation of anatomical structure size, Müller glia in *PTPN20CP*, and the spinal cord in humanized *NPY4R2*. Myeloid cells were also uniquely impacted in *gpr89* and *arhgap11* knockout larvae. 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.

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[59], we sought to characterize mechanisms underlying larval head-size phenotypes in more detail. Generating a stable *gpr89* mutant line (STAR Methods) showed that heterozygous and homozygous knockouts exhibited reduced head size at 3 dpf, verifying results in crispants (Figures 5C and 6A). Further, we observed significantly smaller and larger forebrains in crispant G_0 knockout and humanized zebrafish larvae, respectively, using a neuronal reporter line[121]. Sub-clustering cells from the forebrain, we observed endogenous expression of *gpr89* in telencephalon and inner diencephalon (Figure 6B). DEGs with inverse effects between *GPR89* knockout and humanized models in the telencephalon, a brain structure anatomically equivalent to the mammalian forebrain with roles in higher cognitive functions such as social behavior and associative learning[122,123], were enriched in negative regulation of the DNA replication and cell cycle (Figure 6C, Table S6AB). 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. We estimated the identity of forebrain cells and found humanized *GPR89B* cells more likely to classify as neural progenitors while *gpr89* knockouts more likely to be differentiated neurons (Figure 6D).

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[124]. Loss of function in *Drosophila* leads to global growth deficiencies due to defects in the secretory pathway[125]. In humans, a complete duplication of ancestral *GPR89A* ~4.7 mya produced a derived *GPR89B*[14] (Figure 6E). The two paralogs maintain identical protein similarity but differential and overlapping expression patterns in human brain development, with *GPR89A* evident in pluripotency (C-turquoise module), and *GPR89B* expression turning on slightly later during neural differentiation (C-red module; Figures 2F and 6E). Both genes show evidence of purifying selection (Figure S4), with *GPR89A* exhibiting extreme negative Tajima's D values in individuals of African and American ancestries from the 1KGP cohort (<5th percentile; Figure 1E). These results suggest that *GPR89* paralogs function in early brain development, with delayed expression of *GPR89B* possibly extending expansion of progenitor cells, a feature observed in human cerebral organoids compared with those of other apes[126,127] (Figure 6E). Together with the

increase in forebrain size of "humanized" zebrafish, we propose 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 suggest that the human FRMPD2B acts as a dominant negative to the endogenous Frmpd2. We generated a stable *frmpd2* mutant line (STAR Methods) and observed reduced head size in homozygous knockout larvae validating crispant features (Figures 5C and 6A). Additionally, both the crispant knockout *frmpd2* and humanized *FRMPD2B* larvae exhibit smaller forebrains. Shared upregulated DEGs between *FRMPD2* models function in cell/axon morphogenesis and growth as well as synaptic signaling in telencephalic cells (Figure 6C, Table S6C,D). To better characterize impacts on synaptic signaling, we treated with a low dose of the GABA-antagonizing drug pentylenetetrazol (PTZ) producing a significant increase in high-speed events[128] , indicative of seizures in larvae, in both humanized and knockout *FRMPD2* larvae (4 dpf) versus controls (Figure 6F). These results suggest that Frmpd2 loss of function, through *frmpd2* knockout or antagonism via FRMPD2B, disrupts synapse transmission and amplifies induced seizures, in line with the known interactions of FRMPD2 with glutamate receptors[129].

FRMPD2 (FERM and PDZ domain containing 2) encodes a scaffold protein that participates in cell-cell junction and polarization[130] localized at photoreceptor synapses[131] and the postsynaptic membrane in hippocampal neurons in mice[129]. A partial duplication of the ancestral *FRMPD2* created the 5'-truncated *FRMPD2B* paralog ~2.3 mya[14]. *FRMPD2B* encodes 320 aa of the C-terminus, versus 1,284 aa for the ancestral, maintaining two of three PDZ domains involved in protein binding[132] but lacking the KIND and FERM domains (Figure 6G). Ancestral *FRMPD2* expresses in the human prenatal cortex during upper layer formation, while *FRMPD2B* is evident postnatally[65]. The paralogs also show divergent evolutionary signatures, with *FRMPD2* strongly conserved and *FRMPD2B* exhibiting possible positive selection (Figure 4B,C). 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[129]. 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[133]. 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[134].

Discussion

Our results provide the scientific community with a prioritized and comprehensive set of hundreds of genes to perform functional analyses with the goal to identify drivers of human brain evolution (213 families and 1,002 total paralogs). Compared to a previous assessment of human-specific duplicated genes[14], this represents an approximately fivefold increase, in part because we also included human-expanded gene families and genes with as little as one duplicated exon. These numbers are likely an underestimate, as we excluded 193 high-copy gene families (famCN>10), as well as 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[135], with all three paralogs CN-constrained and brain expressed. 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 due to secondary structural rearrangements that hamper genomic alignments[68,136]. Moving forward, the availability of nonhuman primate T2T genomes will improve orthology and synteny comparisons between species[137–139]. As a resource for the community, we have made available the results of our genome-wide analyses across the complete 1,793 SD98 genes (Table S1A).

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. 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[140], conservation in major brain features make zebrafish well suited to characterize gene functions in neurological traits, including cranial malformations[141], neuronal imbalances[142], and synaptogenesis[143]. Coupled with CRISPR mutagenesis[94,95], zebrafish have been used as a higher-throughput model for human conditions such as epilepsy[128], schizophrenia[144], and autism[78].

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 (Figure 5G). 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[17]. Further, the shared upregulated genes identified in the forebrains of *SRGAP2* mutant larvae point to alterations in axonogenesis and cell migration (Table S5I), matching studies in mice[11,17,18,73,100,145,146]. Alternatively, *ARHGAP11B*, encoding Rho GTPase Activating Protein 11, implicated in the expansion of the neocortex through increased neurogenesis[22,24], 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 S5J). Given that ARHGAP11B impacts the abundance of basal progenitors, a cell type unique to the mammalian neocortex[147], zebrafish may not be suitable to characterize human-specific functions of this gene. This was also evident in *ARHGAP11B*-humanized zebrafish that exhibited similar molecular changes to *arhgap11* morphants (Figure 5E) suggesting antagonism by the human paralog, which is counter to previous studies[148] and possibly a spurious result due to the ectopic human mRNA expression.

Beyond modeling gene functions, our study also highlighted the considerable amount of genetic variation hiding within SD regions. Even with a complete T2T-CHM13, only 10% of SD98 regions are "accessible" to short reads[37] resulting in <10% sensitivity to detect variants and a depletion of GWAS hits (STAR Methods). Analysis of existing assemblies (HPRC and HGSVC) and cHiFi sequencing uncovered some of this hidden variation within 13 pHSD gene families. We note that for some highly identical duplicated genes (*CFC1*), 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[149,150]. Historically, signatures of balancing selection, which include an excess of mid-frequency alleles[151], have been difficult to detect within SDs due to assembly errors[37]. 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[152], 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[153]. SD genes may represent a hidden contributor to disease etiology. Our analysis identified 231 SD98 genes (110 human duplicate paralogs) co-expressed in modules enriched for autism genes (Table S2), including several within disease-associated genomic hotspots. Distinct SD mutational mechanisms, including ~60% higher mutation rate compared to unique regions[154] and interlocus gene conversion that can occur between paralogs[155,156], 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[40]. Human-duplicated gene families include ancestral paralogs *CORO1A*, *TLK2*, and *EIF4E*, with significant genetic associations with autism[68]. 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 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 but re-analysis of transcriptomes from non-brain cells/tissues is required. Meanwhile, duplicate genes, such as those encoding defensins[157–160], mucins[161,162], and amylases[42–44], 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[28]. Our use of a single human T2T-CHM13 haplotype of largely European ancestry[35] could miss some CN-polymorphic genes. As additional T2T genomes are released[29], it will be important to continue curating this list of duplications. Nevertheless, genes CN stratified by human ancestry can be identified using metrics such as V_{st}[163], as has been highlighted in other studies (reviewed here[9] and most recently[164]). 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/shinycn).

Limitations of the study

A notable limitation of our study is its reliance on existing gene annotations, which we used to group human duplicate paralogs into larger multigene families based on shared annotated sequences 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[165]. 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[166] and advancements in protein-prediction[167] and proteomic approaches[168] that will confirm whether or not these genes encode proteins. Similarly, single-cell transcriptomes typically focus on 3′-ends of transcripts, limiting specificity of human paralogs. Generation of single-cell long-read datasets[169] will enable more refined assessments of duplicate genes to discern differences in expression across cell types in the brain and other tissues. Further, for this study we have focused our analysis on gene duplications, but other complex structural variants have high propensity in altering functions and/or regulation of genes[28], with human-specific deletions[170,171] and inversions[172] the focus of recent studies. Finally, we highlight that our zebrafish studies employ transient, ectopic expression of human paralogs to "humanize" larvae and characterize phenotypes, which limited our analysis to early developmental traits (>4 dpf in zebrafish[141]), approximately equivalent to human mid- to late-fetal stages in brain development (Figure S3C), and could result in spurious phenotypes. Moving forward, generating stable transgenic zebrafish and mammalian models that better match endogenous cell/tissue expression of human paralogs will enable more precise delineation of gene functions.

In summary, we interrogated challenging regions of the genome by taking advantage of long-read sequencing in tandem with the new T2T-CHM13 reference genome and demonstrated a method using zebrafish to explore the functions of human-duplicated genes. Among our list of hundreds, we propose duplicate gene paralogs potentially contributing to unique features of the human brain, specifically featuring two: *GPR89B*, with a possible role in expansion of the neocortex, and *FRMPD2B*, with implications in altered synaptic signaling. In the future, additional genetic analyses across modern and archaic humans and experiments utilizing diverse model systems will reveal hidden roles of duplicated genes in human traits and disease.

Resource Availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by lead contact Megan Y. Dennis (mydennis@ucdavis.edu).

Materials availability

- Plasmids generated in this study have been deposited to Addgene.
- Zebrafish lines generated in this study *frmpd2*$^{tup\Delta 5}$ and *gpr89*$^{tup\Delta 8}$ are available from the lead contact upon request.

Data and code availability

- All raw sequencing data generated from this study, including cHiFi sequencing and scRNA-seq from zebrafish, have been deposited to ENA and NCBI (PRJEB82358). Accession numbers of data used from published sources can also be found in the Key Resources Table.

- All original code and processed data associated with this study is publicly available through https://github.com/mydennislab/HSD_brain_evolution and https://github.com/Ricardo-scb/ZebraFish-Diffusion-Model/ and has been deposited at Zenodo (DOIs 10.5281/zenodo.15486469 and 10.5281/zenodo.15485460).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Acknowledgements

We would like to thank the 1KGP, HGDP, T2T, SSC, gnomAD, HPRC, and HGSVC for access to their data and biospecimens. Thank you to Dr. Alfie Gleeson and Cassandra Olivas for support with evolutionary and transcriptome analyses, respectively, as well as Tonia Brown for copyediting. The Cell Biology and Human Anatomy Department at UC Davis provided significant imaging resources. 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.), as well as National Science Foundation CAREER awards (2145885 to M.Y.D., 1846559 to G.Q.). 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). NIH training grants through the National Institute of General Medical Sciences supported A.S. (T32GM007377), N.K.H. (T32GM153586), and N.A.F.M. (R25GM116690). Some figures were made with Biorender.

Author Contributions

D.C.S., J.M.U-S. and M.Y.D. conceived the project; D.C.S., J.M.U-S., G.K., A.S., and M.Y.D. generated sequencing data and analyzed datasets; D.C.S., A.S., A.M.A., and M.Y.D. performed genetic variation and population genetic analysis of human datasets; D.C.S. and Z.J. performed WGCNA analysis; R.V. and G.Q. performed supervised classification of zebrafish larvae; J.M.U-S, N.K.H., K.H., G.L. N.A.F.M., C.I., A.B., S.S., and M.Y.D. contributed to genotyping and phenotyping animal models associated with this study; T.N.T. generated the parCN dataset for the SSC cohort; E.D.G. provided human DNA samples; D.C.S., J.M.U-S., and M.Y.D. drafted the manuscript. All authors read and approved the manuscript.

Declaration of Interests

Authors have nothing to disclose.

