Supplemental Information

Gene expansions contributing to human brain evolution

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

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

[*]These authors contributed equally to this work.

Supplemental Notes

Note S1. Varied confidence in identifying variation across recent duplications

Despite improved representation of duplicated genes in T2T-CHM13, genomic assessment of these regions remains challenging using short-read Illumina data. Duplicated regions are significantly depleted for SNVs in the high-coverage 1KGP dataset compared to unique regions (excluding SD and centromeric satellites [1]) in T2T-CHM13 (SD98: 11.79; SD: 25.9, unique: 37.49 SNVs/kbp; *p*-value<0.05 empirical distribution) (Figure SN1A). The autosomal 2.4 Gbp in T2T-CHM13 accessible for accurate Illumina SNV calling —determined using read depth, mapping quality, and base quality metrics [2]—includes only 37.95% and 10.86% of SD and SD98, respectively, while 95.64% of unique space is accessible (Figure SN2B). In the SD-98 regions, only 56 previously-identified SD98 genes, including 48 protein coding and 8 pseudogenes, are accessible (>90%) to short-reads (Table S1).

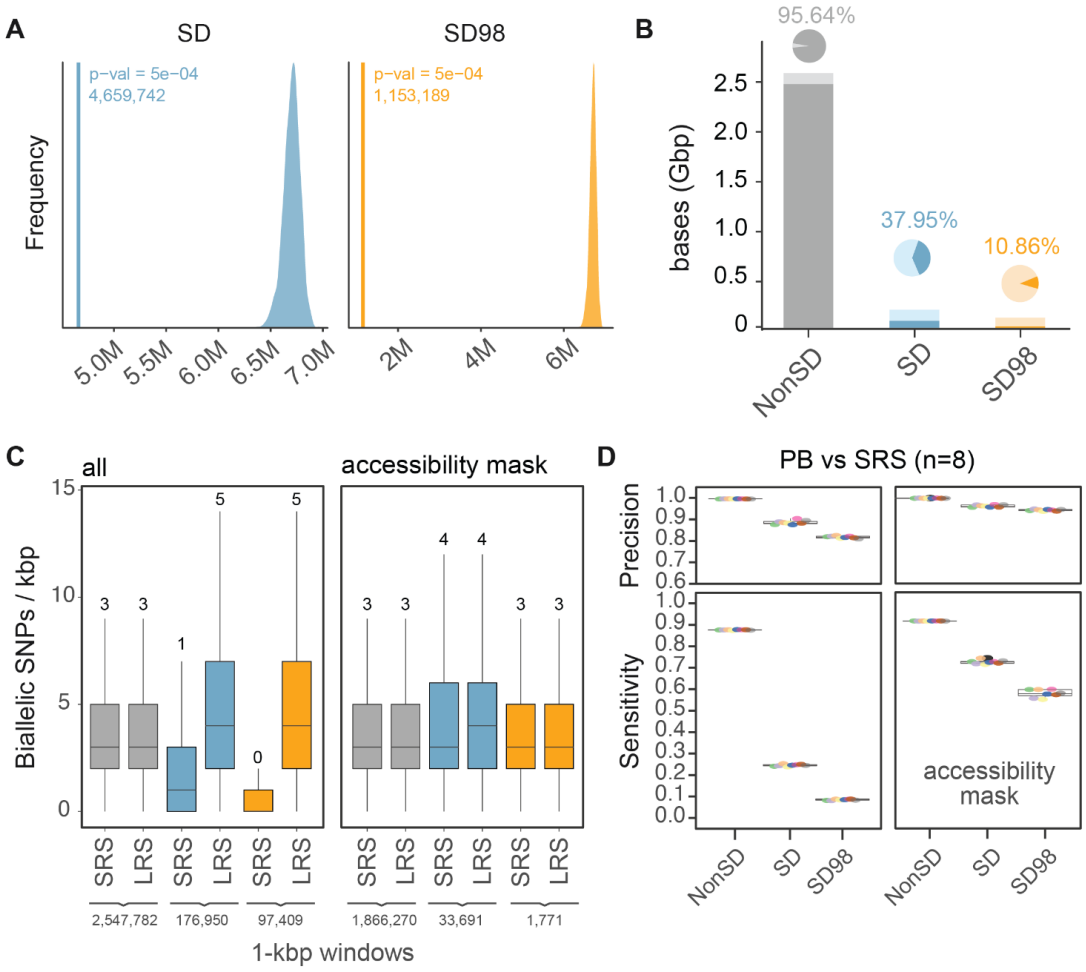

Figure SN1. Assessment of variant calling sensitivity and precision across SD and SD98 regions. (A) SNVs called in SD (blue, left) and SD98 (orange, right) using the T2T-CHM13 regions. Observed values are shown as vertical bars, while empirical distributions of the number of variants observed in randomly sampled regions are represented as density plots. (**B**) Total region size (in Gbp) and accessible sites size (darker colors), for NonSD (gray), SD (blue), and SD98 (orange). (**C**) 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 (right) technologies. Number at the bottom represents the total number of 1-kbp windows defined for each region. (**D**) 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).