Main Figures

Figure 1. Genetic analysis of human-duplicated genes. (**A**) Diagram of segmental duplications (SDs; blue) and subset with >98% identity (SD98; orange) in T2T-CHM13 autosomes, including total number of nucleotides (Mbp) and genes overlapping SD98 regions. (**B**) Copy number (CN) estimation methods, including gene-family CN (famCN) and paralog-specific CN (parCN). Horizontal lines represent short

reads mapping to unique (gray) and duplicated regions (orange and yellow). Heatmaps indicate CN estimates. (**C**) Pipeline for clustering and stratification of SD98 genes based on synteny with the chimpanzee reference 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 including gene models, centromeric satellites (CenSat), SDs (SegDup), and famCN and parCN predictions across sequenced individuals. (**E**) Distribution of Tajima's D values (y-axis) from 1KGP individuals of European (EUR) ancestry genome wide (gray) and SD98 (orange) across human autosomal chromosomes (x-axis). SD98 windows above the 95th (red line) or below the 5th (blue line) percentiles are considered outlier D values (STAR Methods). All human-duplicated gene names with outlier D values in at least one tested ancestry are labeled. Also see Table S1 and Figure S1.

Figure 2. Duplicated gene expression in the developing human brain. (**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), BrainSpan, and CORTECON. Protein-encoding genes are represented in darker shades. (**B**) Counts of expressed (dark orange) and non-expressed (light orange) human-duplicated genes across gene categories. (**C**) Human-duplicated gene expression in the CORTECON dataset stratified by copy number (CN). (**D**) Pipeline used for the weighted gene co-expression analysis (WGCNA). (**E**) The BrainSpan B-turquoise module, exhibiting an enrichment of human-duplicated genes (#) and autism-associated genes (*) plotted over developmental time (post-conception weeks, PCW) and bar colors representing brain regions (see **D**). Gene-ontology (GO) terms overrepresented among the co-expressed B-turquoise genes are depicted on the right. (**F**) Selected CORTECON WGCNA modules with enrichments (see **E**) and overrepresented GO terms indicated below. (**G**) CORTECON module assignment concordance scores are shown on the vertical axis for human-duplicated gene families. The size of each point corresponds to the number of members in the respective gene family. Also see Table S2, Figure S2, and Data S1.

Figure 3. Modeling functions of duplicated genes in brain development. Scaled TPMs from the human 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 with those highlighted with black bars prioritized for additional characterization. Also see Table S3 and Figure S3.

Figure 4. Genetic variation and signatures of selection of top candidate human-duplicated genes. (**A**) Number of likely gene-disruptive (LGD) (red), missense (blue), and synonymous (green) variants identified in pHSD genes. (**B**) Ka/Ks and (**C**) direction of selection (DoS) of pHSD genes with dashed lines indicating average genome-wide values between humans and chimpanzees (red) and neutrality (blue). Differences between matched ancestral and derived paralogs were tested with the Wilcoxon signed-rank test. Paralogs with infinite values or undetermined ancestral/derived state (hollow dots) were excluded from comparisons. (**D**) *CD8B* locus overview, including Tajima's D values derived from 1KGP genome-wide SNVs (top panel). Biallelic SNVs from the Human Pangenome Reference Consortium (HPRC) and the Human Genome Structural Variation Consortium (HGSVC) assemblies are shown with

with a minor allele frequency greater than 0.3 in individuals of African (AFR, n=27) and American (AMR, n=18) ancestry (middle panel) and used to calculate Tajima's D values (bottom panel). (**E**) Scaled transcript per million (TPM) expression of *CD8B* and *CD8B2* in postmortem brain tissue from BrainSpan. Also see Table S4 and Figure S4.

Figure 5. Duplicated gene functions modeled using zebrafish. (A) Functions of each pHSD gene were tested by generating knockout (KO, or morpholino) and 'humanized' models (injection of mRNA). **(B)** The F1 score, generated using a supervised convolutional neural network (CNN), is plotted indicating the effect size of morphological difference between models and controls, either using our batch-corrected images (blue bars) or original data (orange bars). Higher F1 score indicates greater difference. The bars for the control group indicate on average how distinct the controls are from all other groups. A threshold F1 score of 0.2 was used to define models being robustly classified as different from their control group. Pictured as a top inset are feature attribution plots with colors highlighting the region of the image used by the CNN to correctly classify and distinguish those genotypes from controls. **(C)** Measurements of selected pHSD gene families with heatmaps representing the percent change compared to the control group (asterisks indicating a Benjamini Hochberg-corrected *p*-value<0.05). **(D)** t-distributed stochastic neighbor embedding (tSNE) plot highlighting classified cell types from scRNA-seq data at 3 dpf. **(E)** Fold-change comparison between KO 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 (GO) enrichment results for the top overrepresented terms in upregulated genes in forebrain and midbrain across pHSD models, with gray indicating genes with no DEGs. Significant q-value>0.05 indicated with asterisk on color legend. Also see Table S5 and Figure S5.

Figure 6. Neurodevelopmental impact of *GPR89* and *FRMPD2*. (A) Head and brain area assessments at 3 dpf for G_0 crispants and stable knockout lines. *p*-values are indicated above box plots versus controls using an ANCOVA with a rank-transformation (humanized and crispant models) and Wilcoxon signed-rank tests (stable knockout lines). 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 across cell types. **(C)** Log_2 fold change (FC) of gene expression versus controls in cells from the telencephalon between knockout and humanized models. Red and blue colors correspond to DEGs discordant (*GPR89*) or concordant (*FRMPD2*) between the knockout and humanized models and their top representative gene ontology (GO) enrichment. **(D)** Forest plot with the results from the logistic regression for presence of progenitor versus differentiated states in forebrain cells. **(E)** Diagrams of the duplication event of *GPR89* with different expression patterns (**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) with high-speed events (HSE) defined as movement ≥28 mm/s. Colors represent the FC relative to the control group and the asterisk indicates a significant Dunn's test

(p<0.05 Benjamini-Hochberg-adjusted). **(G)** Diagram of the duplication event of *FRMPD2* (see also **E**), with a model of FRMPD2B antagonistic functions resulting in altered synaptic signaling depicted on the right. Also see Table S6.

STAR Methods

Experimental model and study participant details

Study participants

Sequenced genomes from human individuals were included from publicly available resources (1000 Genomes Project [1KGP] [55], Human Genome Diversity Project [HGDP] [173], Simons Genome Diversity Project [SGDP] [45], Genome in a Bottle [174], Human Pangenome Reference Consortium [HPRC)][29], and Human Genome Structural Variation Consortium [HGSVC] [84]) or through controlled access (Simons Simplex Collection [SSC] [60,61]) with known ancestries. Sex was not considered, since we focused on genetic variation across autosomes. This study was reviewed by the Institutional Review Board of the University of California, Davis and deemed minimal risk and human subjects exempt, with all participants de-identified.

Zebrafish procedures

Wild-type NHGRI-1[175], Tg[HuC-GFP] [121], and mutant (generated from this study) adult zebrafish were maintained in a modular system (Aquaneering, San Diego, CA) distributed in tanks at a maximum density of 10 adults per L and all fish were kept in temperature (28±0.5°C) and light (10h dark / 14h light cycle) controlled environment following standard protocols [176] with flowing water filtered via UV (Aquaneering, San Diego, CA). As described previously [105,109], all animals were monitored twice daily for health evaluations and feeding that 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 experiments, NHGRI-1 males and females of at least three months of age were randomly selected and placed in 1L breeding tanks in a 1:1 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 at a density of less than 100 embryos per dish until used. Embryos from all breeding crosses were pooled together and for all experiments embryos were collected at random. CRISPR-generated *frmpd2*$^{tup\Delta 5}$ and *gpr89*$^{tup\Delta 8}$ stable mutant F_1 zebrafish carrying heterozygous alleles were outcrossed twice to wild-type NHGR1-1 adults to remove potential off-target edits and then incrossed to generate batches of wild-type, heterozygous, and homozygous larval siblings for phenotypic assessments, following standard protocols [177]. All phenotypic screening was performed in zebrafish early larvae (up to 5 dpf). At this stage, animals are sexually indifferent. Sex determination in zebrafish is not well understood as it occurs in the absence of sex chromosome but it is generally accepted that sex determination establishes during the late larval stage of development (between approximately 15–25 dpf) [178]. All animal use was approved by the Institutional Animal Care and Use Committee from the Office of Animal Welfare Assurance, University of California, Davis.

Method details

Identification of SD98 genes

Duplicated regions were extracted from previously annotated SDs [36] using T2T-CHM13 (v1.0) coordinates and subsequently merged using BEDTools merge [179]. 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) [35]. 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 [179]. 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 coordinates were lifted over to the T2T-CHM13v2.0 assembly using UCSC liftOver tool [180] and the chain file for v1.0 to v1.1 conversion, obtained from the T2T-CHM13 reference GitHub repository (https://github.com/marbl/CHM13).

Gene family clustering

SD98 genes were grouped into gene families based on shared exons (Table S1C). 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) [181] 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 [36] 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 [36].

Human-duplicated gene 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) [36], 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 [36] 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 [47] for 1KGP 30× high-coverage Illumina individuals [55] and four archaic genomes (including Altai Neanderthal [PRJEB1265] [48], Vindija Neanderthal [PRJEB21157] [49], Mezmaiskaya Neanderthal [PRJEB1757] [48,49], and Denisova [PRJEB3092] [50]), using T2T-CHM13 (v1.0) as reference [35]. 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.

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) [182] and loss-of-function observed/expected upper fraction (LOEUF) [53]. 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.

Assessment of variant-calling performance

Despite improved representation of duplicated genes in T2T-CHM13, genomic assessment of these regions remains challenging using short-read Illumina data. We assessed if SNVs were depleted across duplicated regions using variants from 1KGP individuals mapped to T2T-CHM13 (v1.0) [37] , filtering for biallelic SNPs only, using BCFtools view [183] using parameters --exclude-types indels and --max-alleles 2. We compared observed values to empirical distributions, obtained by randomly sampling regions of identical size as SD and SD98 regions using bedtools shuffle with -noOverlapping -maxTries 10000 -f 0.1 parameters. Previously published centromeric satellites coordinates [184] were also excluded using the flag -excl. We found that duplicated regions are significantly depleted for SNVs in the high-coverage 1KGP dataset compared to unique regions, which do not include SD or centromeric satellites [184] (SD98: 11.79; SD: 25.9, unique: 37.49 SNVs/kbp; p-value<0.05 empirical distribution) (Figure S1D). The autosomal 2.4 Gbp in T2T-CHM13 accessible for accurate Illumina SNV calling—determined using read depth, mapping quality, and base quality metrics [37]—includes only 37.95% and 10.86% of SD and SD98, respectively, while 95.64% of unique space is accessible (Figure S1D). In the SD98 regions, only 56 previously-identified SD98 genes, including 48 protein coding and 8 pseudogenes, are accessible (>90%) to short-reads (Table S1A).

To evaluate our ability to detect variants within duplications using short-read sequencing, we compared SNVs discovered using Illumina short-read versus PacBio HiFi long-read data across eight 1KGP

individuals included in both the 1KGP and HPRC [37,81] (HG01109, HG01243, HG02055, HG02080, HG02145, HG02723, HG03098, and HG03492). Biallelic SNVs were selected using BCFtools view. Concordance between platforms, measured as precision and sensitivity, was obtained with rtg-tools vcfeval [185] for autosomal Non-SDs, SDs, and SD98 regions, using PacBio HiFi variants as a truth-set. Short-read accessible regions were obtained from Aganezov et al. [37]

While no differences in variant density (SNV sites within 1-kbp non-overlapping windows) existed between technologies in non-duplicated regions, we observed reduced mean variant density from short-read versus long-read data across SD (SRS: 1; LRS: 5) and SD98 (SRS: 0; LRS: 5) (Figure S1E). Notably, no differences were observed when considering only short-read accessible regions [37]. Using cHiFi-discovered variants as truth, we next assessed variant calling precision and found that 99.5% of SNVs matched between technologies in non-SD, but decreased to 88.6% and 81.7 % in SD and SD98, respectively (Figure S1F). When considering only short-read accessible regions, SNV precision increased in the three regions assayed to 99.7%, 96.1%, and 94.2% for non-SD, SD, and SD98. Sensitivity—measured as the proportion of HiFi-discovered SNVs also detected using Illumina data—experienced a pronounced decrease of 24.5% in SD and 0.85% in SD98 compared to 87.6% in Non-SD regions. When considering only short-read accessible regions, however, sensitivity is improved to 72.5%, and 57.8% in SD and SD98, respectively. Overall, these results indicate that existing variants identified across duplicated regions from Illumina data are generally accurate, particularly in defined accessible regions, but not comprehensive.

Tajima's D analysis

Additionally, Tajima's D [56] 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) [37]. For each 1KGP continental superpopulation (African, European, East Asian, South Asian, and American), we computed Tajima's D in 25-kbp windows (≥5 SNPs) using VCFtools [186], restricting analyses to short-read–accessible windows (≥50% of bases annotated as accessible in the combined accessibility mask) [37]. Duplicated and non-duplicated genomic loci differ in several ways, including constraint [187] and mutation rates [154]. To mitigate the effects of these differences, outlier D values were defined as the lower 5th and upper 95th percentiles within accessible SD98 windows (i.e., windows overlapping at least 10% of their bases with SD98 regions and meeting the short-read accessibility criterion). Outlier lower and upper 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

Due to difficulties mapping short reads to highly identical regions, as well as lack of SD representation on SNP arrays, variants across SD98 genes and regions are depleted in existing genome-wide studies of phenotypes and diseases. To quantify this underrepresentation, we considered the GWAS catalog v1.0 [188] (mapped to GRCh38.p12), ClinVar [189] (rel. 20200310), and GTEx [190] v8 single-tissue eQTL (dbGaP Accession phs000424.v8.p2; mapped to GRCh38, excluding chromosome Y) 12. From the GWAS catalog, we selected only SNPs significantly associated with brain measurements (p-value < 0.05) and identified GWAS "mapped genes" that overlapped with our SD98 gene list using gene symbols. We

observed significant depletion across the GWAS catalog (SD98: 0.29 variants/100kbp; genome-wide: 1.5 variants/100 kbp), ClinVar (SD98: 20.81 variants/100 kbp; genome-wide: 9.95 variants/100 kbp), and GTEx expression quantitative trait loci (eQTL) databases (SD98: 398.7 variants/100 kbp; genome-wide: 70.14 variants/100 kbp) (Figure S1J).

Additionally, we downloaded published associations between CNVs and neural traits in the UKBB [58]. 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 [180]. 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[179].

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 [191] 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 analysis

Previously published brain transcriptomic datasets, including post-mortem tissue and cell lines, were obtained. These datasets included neocortical germinal zones [62], neural stem and progenitor cells [22], a neuroblastoma cell line SHSY5Y [63], and two longitudinal studies of *in vitro* induced neurogenesis from human embryonic stem cells [64] (CORTECON), and post-mortem brain (BrainSpan) [65]—the latter of which was separated into prenatal and postnatal samples. Transcriptomic data from lymphoblastoid cell lines from 69 individuals were also included for comparison [66]. 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 [192] with the flags "--validateMappings --gcBias". The CAT/Liftoff transcriptome was converted to fasta format using gffread [193]. Transcripts per million (TPM) values and raw counts were summed to the gene level using tximport [194]. 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.

Weighted gene co-expression analysis

WGCNA was performed using the R package [67] with two longitudinal brain datasets: prenatal BrainSpan [65] and CORTECON datasets [64]. First, we modeled gene expression across the developmental brain using BrainSpan samples available in NCBI BioProject PRJNA242448. The BrainSpan dataset includes 607 samples from 25 brain regions, spanning early prenatal to adulthood developmental epochs. We filtered for high-quality samples used in a previous gene co-expression analysis [195] and focused exclusively on prenatal samples from the frontal cortex, including the dorsolateral prefrontal cortex (DFC; n = 12), ventrolateral prefrontal cortex (VFC; n = 1), medial prefrontal cortex (MFC; n = 12), and the orbital prefrontal cortex (OFC; *n* = 12) (Data S1). Overall, our dataset included 47 samples from four brain regions, spanning three developmental epochs—early prenatal, early-mid prenatal, and late-mid prenatal—covering post-conception weeks (PCW) 8 to 22. We selected genes expressed (≥1 TPM) in at least 80% of the samples for each developmental epoch and brain region, resulting in 17,388 genes which were used as input for WGCNA analysis. We used Principal Component Analysis (prcomp function in R) to cluster samples based on their expression profiles, and no outliers were identified (Data S1). However, the first two principal components distinctly separated the early prenatal samples at 8 weeks post-conception (8 PCW) from the rest, indicating that post-conception age exerts a stronger influence on gene expression than the broader developmental epoch.

To overcome coverage and batch effects, we performed variance stabilizing transformation on the raw counts using function varianceStabilizingTransformation() from the R package DESeq2 with a blind design to preserve variability within each developmental epoch. We use the function pickSoftThreshold() from the R package WGCNA to estimate the power parameter that generates a scale-free topology network, choosing the minimum value where the scale-free topology fit index is around 0.8, which in the BrainSpan dataset was 24. Normalized counts were used as input to construct a signed network using function blockwiseModules() with parameters networkType="signed", deepSplit=4, detectCutHeight=0.995, minModuleSize=30, mergeCutHeight=0.25, and softPower=24, and default parameters otherwise. This yielded 23 modules represented by their respective eigengenes (Data S1) named "B-" followed by a random color, ranging from 47 to 3,126 genes each, with a median module size of 428. The largest modules were B-turquoise (n=3,126), B-blue (n=2,727), B-brown (n=2,102), and B-yellow (n=1,802). Despite the larger numbers of genes assigned with these clusters, the median module memberships remained high (B-turquoise: 0.72, B-blue: 0.74, B-brown: 0.73, B-yellow: 0.72). The B-turquoise and B-yellow were negatively correlated with early prenatal stages, while module B-blue and B-brown were correlated with prenatal stages. Seventeen modules included paralogs from human-specific gene families. In the BrainSpan dataset, 2,320 genes remained unclustered and were assigned to module B-Grey, reflecting the expression noise inherent in complex tissues composed of multiple cell types.

We then modeled gene expression using the CORTECON dataset, which tracks *ex-vivo*-induced neurogenesis from human embryonic stem cells. This dataset includes 23 transcriptome samples (our analysis excluded the test set), spanning the stages of pluripotency, neural differentiation, cortical specification, deep-layer formation, and upper-layer formation (Data S1. We clustered samples based on their raw expression counts using hierarchical clustering and Principal Component Analysis (prcomp function in R), which flagged two samples as outliers (SRR1238515 and SRR1238516) and therefore they were removed from downstream analyses. To reduce noise, we removed gene features with consistently

low counts, meaning less than 10 counts across 90% of the CORTECON samples, resulting in 15,697 gene features. We transformed raw counts using the varianceStabilizingTransformation() function from the DESeq2 package with a non-blind design that aimed to remove differences from developmental stages. To determine the optimal softPower parameter for the CORTECON dataset, we used the pickSoftThreshold() function from the WGCNA package, which identified a value of 24.

Co-expressed modules were obtained with the function blockwiseModules() from the WGCNA package, using networkType="signed", deepSplit=4, detectCutHeight = 0.995, minModuleSize=30, mergeCutHeight = 0.15, and softPower=24, and default parameters otherwise. This approach yielded 37 co-expressed modules (Data S1), represented by their respective eigengenes and named as "C-" followed by a random color. The number of genes assigned to a module ranged from 33 to 3,988, with a median value of 424.1 genes per module. The most genes were assigned to modules C-turquoise (n=3,988), C-blue (n=2433), C-yellow (n=1,149), C-green (n=954), and C-red (n=709). The median module memberships for these larger modules were higher than the BrainSpan dataset, likely reflecting reduced variability in the *in vitro* neuronal samples versus the more heterogeneous post-mortem brain tissue (C-turquoise: 0.81, C-blue: 0.84, C-yellow: 0.85, C-green: 0.86, and C-red: 0.83). These modules showed the strongest associations with developmental stages, where C-turquoise and C-green were strongly associated with pluripotency (Pearson r = 0.87 and 0.97, respectively), C-yellow and C-blue were strongly anti-correlated with pluripotency (Pearson r = -0.98 and -0.9, respectively), and C-red was correlated with neural differentiation and anti-correlated with upper layer formation (Pearson r = 0.67 and -0.56). Twenty-one co-expression modules included paralogs from human-specific gene families. Importantly, 871 genes were assigned to the C-grey module, which corresponds to unclustered genes (Data S1). To verify CORTECON module assignments, we also assessed human-duplicated genes with known functions. *ARHGAP11B*, which induces cortical neural progenitor amplification by altering glutaminolysis in the mitochondria [24], is a member of the C-turquoise module. 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. Additionally, the hominoid-specific gene *TBC1D3*, known to promote basal progenitor amplification in the outer radial glia resulting in cortical folding in mice [25] is a member of the C-purple module, which is associated with regulation of neural differentiation.

Module concordance was calculated for each gene family as the proportion of its members assigned to the same module, defined as the maximum number of co-assigned members divided by the total number of members in the family. A concordance score of 1 indicates that all members were assigned to the same module, while a score of 0 indicates that no members shared a module assignment. 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. GO terms enrichment analysis was performed using the R package clusterProfiler ego function, using an adjusted *p*-value threshold of 0.05 [196]. Enrichment of gene categories were performed using the hypergeometric test in R for autism

genes [68], expanded genomic hotspots [191], and cell markers [195], as well as for SD98 genes and human duplicated genes.

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) [75] and zebrafish (GSE158142) [76], calculating Z-score normalized TPM values. Matching of developmental stages across human, mouse, and zebrafish was done as previously described [77]. 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 sequencing

We performed cHiFi sequencing of 172 individuals from the 1KGP, two trios from Genome in a Bottle [174], and 22 HGDP individuals with available linked-read data via the 10X Genomics platform [173], totaling 200 samples and 18 family trios (Table S4B). 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 [197]. 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 S4D). 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 capture sequencing protocol included tiled baits across all duplicated regions of interest and only exons in non-duplicated space (Figure S4B). As a result, unique exons exhibited significantly lower coverage compared to duplicated exons (Mann-Whitney U test, p-value=2.2×10^{-16}). Importantly, we did not observe significant differences in coverage between ancestral and derived paralogs, despite the baits being designed based on the ancestral sequence (Mann-Whitney U test, p-value>0.05). cHiFi coverage across regions of interest was calculated using samtools depth [183] with --min-MQ 10. Globally, considering a cutoff MAPQ score greater than 10, we achieved ~3 kbp reads with an average coverage of 27× within regions of interest (Table S4E, Figure S4C). We also assessed for the occurrence of PCR duplicates given that they pose three problems: 1) the true output of diverse representation of reads that are sequenced is reduced, 2) lead to false positive variant calls skewing allele frequencies, and 3) may introduce erroneous mutations that do not reflect true population variants. We found 66% of sequenced reads to be unique genome-wide, and within the intended capture space, 34% of the total unique reads mapped to the regions of interest.

Long-read genetic variation

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 [84] 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 [198] pileup with parameters -q 0 -evcf and converted into diploid calls using dipcall-aux.js [199] 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.

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. For each sample, read groups were added with Picard AddOrReplaceReadGroups and variants were called on each sample using GATK HaplotypeCaller [200], 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 genotyping confidence thresholds of 0, 10, 20 and 30, and variants were subsequently filtered using hard-filtering thresholds for both genotyping quality (0, 20, 50, 70) and depth (0, 4, 8, 12, 16).

We optimized variant genotyping and hard-filtering parameters by benchmarking minimum thresholds using both population and trio-based analysis, focusing exclusively on biallelic SNPs selected with bcftools view --max-alleles 2 and bcftools view --exclude-types indels. Specifically, we assessed deviations from Hardy-Weinberg equilibrium calculating inbreeding coefficients from the founder population (excluding offspring), with an inbreeding coefficient below -0.3 considered indicative of excess heterozygosity. Additionally, we evaluated Mendelian concordance within trios, calculated for each threshold combination using rtg mendelian [185], excluding trios where any of the members had a missing genotype with bcftools view -i 'F_MISSING=0'. Total number of variant sites were obtained with BCFtools stats.

As MAPQ and genotyping confidence thresholds became more stringent, the total number of variant sites decreased, while biological metrics improved, including increased Mendelian concordance and reduced excess heterozygosity (Figure S4D,E). A minimum MAPQ threshold of 20 reduced sites with excess heterozygosity and improved Mendelian concordance across all genotyping confidence levels, while only marginally reducing the number of detected variants. Therefore, we conservatively proceeded with a minimum MAPQ of 20 and a minimum genotyping confidence threshold of 30. We next optimized hard-filtering parameters, observing that Mendelian concordance increased significantly with higher read-depth and genotype-quality thresholds. We achieved near 100% concordance using either genotype quality of 50 (at any read depth) or the combination of read depth 8 with genotype quality 20. Since the latter combination provided similar performance while retaining more variants, we selected these as our hard-filtering parameters, resulting in the identification of 28,476 biallelic SNVs across 200 individuals. For downstream population genetics analyses, we retained the 144 unrelated individuals that passed our quality thresholds and merged their cHiFi variants with variants from 56 non-redundant HPRC/HGSVC individuals using BCFtools merge, creating a unified cohort of 200 genomes. Functional consequences in the combined dataset were annotated with the Ensembl Variant Effect Predictor (VEP).

Haplotype networks for *CD8B* were constructed using HPRC/HGSVC continuous haplotypes extracted with BEDtools getfasta and aligned with Muscle using Mega Software [201]. Networks were generated using a minimum spanning tree with the software PopArt [202].