To evaluate our ability to detect variants within duplications, we compared SNVs discovered in the T2T-CHM13 reference genome using Illumina short-read and PacBio HiFi long-read data across eight 1KGP individuals included in the Human Pangenome Reference Consortium (HPRC+) [2,3]. While no differences in density (SNV sites within 1-kbp non-overlapping windows) existed between data types in non-duplicated, 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 SN1C). Notably, no differences were observed between technologies when considering only T2T-CHM13 accessible regions [4].

Using HiFi-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 SN1D). 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.

Note S2. Underrepresentation of phenotype and disease associations across SDs

Due to difficulties mapping short reads to highly identical regions, as well as lack of SD representation on SNP arrays, associated variants and genes across SD-98 regions are depleted in existing genome-wide studies of phenotypes and diseases, including GWAS catalog (SD-98: 0.29 variants/100kbp; GW: 1.5 variants/100 kbp), ClinVar (SD-98: 20.81 variants/100 kbp; GW: 9.95 variants/100 kbp), and GTEx expression quantitative trait loci (eQTL) databases (SD-98: 398.7 variants/100 kbp; GW: 70.14 variants/100 kbp) (Figure SN2).

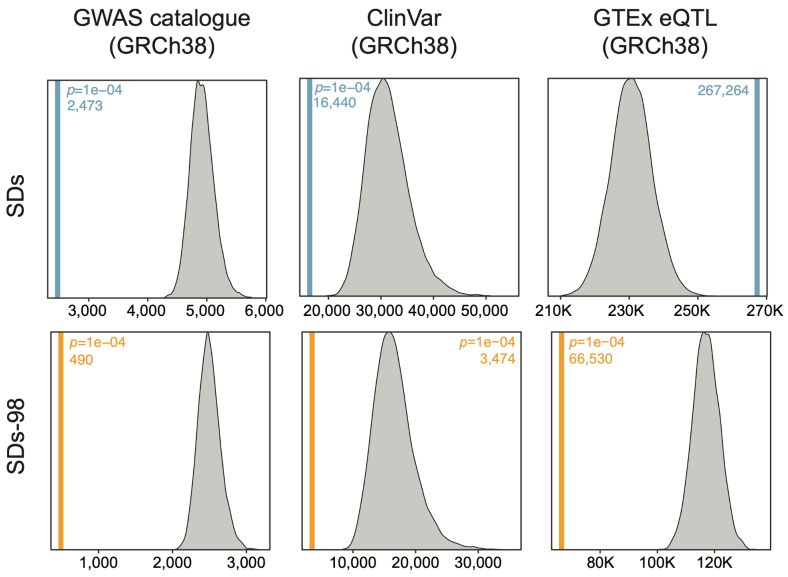

Figure SN2. Assessment of variant association depletion in SD and SD98 regions in short-read-based databases, including the GWAS catalog, ClinVar, and GTEx eQTL. Observed variation is represented in vertical lines for SD (blue) and SD98 (orange) regions, and density plots represent empirical distribution of randomly sampled sites of the same size as SD or SD98 regions.

Note S3. Benchmarking and variant calling with cHiFi data

To gain a more comprehensive understanding of genetic variation across pHSD genes, we performed capture HiFi (cHiFi) sequencing of 172 human samples including individuals from the 1000 Genomes Project and the Genome in a Bottle (GIAB) CEPH trio (YRI n=50, LWK, n=30, GIH n=29, TSI n=30, CHB n=30; CEU n=3; Table S13). Additionally, we sequenced 26 HGDP individuals from 12 additional populations, totaling 200 cHiFi samples from 18 different populations, representing five continental superpopulations, and 18 father-mother-child trios.

The capture sequencing protocol included tiled baits across all duplicated regions of interest and only exons in non-duplicated space (Figure S12). As a result, unique exons exhibited significantly lower coverage compared to duplicated exons (Mann-Whitney U test, *p*-value=2.2e-16) (Figure SN3A). 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). Globally, considering a cutoff MAPQ score greater than 10, we achieved an average coverage of 27× within regions of interest. 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.

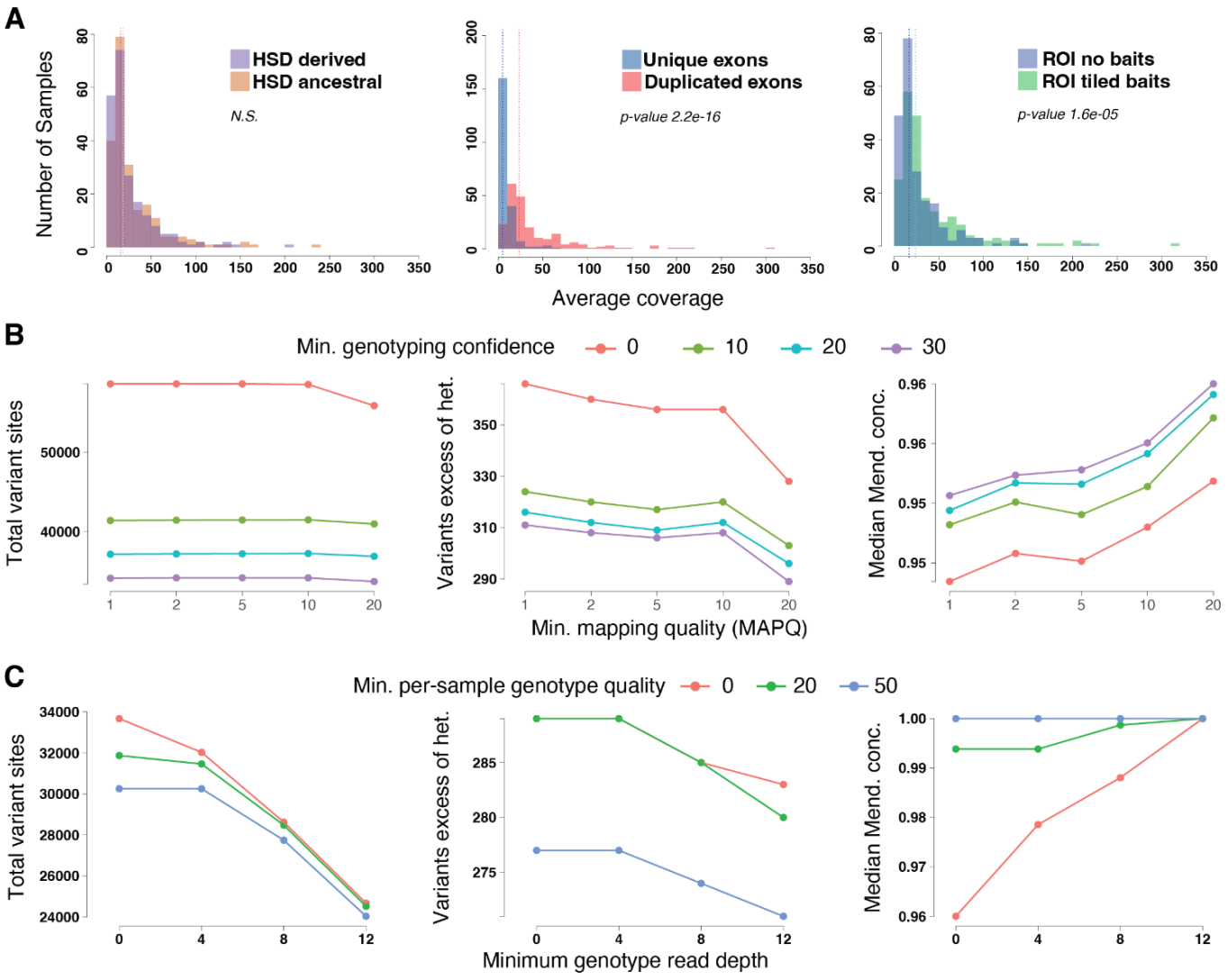

Figure SN3. Benchmarking capture HiFi sequencing variants. (**A**) Comparison of sequencing coverage between derived and ancestral paralogs (left), unique exons and duplicated exons (middle), and tiled versus untiled regions (right). (**B**) Impact of mapping quality (MAPQ) and genotyping confidence thresholds on the total number of variant sites (left), variant sites with excess heterozygosity (middle), and median Mendelian concordance across 18 trios (right). (**C**) Effect of 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).

We next identified small variants using a GATK joint genotyping pipeline. Focusing on biallelic SNPs, we optimized parameters for variant calling across SDs by benchmarking minimum thresholds for mapping quality (MAPQ) and genotyping confidence, using both population data and trio-based analysis. Specifically, we assessed deviations from Hardy-Weinberg equilibrium by 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. As MAPQ and genotyping confidence thresholds tighten, the total number of variant sites decreased, while biological metrics improved, including increased Mendelian concordance and reduced excess heterozygosity (Figure SN3B). The greatest improvement in variant accuracy was observed when increasing genotyping confidence from 0 to 10, with a threshold of 30 yielding the most accurate results. A minimum MAPQ threshold of 20 further reduced sites with excess heterozygosity and improved Mendelian concordance across all genotyping confidence levels, while only marginally reducing the number of detected variants.