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) [88]. 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 [89]) from generated multiple-species alignments for each gene family

(MAFFT [203]), 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 [14]. 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 [204] 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

Creation of CRISPR lines to knockout genes of interest was done as previously described [105,109,205]. 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 S5A. 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 S5A). Knockout alleles in stable lines corresponded to a 5-bp deletion in *frmpd2* (named $frmpd2^{tup\Delta 5}$) and an 8-bp deletion in *gpr89* (named $gpr89^{tup\Delta 8}$, allele sequences can be found in Table S6E). 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-cell-stage embryos. Assessments of potential off-target sites for all sgRNAs used in this study were performed with the CIRCLE-seq protocol [206,207] and top potential off-target sites were evaluated via Sanger sequencing as previously described [105]. 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*), pCR-TOPO (*NPY4R* and *FAM72B*), and pCMV-SPORT6 (*FRMPD2B*). The cDNA inserts of two genes were synthesized (Twist Biosciences, San Francisco, CA) based on transcript evidence from IsoSeq data from the ENCODE [208] project (*PDZK1P1*: ENCFF158KCA, ENCFF939EUU; *PTPN20CP*: ENCFF305AFY). All plasmids were sequenced through either Azenta or Plasmidsaurus. 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/µl concentration and ~1 nl of the mix microinjected into one-cell stage embryos, as described above. Presence of the human mRNA transcripts was observed by using 500 ng of extracted RNA from 3 dpf-injected larvae used in the sciRNA-seq experiment with the SuperScript IV Reverse Transcriptase kit (Thermo Fisher, Waltham, MA) followed by PCR amplification with DreamTaq PCR Master Mix (Thermo Fisher, Waltham, MA) and primers listed in Table S5A.

Morphometric assessments

High-throughput imaging of the zebrafish larvae was performed using the VAST BioImager system (Union Biometrica, Holliston, MA) as previously described [108,109]. 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 [107] 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] [121] 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 [209] 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 was not included.

Supervised classification of images

As an alternative to performing statistical tests to identify changes in predefined morphological measurements between mutants and controls, we employed a convolutional neural network (CNN) to identify differences between mutants and controls without the need to measure predefined features [110]. 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 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 first transforming the original image x into a latent representation z_0 through a variational autoencoder function ε that reduces the dimensionality of x but does not remove any batch effect:

$$z_0 = \varepsilon(x)$$

We then pass the encoded image z_0 through an LDM forward process, in which we repeatedly sample new latent variables $z_1,\ldots,z_T$ by effectively repeatedly adding Gaussian noise to the original image z_0:

$$q(z_t|z_{t-1}) = N\left(z_t; z_{t-1}\sqrt{1-\beta_t}, \beta_t \boldsymbol{I}\right)$$

Where β_t is the magnitude of noise decided by a noise scheduler at time step t. This ultimately transforms the original (VAE-transformed) image 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 by successively transforming 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:

$$p(z_{t-1}|z_t) = N\big(z_{t-1}; \mu_\theta(z_t, t, c), \Sigma_\theta(z_t, t, c)\big)$$

Where $\mu_\theta(z_t,t,c)$ and $\Sigma_\theta(z_t,t,c)$ represents the predicted mean and covariance functions. 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 to minimize negative log likelihood using 350 diffusion steps with a linear noise scheduler [210,211]. 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 [212]. 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 a five-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 5B), 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. All code related to this analysis is available [213].

Single-cell RNA-seq

We performed cellular assessments using the single-cell combinatorial indexing RNA sequencing (sciRNA-seq) protocol [113]. 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 [214] 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 in 5-µl increments, and then slowly frozen 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 [113], 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 [215] (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 [216] and fastqc [217]. Reads were then mapped to the improved zebrafish transcriptome [218] with STAR [219] 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 split by their UMI sequences (corresponding to individual cells) and gene-cell count matrices constructed by mapping reads to the zebrafish v4 GTF file [218].

Gene-cell count matrices were loaded into R to generate Seurat v4 [220] 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 [215,221] and estimated using DoubletFinder [222]) were removed (Figure S5E). Cells from different libraries were normalized using SCTransform [223] with the glmGamPoi method and regressing by the percentage of mitochondrial and ribosomal counts. Then, normalized counts across sequencing libraries were integrated with Harmony [224] 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 [76,119] and the Zebrafish Information Network (ZFIN [98]) website. Focusing on neuronal, glial, and eye-related clusters left a total of 95,555 cells for further analysis (Table S5F). 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 [225] 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.

Seizure susceptibility

To assess changes to chemically induced seizure susceptibility, we employed an optimized published protocol [128]. 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 [128] to compare against batch-sibling controls.

Quantification and statistical analysis

Gene ontology overrepresentation

Overrepresentation of gene ontology terms across human duplicated genes and copy-number constrained genes was assessed with clusterProfiler [196], using all human-genes as background and considering terms with Benjamini-Hochberg adjusted p-value ≤ 0.05 as significant. Enrichment of DEGs in GO terms for the zebrafish pseudo-bulk analysis was also estimated with clusterProfiler [196] using only the expressed genes as the background list for the tests and p-values below 0.05 after a Benjamini-Hochberg adjustment were considered as significant.

Genetic association

To assess depletion levels of associated variants from published databases, observed values were compared to empirical null distributions, built from 10,000 non-overlapping, size-matched random regions generated using bedtools shuffle -noOverlapping -maxTries 10000 -f 0.1. One-tailed empirical p-values were calculated as: $p = (M + 1) / (N + 1)$, where M is the number of iterations yielding a number of features less than (depletion) the observed value and N is the number of iterations. Empirical p-values were calculated using 10,000 permutations. Significant differences between CN across probands and unaffected siblings were calculated using a paired design for each sibling pair, using a Wilcoxon signed-rank test corrected for multiple testing using false discovery rate, considering q-values ≤ 0.05 as significant.

Brain expression

Significant differences in expression across developmental stages of the CORTECON datasets stratified by copy-number status (polymorphic, nearly fixed and fixed) were obtained using a Mann-Whitney U test. Significance thresholds were defined as: ns: non-significant; * ≤ 0.05, ** ≤ 0.01 and *** ≤ 0.001. Enrichment of gene categories (human duplicated, SD98, autism, human hotspots genes) across BrainSpan and CORTECON modules was assessed with a Hypergeometric test (phyper function in R) using *p*-value ≤ 0.05 as significance cut-off. Gene ontology overrepresentaton analysis across BrainSpan and CORTECON modules was performed with clusterProfiler [196], employing all human genes as the background and considering terms with a Benjamini-Hochberg *p*-value ≤ 0.05 as significant.

Signatures of natural selection

Outlier genome-wide Tajima's D values were obtained as the 5th and 95th percentiles across each 1KGP continental population. Differences in heterozygous sites distribution between paralogs across pHSD families were calculated using a Mann-Whitney U test, considering *p*-value ≤ 0.05 as significant. Differences in allele frequency distributions between 35-kbp windows overlapping *CD8B* and *CD8B2* loci were obtained using Kolmogorov-Smirnov test. Global differences in Ka/Ks and Direction of Selection values between ancestral and derived pHSD genes were obtained using a paired-design between ancestral and derived paralogs, using a Wilcoxon signed-rank test. For gene families with more than two paralogs, all derived paralogs were compared to the ancestral one.

Single-cell transcriptomics

To define cellular identities of each obtained cluster, we extracted DEGs between clusters using the MAST test with a Bonferroni correction. For the cell-type differential expression between genotypes, we balanced the number of cells between groups (maintaining the number of cells from the smaller group) and gathered gene counts to obtain pseudo-bulk values that we then compared using the Wald test with a Benjamini-Hochberg adjustment. Fold-change correlations between DEGs across groups was assessed using the spearman method. Evaluation of neuronal classification in progenitor-like or mature was done using a generalized linear model after each neuronal cell was assessed for the presence of counts for the defined progenitor or mature neuronal markers described above. In all analyses, adjusted *p*-values below 0.05 were classified as significant.

Morphometrics

For each morphometric measurement, values larger than the 75% quantile plus two times the interquartile range or smaller than the 25% quantile minus two times the interquartile range were considered as technical outliers and removed from the distribution. Normality of each morphometric value was assessed using a qq-plot and a Shapiro-Wilk test. Due to non-normality of residuals, an ANCOVA with a rank-transformation for each measurement was used with the date of the experiment and the plate as covariates. To control for potential confounding technical biases, for each mutant evaluated, we compared all larvae to the controls obtained in the same experimental batch. *p*-values from paired tests between mutants and controls were adjusted using the Benjamini-Hochberg method and significance was determined as adjusted *p*-values below 0.05. All numbers of larvae included in each comparison can be found in Table S5E, as well as each raw and adjusted *p*-values, the mean values for each group, standard deviation, and delta. For the stable zebrafish lines, direct comparisons between mutants and controls were

performed using the Wilcoxon test due to non-normality of residuals and no additional covariables were included given that all values were obtained in the same experiment. No multiple testing adjustment was added to these Wilcoxon rank-test *p*-values. The number of larvae included in each group for these tests can be found in Figure 6 at their corresponding boxplots, which include the median and quantile values. The values above each box is the *p*-value from the Wilcoxon tests against controls.

Seizure susceptibility

Frequencies of HSE between control and mutant groups was performed using a Kruskal-Wallis test with a Dunn's post-hoc test given that values were non-normally distributed following an assessment via qq-plot and Shapiro-Wilk tests. The number of larvae included in the reported groups in Figure 6F for the control treatment (0 mM PTZ) were 666 controls, 54 *frmpd2* knockouts and 65 *FRMPD2B*-injected, while the 2.5 mM PTZ treatment included 693 controls, 84 *frmpd2* knockout and 71 *FRMPD2B*-injected larvae. Tests with a *p*-value below 0.05 are highlighted in the heatmap with an asterisk.

Additional resources

- parCN estimates for 1KGP individuals available at https://dcsoto.shinyapps.io/shinycn.
- Gene expression from human brain datasets available at https://dcsoto.shinyapps.io/shinybrain/.

Supplemental Information

Figure S1. Detailed genetic analysis of human duplicated genes related to Figure 1 and STAR Methods. (A) Pipeline to group SD98 genes into gene families. (**B**) Distribution of number of gene members within duplicate gene families. (**C**) gnomAD pLI versus LOEUF scores for all SD98 genes with available scores. (**D**) From top to bottom: 1KGP short-read SNVs [37] in SD (blue, left) and SD98 (orange, right) using the T2T-CHM13 regions. Observed values are shown as vertical bars, while empirical distributions of the number of variants observed in randomly sampled regions are represented as density plots. Total region size (in Gbp) and accessible sites size (darker colors), for NonSD (gray), SD (blue), and SD98 (orange). **(E)** Distribution of biallelic SNVs across non-overlapping 1-kbp windows across Non-SD (gray), SD (blue), and SD98 (orange), discovered with short-read sequencing (SRS, left) and long-read sequencing (LRS, right) technologies. Number at the bottom represents the total number of 1-kbp windows defined for each region. **(F)** Assessment of precision and recall across eight individuals sequenced with Illumina short-read sequencing and PacBio long-read sequencing reads, for all regions (left) and only accessible sites (right). (**G)** Percentage of short-read accessible bases versus percentage of bases within SD98 regions for 25-kbp windows genome-wide used in Tajima's D calculations. **(H)** Distribution of Tajima's D values calculated using 1KGP SNPs from individuals of African ancestry across 25-kbp overlapping protein-coding genes (green), unprocessed pseudogenes (purple), other genes (blue) and no genes (red), in non-duplicated (nonSD) and SD98 regions. *p*-values were calculated using a Mann-Whitney U test. ns: non-significant; * ≤ 0.05, **** ≤ 0.0001. **(I)** Tajima's D values from individuals of the 1000 Genomes Project were calculated across 25-kbp windows genome-wide (gray) and in SD98 region (orange) divided per superpopulation. Only outlier values in the upper 95th percentile or bottom 5th percentile are shown, plotted across human autosomal chromosomes on the x-axis. Human duplicated genes within windows with outlier D values are highlighted. Ancestries depicted include

African (AFR), East Asian (EAS), South Asian (SAS), and American (AMR). **(J)** Assessment of variant association depletion in SD and SD98 regions in short-read-based databases. Included databases: GWAS catalog, ClinVar, and GTEx eQTL. Observed variation is represented in vertical lines for SD (blue) and SD98 (orange) regions, and density plots represent empirical distribution of randomly sampled sites of the same size as SD or SD98 regions. **(K)** Human duplicated genes with significant copy-number differences between autistic probands and unaffected siblings from the Simons Simplex Collection. Significant differences were obtained using a Wilcoxon signed-rank test FDR-adjusted q-value < 0.05.