We conservatively proceeded with a minimum MAPQ of 20 and a minimum genotyping confidence threshold of 30. Given that duplicated regions are recalcitrant to filtering techniques such as VQSR, we applied hard filtering thresholds based on per-sample genotype quality and read depth. Starting with a median Mendelian concordance of 0.96 prior to filtering, we observed a significant increase in concordance with increasing read-depth and genotype-quality thresholds, achieving near 100% concordance at a genotype quality of 50, or with the combination of read depth 8 and genotype quality 20 (Figure SN3C). To balance genotype accuracy with sensitivity, we opted for a more lenient minimum genotype quality threshold of 20 combined with a read depth of 8, resulting in the identification of 28,476 biallelic SNVs across the 200-individual cohort.

Supplemental Note Methods

Variant depletion across duplicated regions

Variants from 1KGP individuals mapped to T2T-CHM13 (v1.0) [2] were filtered for biallelic SNPs using bcftools view using parameters --exclude-types indels and --max-alleles 2. Observed values were obtained by intersecting SD and SD98 coordinates with the variant file using bedtools intersect. Empirical distributions were 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 [1] were also excluded using the flag -excl.

Short-read and long-read variant calling benchmarking

Comparison between SNVs discovered with PacBio HiFi and Illumina sequencing were obtained for eight individuals of the 1KGP and HPRC+ datasets mapped to T2T-CHM13 (v1.0) [2], including individuals HG01109, HG01243, HG02055, HG02080, HG02145, HG02723, HG03098, and HG03492. Biallelic SNVs were selected using bcftools view [5]. Concordance between platforms, measured as precision and sensitivity, was obtained with rtg-tools vcfeval [6] for autosomal Non-SDs, SDs, and SD-98 regions, using PacBio HiFi variants as a truth-set. Short-read accessible regions were obtained from Aganezov et al. [2]

Variant-phenotype associations depletion

Databases of genetic analyses were obtained from GWAS Catalog v1.0 (mapped to GRCh38.p12) [7], ClinVar (rel. 20200310) [8], and GTEx v8 single-tissue eQTL (dbGaP Accession phs000424.v8.p2; mapped to GRCh38, excluding chromosome Y) [9]. Empirical distributions were generated by intersecting each dataset with randomly sampled regions of identical size to SD and SD-98 generated with bedtools shuffle -noOverlapping -maxTries 10000 -f 0.1.