Figure S2. Human brain expression of duplicated genes related to Figure 2. (A) Intersection between human duplicated genes expressed (TPM≥1) across prenatal datasets. **(B)** Gene expression across human-duplicated gene subsets in $\log_2$(TPM) in the CORTECON dataset, spanning pluripotency to upper layer formation, and lymphoblastoid cell line data (n=69) [66], stratified by copy number (CN) category. **(C, D)** Module eigengenes from weighted gene co-expression network analysis (WGCNA) of prenatal BrainSpan samples from the prefrontal cortex **(C)** and CORTECON **(D)**. Each module is represented by a randomly assigned color stated above each plot. Numbers in parentheses represent the total number of genes assigned to the module. Stars represent modules enriched on different gene categories, including gene ontology (GO) terms (red), SD98 genes (light blue), human-duplicated genes (dark blue), autism-associated (ASD) genes (yellow), and genomic hotspots from Sattertrom et al. [191] (green). Colored bars at the bottom indicate different ages in post-conception weeks (PCW) for BrainSpan and different developmental stages of *ex vivo* neurogenesis for CORTECON. **(E)** Network diagram of the C-yellow module. Only genes within human-duplicated gene families (red), SD98 (pink) and autism-associated (yellow) categories with high module membership are depicted. Genes with asterisks are non-syntenic with the chimpanzee reference (PanTro6) and bold borders are within ±500-kbp of a genomic disorder hotspot.

Figure S3. Matched neurodevelopment staging of human, mouse, and zebrafish related to Figure 3 and Discussion. Depicted are principal component analyses of brain single-cell RNA-sequencing samples from **(A)** mouse [75] and **(B)** zebrafish [76], and (**C**) matched mouse and zebrafish samples to human developmental stages from the BrainSpan dataset.

Figure S4. Detailed genetic analysis of priority human-duplicated genes related to Figure 4 and STAR Methods. (A) Sequenced samples used for pHSD variant analysis from **(A)** draft human diploid assemblies included in the Human Pangenome Reference Consortium (HPRC, n=47) and Human Genome Structural Variation Consortium (HGSVC, n=9) and **(B)** capture strategy followed by PacBio HiFi long-read sequencing (cHiFi) from the 1000 Genomes Project (1KGP) and Human Pangenome Reference Consortium (n=200; n=144 unrelated). World maps represent sample sites for each ancestry with counts depicted. **(C)** Benchmarking cHiFi sequencing variants by comparing sequencing coverage between derived and ancestral paralogs (left), unique exons and duplicated exons (middle), and tiled versus untiled regions (right). **(C, D)** Impact of **(C)** mapping quality (MAPQ) and genotyping confidence thresholds, and **(D)** per-sample genotype quality and minimum read depth thresholds on the total number of variant sites (left), variant sites with excess heterozygosity (middle), and median Mendelian concordance across 18 trios (right) using cHiFi reads. **(F)** Heterozygous-site densities across duplicated portions of pHSD captured loci. Variants were identified for HPRC and HGSVC samples (top; n=56) and non-redundant unrelated cHiFi individuals (bottom; n=144). Ancestries depicted include African (AFR), European

(EUR), East Asian (EAS), South Asian (SAS), and American (AMR). **(G)** The human genetic variation landscape across *SRGAP2C* locus with 1KGP genome-wide outlier Tajima's D value (shaded region) as well as and Tajima's D plots derived from HPRC/HGSVC assembly-derived SNVs using 6-kbp windows and 500-bp steps. **(H, I, J)** Assembly-derived SNVs were also used to characterize nucleotide diversity π (top) and Tajima's D (bottom) across corresponding duplicated exons, calculated in 15-kbp sliding windows with 1-kbp steps, for human duplicated gene paralogs **(H)** *GPR89*, **(I)** *ROCK1*, and **(J)** *FAM72*. **(K)** Human genetic variation landscape of the *CD8B* locus in T2T-CHM13v1.0 reference genome in the UCSC browser, with HPRC/HGSVC intermediate allele frequency variants, and derived Tajima's D values calculated in 6-kbp windows with 500-bp steps. Haplotype networks for all HPRC/HGSCV continuous haplotypes in addition to chimpanzee (panTro6) are plotted for each highlighted region, encompassing 6-kbp of sequence. **(L)** Folded Site frequency spectrum with minor allele frequency (MAF) calculated across 35-kbp regions overlapping *CD8B* (light gray) and *CD8B2* (dark gray) from variants detected in the combined dataset including long-read assemblies and capture PacBio HiFi sequencing from individuals of AFR ancestry (n=88 individuals), EUR ancestry (n=29 individuals), and AMR ancestry (n=18 individuals). Three individuals of EUR ancestry (NA20582, NA20525, NA20542) were excluded from this analysis. *p*-values were obtained comparing *CD8B* MAF distribution between populations using Kolmogorov-Smirnov test.

Figure S5. Analysis of human duplicated priority genes using zebrafish related to Figure 5 and STAR Methods. (A–C) Endogenous gene expression of pHSD zebrafish orthologs during development. **(A)** Temporal expression between 0 and 120 hours post-fertilization using published data [106] of the zebrafish orthologs of the selected pHSDs. Shaded area corresponds to the brain development period in zebrafish embryos [111] of the zebrafish orthologs of the selected pHSDs. **(B)** Expression of the selected genes in embryonic or adult tissues (data from [97]). **(C)** Available expression patterns via *in situ* hybridization in the Zebrafish Information Network (ZFIN) [98]. **(D)** Detection of human mRNA post-injection in 'humanized' zebrafish models using RT-PCR of RNA extracted from 3 dpf injected "humanized" larvae (denoted by white stars) and controls (denoted by black stars) with primers targeting eight human-specific mRNAs (*SRGAP2C*, *ARHGAP11B*, *GPR89B*, *PDZK1P1*, *PTPN20CP*, *NPY4R*, *FAM72B*, *FRMPD2B*). **(E)** Description of the number of cells (n) per zebrafish mutant model used for single-cell transcriptomic analysis.

Table S1. Genetic variants analysis of SD98 genes related to Figure 1, Discussion, and STAR Methods. (A) SD-98 genes (>1 exon overlapping segmental duplications with over 98% identity) in T2T-CHM13v1.0, database intersections, and brain RNA-seq expression (TPM: transcripts per million). **(B)** SD-98 genes in chromosomes T2T-CHM13 (v1.0) X and T2T-HG002Y (hs1). **(C)** SD98 gene clustering into gene families based on shared exons and similar famCN (MAD<1) between paralogs. Copy number was calculated only for protein coding and unprocessed pseudogenes, but other overlapping gene features (i.e. lncRNA) were reported. **(D)** Predicted evolutionary status of SD98 gene families. **(E)** Copy-number variation analysis in SD98 regions. parCN: paralog-specific copy-number. **(F)** Outlier Tajima's D values across SD98 windows (>10% SD98) at least 50% accessible and carrying 5 or more SNPs. **(G)** *De novo* copy-number events of SD98 genes identified in the SSC.

Table S2. Human brain expression of SD98 genes related to Figure 2. (A) BrainSpan WGCNA module assignment. **(B)** Gene ontology overrepresentation of BrainSpan WGCNA modules. **(C)** Cortecon

WGCNA gene-module assignment. **(D)** Gene family co-expression concordance using CORTECON WGCNA. **(E)** Gene ontology overrepresentation of CORTECON WGCNA modules.

Table S3. Modeling SD98 genes in mouse and zebrafish related to Figure 3 and Discussion. (A) Mouse and zebrafish orthologs of SD98 gene families. **(B)** SD98 genes fetal brain expression in human, mouse and zebrafish orthologs.

Table S4. Sequence and variant analysis of priority human-specific duplicated (pHSD) genes related to Figure 4 and STAR Methods. (A) Summary of selected pHSD genes, canonical transcripts, variant calling, and variant effect prediction. **(B)** Individuals sequenced with capture PacBio HiFi long read sequencing. **(C)** Coordinates of captured pHSD regions and summary of long-read sequencing variant discovery. **(D)** Oligonucleotide baits design for cHiFi sequencing. **(E)** Summary statistics of cHiFi sequencing. **(F)** Variant effect prediction across pHSD paralogs. **(G)** pHSD coding variants and allele frequencies. **(H)** Ka/Ks, biallelic SNPs, pN/pS, and Direction of Selection across pHSD paralogs. **(I)** Comparison of dN/dS under different models using codeml.

Table S5. Details of zebrafish models of priority human-duplicated (pHSD) genes related to Figure 5, Discussion, and STAR Methods. (A) Oligonucleotide sequences used in this study. **(B)** Survival of mutant zebrafish larvae. **(C)** Distribution of the 3,146 images of larvae for morphological assessments. **(D)** Raw morphometric data for all zebrafish models of the selected pHSD genes for functional characterizations. **(E)** Statistical results from the morphological comparisons across zebrafish models of the selected pHSD genes.. **(F)** Description of the sci-RNA-seq identified clusters from heads of 3 dpf zebrafish larvae. **(G)** Marker genes for all identified clusters in the sci-RNA-seq data from heads of 3 dpf zebrafish larvae. **(H)** Gene ontology (GO) terms enriched in DEGs across zebrafish models of the selected pHSD genes for forebrain and midbrain. **(I)** GO terms enriched in DEGs for *SRGAP2* mutant zebrafish models. **(J)** GO terms enriched in DEGs for *ARHGAP11B* humanized zebrafish model.

Table S6. Zebrafish mutant models *gpr89* and *frmpd2* related to Figure 6 and STAR Methods. (A) Differentially expressed genes (DEGs) between *GPR89B* and *gpr89* knockout (KO) models and their respective controls. **(B)** Gene ontology (GO) terms enriched in DEGs between *GPR89B* and *gpr89* KO models and their respective controls. **(C)** DEGs between *FRMPD2*B and *frmpd2* KO models and their respective controls. **(D)** GO terms enriched in DEGs between *FRMPD2B* and *frmpd2* KO models and their respective controls. **(E)** Stable mutant zebrafish alleles.

Data S1. Weighted gene co-expression analysis of human brain samples related to Figure 2 and STAR Methods.

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. Ohno, S. (1970). Evolution by gene duplication (Allen & Unwin; Springer-Verlag).

13. Porubsky, D., and Eichler, E.E. (2024). A 25-year odyssey of genomic technology advances and structural variant discovery. Cell *187*, 1024–1037.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

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

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. Yilmaz, F., Karageorgiou, C., Kim, K., Pajic, P., Scheer, K., Human Genome Structural Variation Consortium, Beck, C.R., Torregrossa, A.-M., Lee, C., Gokcumen, O., et al. (2024). Reconstruction of the human amylase locus reveals ancient duplications seeding modern-day variation. Science *386*,

eadn0609.

45. 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.

46. 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.

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

48. 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.

49. 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.

50. 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.

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

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

53. 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.

54. 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.

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

57. Soto, D.C., Uribe-Salazar, J., and Dennis, M. mydennislab/HSD_brain_evolution: HSD-brain. https://doi.org/10.5281/zenodo.15486469.

58. 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.

59. 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.

60. An, J.-Y., Lin, K., Zhu, L., Werling, D.M., Dong, S., Brand, H., Wang, H.Z., Zhao, X., Schwartz, G.B., Collins, R.L., et al. (2018). Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science *362*. https://doi.org/10.1126/science.aat6576.

61. Zhou, J., Park, C.Y., Theesfeld, C.L., Wong, A.K., Yuan, Y., Scheckel, C., Fak, J.J., Funk, J., Yao, K., Tajima, Y., et al. (2019). Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. Nat Genet *51*, 973–980.

62. 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.

63. 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.

64. 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.

65. 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.

66. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y., and Pritchard, J.K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature *464*, 768–772.

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

68. 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.

69. Stühmer, T., Puelles, L., Ekker, M., and Rubenstein, J.L.R. (2002). Expression from a Dlx gene enhancer marks adult mouse cortical GABAergic neurons. Cereb Cortex *12*, 75–85.

70. Stühmer, T., Anderson, S.A., Ekker, M., and Rubenstein, J.L.R. (2002). Ectopic expression of the Dlx genes induces glutamic acid decarboxylase and Dlx expression. Development *129*, 245–252.

71. Song, Y.-H., Yoon, J., and Lee, S.-H. (2021). The role of neuropeptide somatostatin in the brain and its application in treating neurological disorders. Exp Mol Med *53*, 328–338.

72. 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.

73. 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.

74. 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.

75. 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.

76. 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.

77. 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.

78. 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.

79. 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.

80. 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.

81. 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.

82. 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). Semi-automated assembly of high-quality diploid human reference genomes. Nature *611*, 519–531.

83. 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. (2023). A draft human pangenome reference. Nature *617*, 312–324.

84. 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.

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

86. 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.

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

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

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

90. 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.

91. 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.

92. 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.

93. 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.

94. 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.

95. 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.

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

97. Yang, H., Luan, Y., Liu, T., Lee, H.J., Fang, L., Wang, Y., Wang, X., Zhang, B., Jin, Q., Ang, K.C., et al. (2020). A map of cis-regulatory elements and 3D genome structures in zebrafish. Nature *588*, 337–343.

98. 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.

99. 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.

100. 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.

101. 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.

102. 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.

103. 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.

104. 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.

105. 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.

106. White, R.J., Collins, J.E., Sealy, I.M., Wali, N., Dooley, C.M., Digby, Z., Stemple, D.L., Murphy, D.N., Billis, K., Hourlier, T., et al. (2017). A high-resolution mRNA expression time course of embryonic development in zebrafish. Elife *6*. https://doi.org/10.7554/eLife.30860.

107. 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.

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

109. 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.

110. Valdarrago, R.M., Hu, H., Li, R., Uribe-Salazar, J.M., Dennis, M.Y., and Quon, G. (2025). A conditional generative model to disentangle morphological variation from batch effects in model organism imaging studies. bioRxiv. https://doi.org/10.1101/2025.06.04.657966.

111. 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.

112. 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.

113. 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.

114. 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.

115. Birchler, J.A., and Veitia, R.A. (2012). Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines. Proc Natl Acad Sci U S A *109*, 14746–14753.

116. Papp, B., Pál, C., and Hurst, L.D. (2003). Dosage sensitivity and the evolution of gene families in yeast. Nature *424*, 194–197.

117. Milind, N., Smith, C.J., Zhu, H., Spence, J.P., and Pritchard, J.K. (2024). Buffering and non-monotonic behavior of gene dosage response curves for human complex traits. medRxiv. https://doi.org/10.1101/2024.11.11.24317065.

118. 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.

119. 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.

120. 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.

121. 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.

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

123. 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.

124. 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.

125. 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.

126. 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.

127. 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.

128. 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.

129. 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.

130. 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.

131. 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.

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

133. 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.

134. 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.

135. 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.

136. 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.

137. 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.

138. 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.

139. 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. (2025). Complete sequencing of ape genomes. Nature *641*, 401–418.

140. 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.

141. 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.

142. 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.

143. 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.

144. 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.

145. 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.

146. 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.

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

148. Fischer, J., Fernández Ortuño, E., Marsoner, F., Artioli, A., Peters, J., Namba, T., Eugster Oegema, C., Huttner, W.B., Ladewig, J., and Heide, M. (2022). Human-specific ARHGAP11B ensures human-like basal progenitor levels in hominid cerebral organoids. EMBO Rep *23*, e54728.

149. 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.

150. 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.

151. 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.

152. 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.

153. 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.

154. Vollger, M.R., Dishuck, P.C., Harvey, W.T., DeWitt, W.S., Guitart, X., Goldberg, M.E., Rozanski, A.N., Lucas, J., Asri, M., Human Pangenome Reference Consortium, et al. (2023). Increased mutation and gene conversion within human segmental duplications. Nature *617*, 325–334.

155. Dumont, B.L. (2015). Interlocus gene conversion explains at least 2.7% of single nucleotide variants

in human segmental duplications. BMC Genomics *16*, 456.

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

157. 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.

158. 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.

159. 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.

160. 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.

161. 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.

162. 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.

163. 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.

164. Jeong, H., Dishuck, P.C., Yoo, D., Harvey, W.T., Munson, K.M., Lewis, A.P., Kordosky, J., Garcia, G.H., Yilmaz, F., Hallast, P., et al. (2025). Structural polymorphism and diversity of human segmental duplications. Nature Genetics *57*, 390–401.

165. 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.

166. 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. (2024). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Nature Methods *21*, 1349–1363.

167. 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.

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

169. Gupta, P., O'Neill, H., Wolvetang, E.J., Chatterjee, A., and Gupta, I. (2024). Advances in single-cell long-read sequencing technologies. NAR Genom Bioinform *6*, lqae047.

170. Xue, J.R., Mackay-Smith, A., Mouri, K., Garcia, M.F., Dong, M.X., Akers, J.F., Noble, M., Li, X., Zoonomia Consortium†, Lindblad-Toh, K., et al. (2023). The functional and evolutionary impacts of human-specific deletions in conserved elements. Science *380*, eabn2253.

171. Fair, T., Pavlovic, B.J., Swope, D., Castillo, O.E., Schaefer, N.K., and Pollen, A.A. (2024). Mapping - and -regulatory target genes of human-specific deletions. bioRxiv. https://doi.org/10.1101/2023.12.27.573461.

172. Ding, W., Li, X., Zhang, J., Ji, M., Zhang, M., Zhong, X., Cao, Y., Liu, X., Li, C., Xiao, C., et al. (2024). Adaptive functions of structural variants in human brain development. Sci Adv *10*, eadl4600.

173. 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.

174. 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.

175. 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.

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

177. 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.

178. Liew, W.C., and Orbán, L. (2014). Zebrafish sex: a complicated affair. Brief Funct Genomics *13*, 172–187.

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

180. 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.

181. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

182. 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.

183. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. Gigascience *10*. https://doi.org/10.1093/gigascience/giab008.

184. Altemose, N., Logsdon, G.A., Bzikadze, A.V., Sidhwani, P., Langley, S.A., Caldas, G.V., Hoyt, S.J., Uralsky, L., Ryabov, F.D., Shew, C.J., et al. (2022). Complete genomic and epigenetic maps of human centromeres. Science *376*, eabl4178.

185. Cleary, J.G., Braithwaite, R., Gaastra, K., Hilbush, B.S., Inglis, S., Irvine, S.A., Jackson, A., Littin, R., Rathod, M., Ware, D., et al. (2015). Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. bioRxiv, 023754. https://doi.org/10.1101/023754.

186. 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.

187. 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.

188. 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.

189. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067.

190. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

191. 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.

192. 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.

193. Pertea, G., and Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. F1000Res *9*. https://doi.org/10.12688/f1000research.23297.2.

194. 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.

195. 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.

196. 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.

197. 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.

198. Bonfield, J.K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R.M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience *10*. https://doi.org/10.1093/gigascience/giab007.

199. Li, H., Bloom, J.M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., and MacArthur, D. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. Nat. Methods *15*, 595–597.

200. 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.

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

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

203. 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.

204. Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S.E., and Lercher, M.J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol. Biol. Evol. *31*, 1929–1936.

205. 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.

206. 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.

207. 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.

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

209. 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.

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

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

212. 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.

213. Valdarrago, R. Ricardo-scb/ZebraFish-Diffusion-Model: Initial Release. https://doi.org/10.5281/zenodo.15485460.

214. 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.

215. 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.

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

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

218. 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.

219. 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.

220. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573–3587.e29.

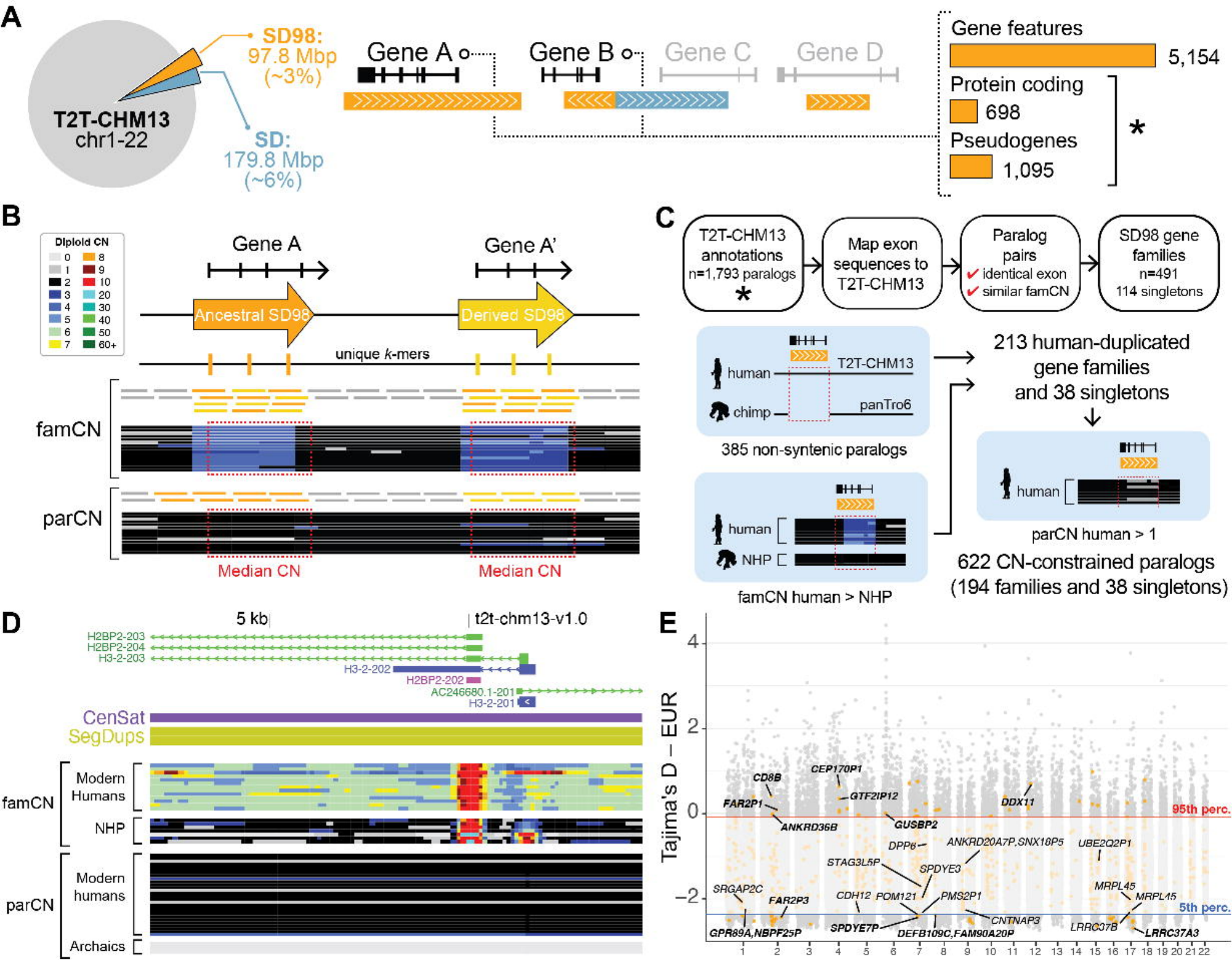
221. Tran, V., Papalexi, E., Schroeder, S., Kim, G., Sapre, A., Pangallo, J., Sova, A., Matulich, P., Kenyon, L., Sayar, Z., et al. (2022). High sensitivity single cell RNA sequencing with split pool barcoding. bioRxiv, 2022.08.27.505512. https://doi.org/10.1101/2022.08.27.505512.