Empirical *p*-value calculation

One-tailed empirical p-values were calculated as: p-value = (M + 1) / (N + 1), where M is the number of iterations yielding a number of features less than (depletion) observed and N is the number of iterations. Empirical *p*-values were calculated using 10,000 permutations.

cHiFi variant calling benchmarking

cHiFi coverage across regions of interest was calculated using samtools depth with --min-MQ 10. Inbreeding coefficients and Mendelian concordance were calculated for biallelic SNPs only, selected with bcftools view --max-alleles 2 and bcftools view --exclude-types indels. Inbreeding coefficients were extracted from GATK Joint Genotyping output [10], which were originally calculated considering only the parental samples. Mendelian concordance was calculated for each threshold combination using rtg mendelian, 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.

Supplemental Figures

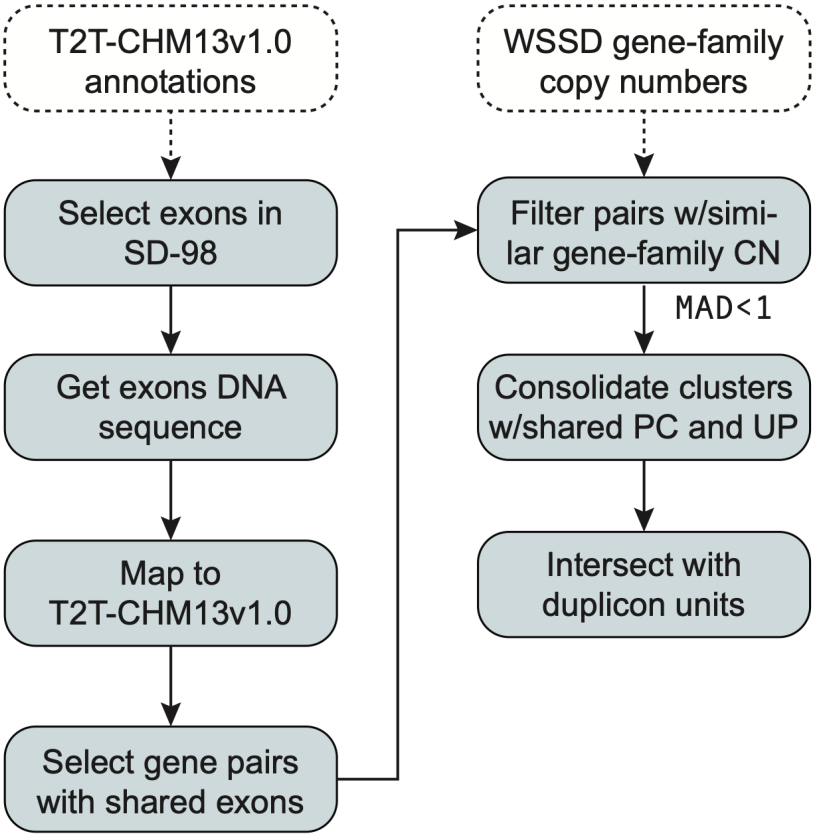

Figure S1. Pipeline to group SD98 genes into gene families.

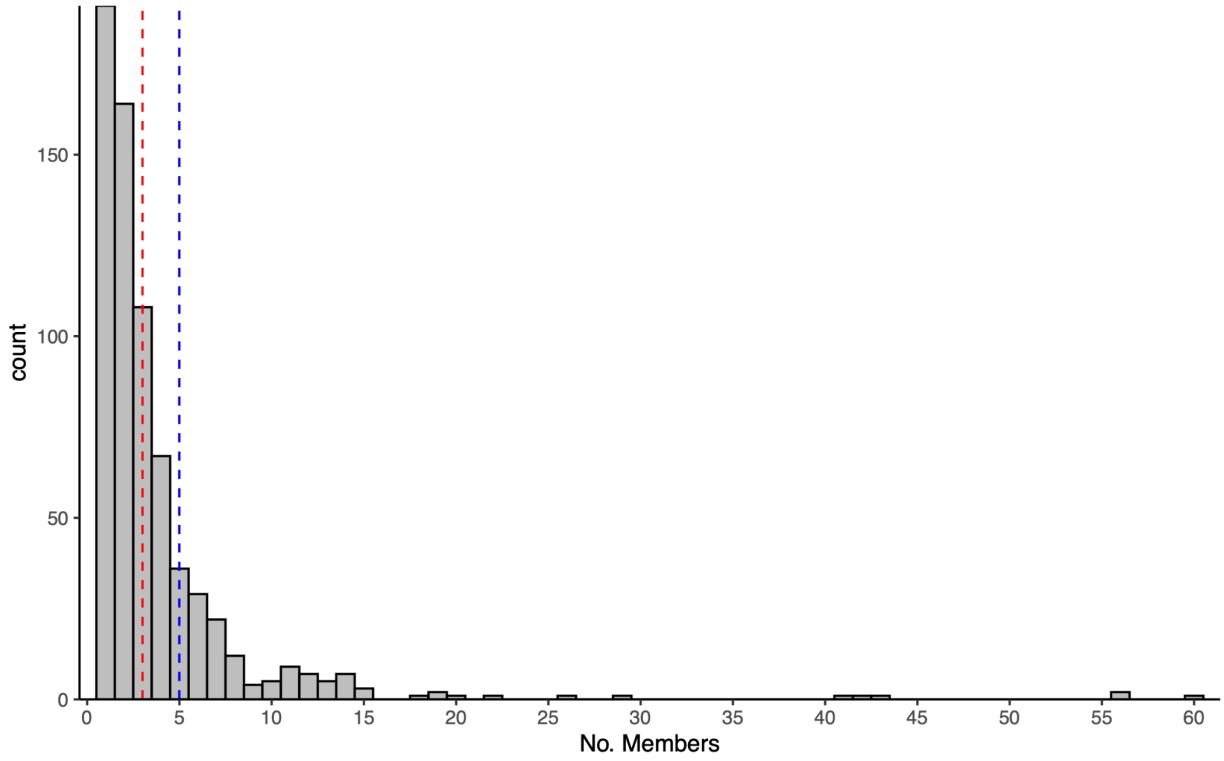

Figure S2. Distribution of number of gene members within duplicate gene families.