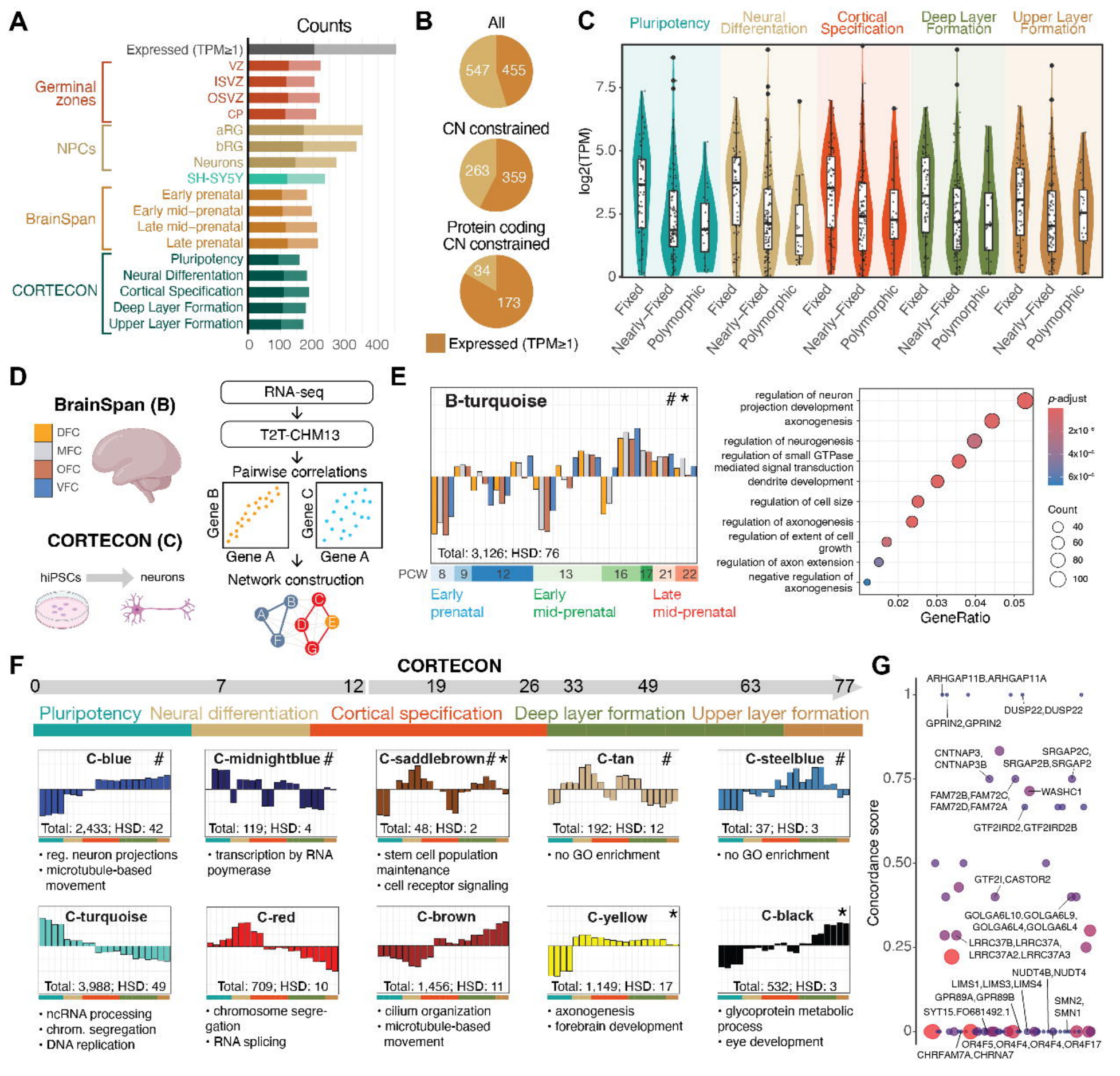
222. McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst *8*, 329–337.e4.

223. 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.

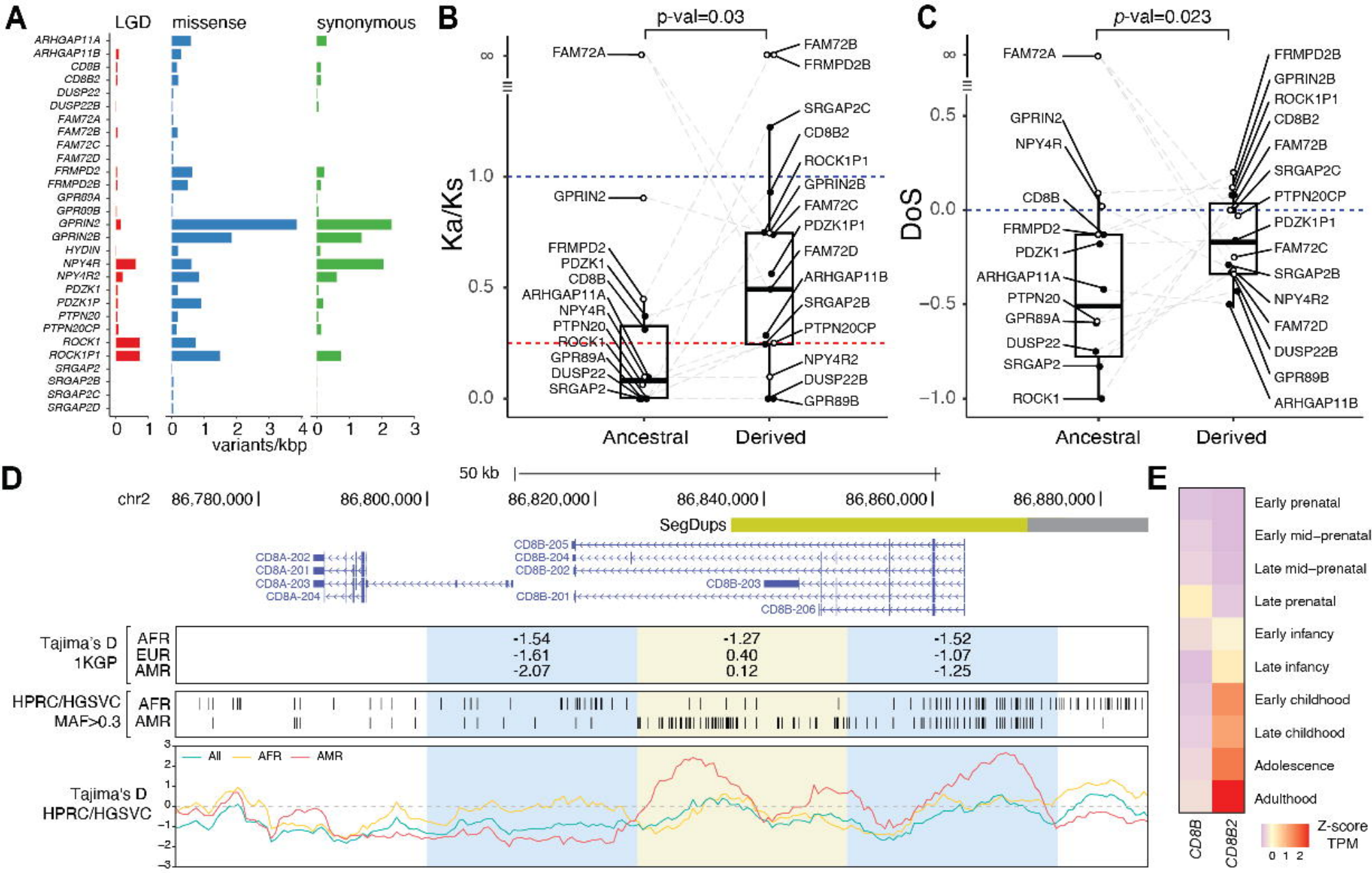
224. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

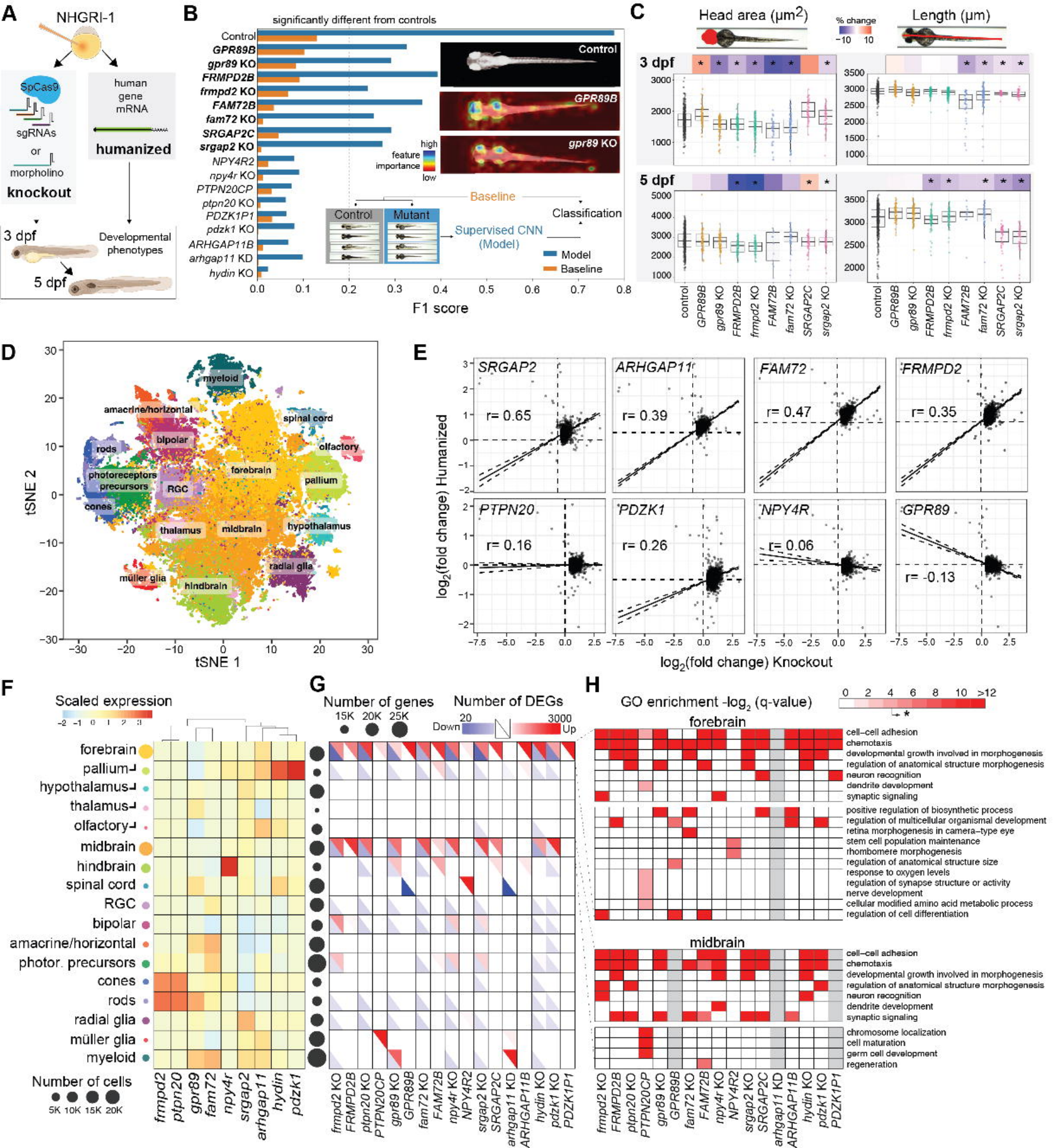
225. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