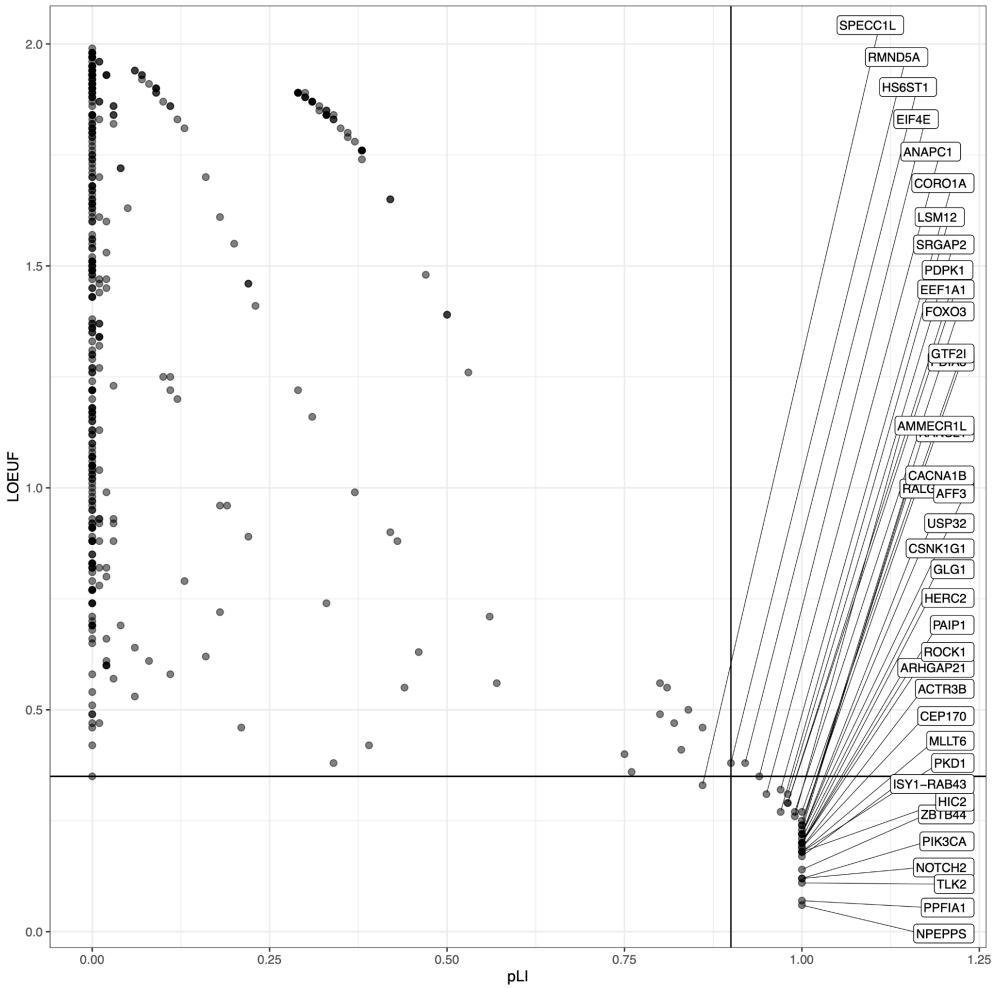

Figure S3. genomAD pLI versus LOEUF scores for all SD98 genes with available scores.

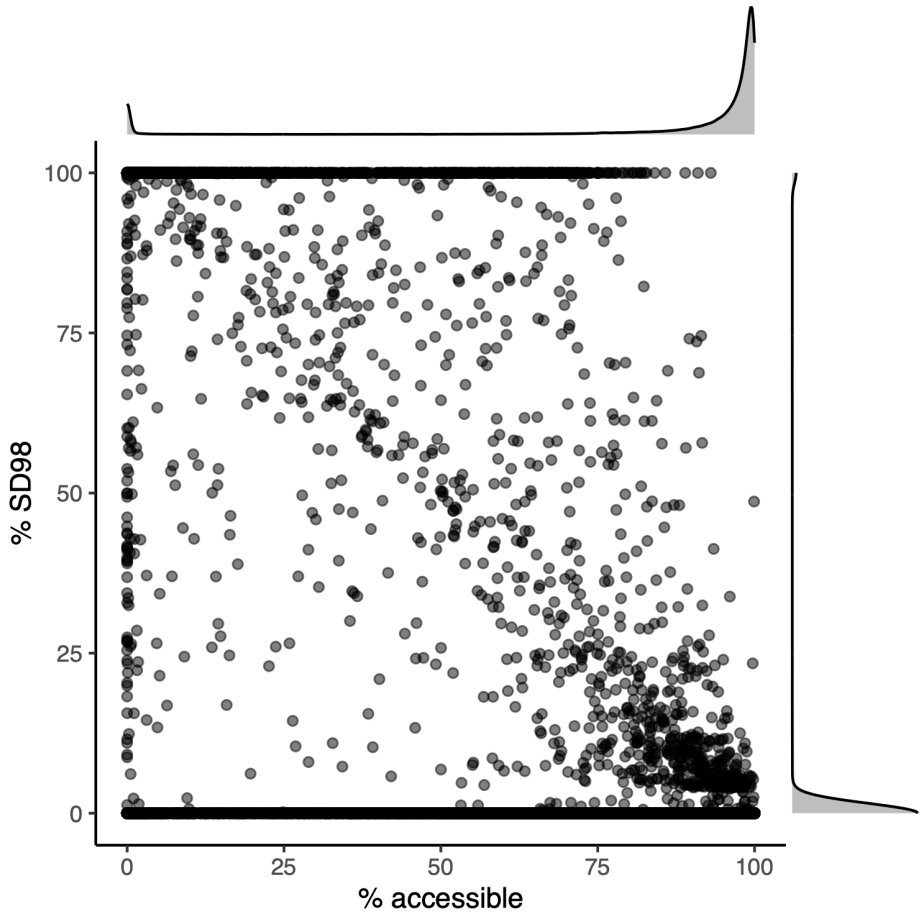

Figure S4. Percentage of short-read accessible bases versus percentage of bases within SD98 regions for 25-kbp windows genome-wide used in Tajima's D calculations.