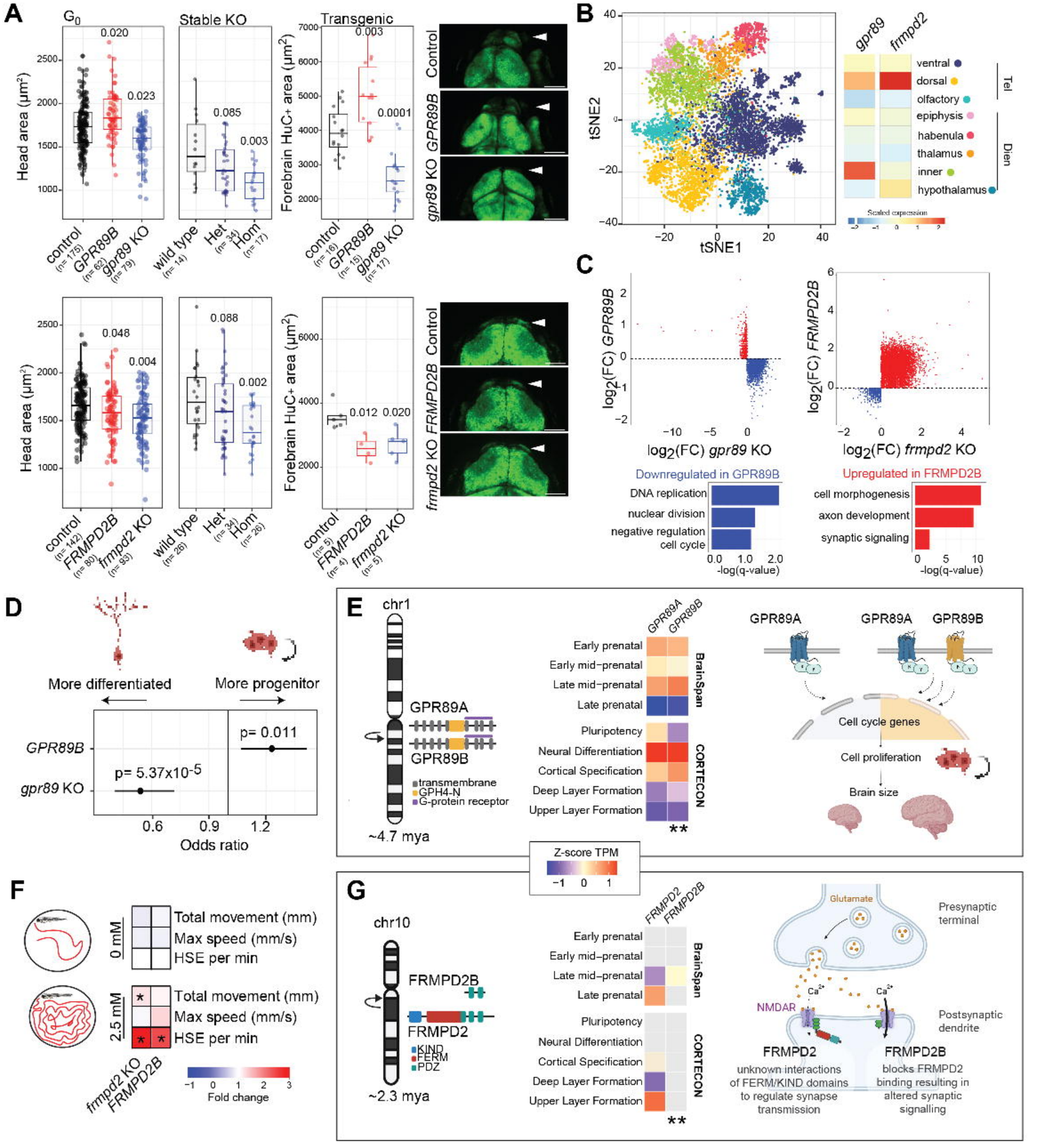

A
T2T-CHM13
chr1-22
SD98:
97.8 Mbp
(~3%)
SD:
179.8 Mbp
(~6%)
Gene A
Gene B
Gene C
Gene D
Gene features
5,154
Protein coding
698
Pseudogenes
1,095
*
B
Diploid CN
0
1
2
3
4
5
6
7
8
9
10
20
30
40
50
60+
Gene A
Gene A'
Ancestral SD98
Derived SD98
unique k-mers
famCN
parCN
Median CN
Median CN
C
T2T-CHM13 annotations
n=1,793 paralogs
*
Map exon sequences to T2T-CHM13
Paralog pairs
identical exon
similar famCN
SD98 gene families
n=491
114 singletons
human
T2T-CHM13
chimp
panTro6
385 non-syntenic paralogs
213 human-duplicated gene families and 38 singletons
human
parCN human > 1
human
NHP
famCN human > NHP
622 CN-constrained paralogs
(194 families and 38 singletons)
D
5 kb
t2t-chm13-v1.0
H2BP2-203
H2BP2-204
H3-2-203
H3-2-202
H2BP2-202
AC246680.1-201
H3-2-201
CenSat
SegDups
famCN
Modern Humans
NHP
parCN
Modern humans
Archaics
E
Tajima's D – EUR
4
2
0
−2
95th perc.
5th perc.
CD8B
CEP170P1
FAR2P1
GTF2IP12
DDX11
ANKRD36B
GUSBP2
DPP6
ANKRD20A7P,SNX18P5
UBE2Q2P1
STAG3L5P
SPDYE3
SRGAP2C
FAR2P3
CDH12
POM121
PMS2P1
MRPL45
MRPL45
CNTNAP3
GPR89A,NBPF25P
SPDYE7P
DEFB109C,FAM90A20P
LRRC37B
LRRC37A3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22


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 Differentation
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 Differentation
Cortical Specification
Deep Layer Formation
Upper Layer Formation
log2(TPM)
7.5
5.0
2.5
0
Fixed
Nearly-Fixed
Polymorphic
D
BrainSpan (B)
DFC
MFC
OFC
VFC
CORTECON (C)
hiPSCs
neurons
RNA-seq
T2T-CHM13
Pairwise correlations
Gene B
Gene A
Gene C
Gene A
Network construction
E
B-turquoise
Total: 3,126; HSD: 76
PCW 8 9 12 13 16 17 21 22
Early prenatal
Early mid-prenatal
Late mid-prenatal
regulation of neuron projection development
axonogenesis
regulation of neurogenesis
regulation of small GTPase mediated signal transduction
dendrite development
regulation of cell size
regulation of axonogenesis
regulation of extent of cell growth
regulation of axon extension
negative regulation of axonogenesis
0.02 0.03 0.04 0.05
GeneRatio
p-adjust
Count
40
60
80
100
F
CORTECON
0 7 12 19 26 33 49 63 77
Pluripotency
Neural differentiation
Cortical specification
Deep layer formation
Upper layer formation
C-blue
Total: 2,433; HSD: 42
• reg. neuron projections
• microtubule-based movement
C-midnightblue
Total: 119; HSD: 4
• transcription by RNA poymerase
C-saddlebrown
Total: 48; HSD: 2
• stem cell population maintenance
• cell receptor signaling
C-tan
Total: 192; HSD: 12
• no GO enrichment
C-steelblue
Total: 37; HSD: 3
• no GO enrichment
C-turquoise
Total: 3,988; HSD: 49
• ncRNA processing
• chrom. segregation
• DNA replication
C-red
Total: 709; HSD: 10
• chromosome segregation
• RNA splicing
C-brown
Total: 1,456; HSD: 11
• cilium organization
• microtubule-based movement
C-yellow
Total: 1,149; HSD: 17
• axonogenesis
• forebrain development
C-black
Total: 532; HSD: 3
• glycoprotein metabolic process
• eye development
G
Concordance score
1
0.75
0.50
0.25
0
ARHGAP11B,ARHGAP11A
DUSP22,DUSP22
GPRIN2,GPRIN2
CNTNAP3, CNTNAP3B
SRGAP2C, SRGAP2B,SRGAP2
FAM72B,FAM72C, FAM72D,FAM72A
WASHC1
GTF2IRD2,GTF2IRD2B
GTF2I,CASTOR2
GOLGA6L10,GOLGA6L9, GOLGA6L4,GOLGA6L4
LRRC37B,LRRC37A, LRRC37A2,LRRC37A3
NUDT4B,NUDT4
LIMS1,LIMS3,LIMS4
GPR89A,GPR89B
SYT15,FO681492.1
SMN2, SMN1
OR4F5,OR4F4,OR4F4,OR4F17
CHRFAM7A,CHRNA7

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
IQSEC3
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
HERC2P2
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
GOLGA8G
GOLGA8F
GOLGA8S
GOLGA8DP
GOLGA8IP
GOLGA8EP
GOLGA8CP
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
1.5
1.0
0.5
0.0
−0.5
Scaled expression


A
LGD
missense
synonymous
ARHGAP11A
ARHGAP11B
CD8B
CD8B2
DUSP22
DUSP22B
FAM72A
FAM72B
FAM72C
FAM72D
FRMPD2
FRMPD2B
GPR89A
GPR89B
GPRIN2
GPRIN2B
HYDIN
NPY4R
NPY4R2
PDZK1
PDZK1P
PTPN20
PTPN20CP
ROCK1
ROCK1P1
SRGAP2
SRGAP2B
SRGAP2C
SRGAP2D
variants/kbp
B
p-val=0.03
Ka/Ks
Ancestral
Derived
C
p-val=0.023
DoS
Ancestral
Derived
D
50 kb
chr2
86,780,000
86,800,000
86,820,000
86,840,000
86,860,000
86,880,000
SegDups
CD8A-202
CD8A-201
CD8A-203
CD8A-204
CD8B-205
CD8B-204
CD8B-202
CD8B-203
CD8B-201
CD8B-206
Tajima's D
1KGP
AFR
EUR
AMR
-1.54
-1.27
-1.52
-1.61
0.40
-1.07
-2.07
0.12
-1.25
HPRC/HGSVC
MAF>0.3
Tajima's D
HPRC/HGSVC
All
AFR
AMR
E
Early prenatal
Early mid−prenatal
Late mid−prenatal
Late prenatal
Early infancy
Late infancy
Early childhood
Late childhood
Adolescence
Adulthood
CD8B
CD8B2
Z-score
TPM


A
NHGRI-1
SpCas9
sgRNAs
or
morpholino
knockout
human
gene
mRNA
humanized
3 dpf
5 dpf
Developmental
phenotypes
B
significantly different from controls
Control
GPR89B
gpr89 KO
FRMPD2B
frmpd2 KO
FAM72B
fam72 KO
SRGAP2C
srgap2 KO
NPY4R2
npy4r KO
PTPN20CP
ptpn20 KO
PDZK1P1
pdzk1 KO
ARHGAP11B
arhgap11 KO
hydin KO
high
feature
importance
low
Baseline
Control
Mutant
Classification
Supervised CNN
(Model)
Model
Baseline
F1 score
C
Head area (μm2)
% change
−10
10
Length (μm)
3 dpf
5 dpf
control
GPR89B
gpr89 KO
FRMPD2B
frmpd2 KO
FAM72B
fam72 KO
SRGAP2C
srgap2 KO
D
tSNE 1
tSNE 2
myeloid
amacrine/horizontal
spinal cord
bipolar
rods
olfactory
photoreceptors
precursors
forebrain
pallium
RGC
cones
thalamus
midbrain
hypothalamus
radial glia
müller glia
hindbrain
E
SRGAP2
r= 0.65
ARHGAP11
r= 0.39
FAM72
r= 0.47
FRMPD2
r= 0.35
PTPN20
r= 0.16
PDZK1
r= 0.26
NPY4R
r= 0.06
GPR89
r= -0.13
log2(fold change) Humanized
log2(fold change) Knockout
F
Scaled expression
forebrain
pallium
hypothalamus
thalamus
olfactory
midbrain
hindbrain
spinal cord
RGC
bipolar
amacrine/horizontal
photor. precursors
cones
rods
radial glia
müller glia
myeloid
Number of cells
5K 10K 15K 20K
frmpd2
ptpn20
gpr89
fam72
npy4r
srgap2
arhgap11
hydin
pdzk1
G
Number of genes
15K 20K 25K
Number of DEGs
20
3000
Down
Up
H
GO enrichment -log2 (q-value)
0 2 4 6 8 10 >12
forebrain
cell–cell adhesion
chemotaxis
developmental growth involved in morphogenesis
regulation of anatomical structure morphogenesis
neuron recognition
dendrite development
synaptic signaling
positive regulation of biosynthetic process
regulation of multicellular organismal development
retina morphogenesis in camera–type eye
stem cell population maintenance
rhombomere morphogenesis
regulation of anatomical structure size
response to oxygen levels
regulation of synapse structure or activity
nerve development
cellular modified amino acid metabolic process
regulation of cell differentiation
midbrain
cell–cell adhesion
chemotaxis
developmental growth involved in morphogenesis
regulation of anatomical structure morphogenesis
neuron recognition
dendrite development
synaptic signaling
chromosome localization
cell maturation
germ cell development
regeneration


A
G0
Stable KO
Transgenic
Head area (μm²)
Forebrain HuC+ area (μm²)
0.020
0.023
0.085
0.003
0.003
0.0001
control (n= 175)
GPR89B (n= 62)
gpr89 KO (n= 79)
wild type (n= 14)
Het (n= 34)
Hom (n= 17)
control (n= 16)
GPR89B (n= 15)
gpr89 KO (n= 17)
Control
GPR89B
gpr89 KO
0.048
0.004
0.088
0.002
0.012
0.020
control (n= 142)
FRMPD2B (n= 80)
frmpd2 KO (n= 93)
wild type (n= 26)
Het (n= 34)
Hom (n= 26)
control (n= 5)
FRMPD2B (n= 4)
frmpd2 KO (n= 5)
FRMPD2B
frmpd2 KO
B
tSNE1
tSNE2
gpr89
frmpd2
ventral
dorsal
olfactory
epiphysis
habenula
thalamus
inner
hypothalamus
Tel
Dien
Scaled expression
C
log2(FC) GPR89B
log2(FC) gpr89 KO
log2(FC) FRMPD2B
log2(FC) frmpd2 KO
Downregulated in GPR89B
DNA replication
nuclear division
negative regulation cell cycle
-log(q-value)
Upregulated in FRMPD2B
cell morphogenesis
axon development
synaptic signaling
-log(q-value)
D
More differentiated
More progenitor
GPR89B
gpr89 KO
p= 0.011
p= 5.37x10-5
Odds ratio
E
chr1
GPR89A
GPR89B
transmembrane
GPH4-N
G-protein receptor
~4.7 mya
GPR89A
GPR89B
Early prenatal
Early mid-prenatal
Late mid-prenatal
Late prenatal
BrainSpan
Pluripotency
Neural Differentiation
Cortical Specification
Deep Layer Formation
Upper Layer Formation
CORTECON
**
Cell cycle genes
Cell proliferation
Brain size
Z-score TPM
F
0 mM
2.5 mM
Total movement (mm)
Max speed (mm/s)
HSE per min
frmpd2 KO
FRMPD2B
Fold change
G
chr10
FRMPD2B
FRMPD2
KIND
FERM
PDZ
~2.3 mya
FRMPD2
FRMPD2B
Early prenatal
Early mid-prenatal
Late mid-prenatal
Late prenatal
BrainSpan
Pluripotency
Neural Differentiation
Cortical Specification
Deep Layer Formation
Upper Layer Formation
CORTECON
**
Glutamate
Presynaptic terminal
NMDAR
Postsynaptic dendrite
FRMPD2
unknown interactions of FERM/KIND domains to regulate synapse transmission
FRMPD2B
blocks FRMPD2 binding resulting in altered synaptic signalling