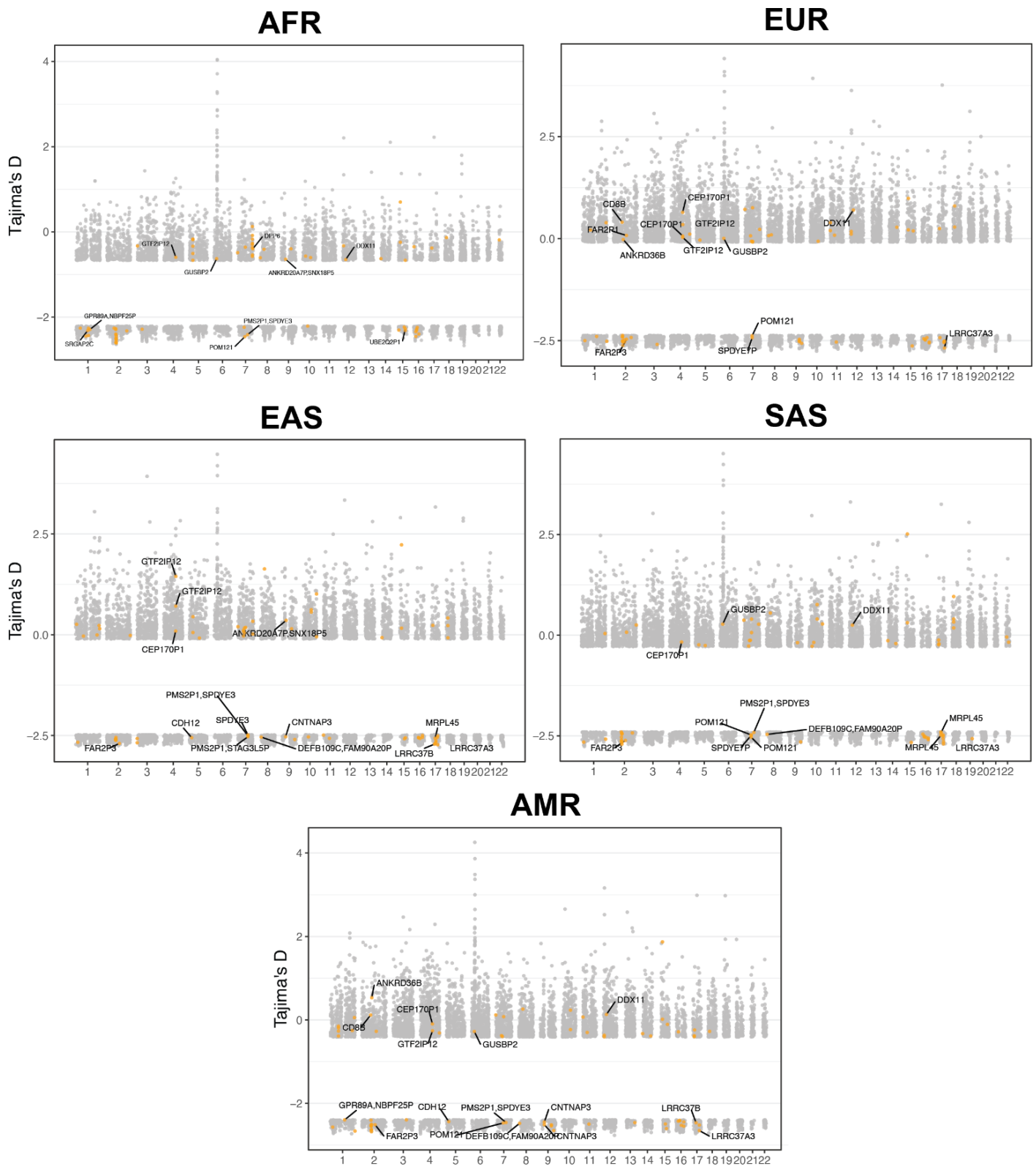

Figure S5. Tajima's D values from individuals of the 1000 Genomes Project across 25-kbp windows genome-wide (gray) and in SD98 region (orange) divided per superpopulation. Only outlier values in the upper 95th percentile or bottom 5th percentile are shown, plotted across human autosomal chromosomes on the x-axis. Human duplicated genes within windows with outlier D values are highlighted. Ancestries depicted include African (AFR), European (EUR), East Asian (EAS), South Asian (SAS), and American (AMR).

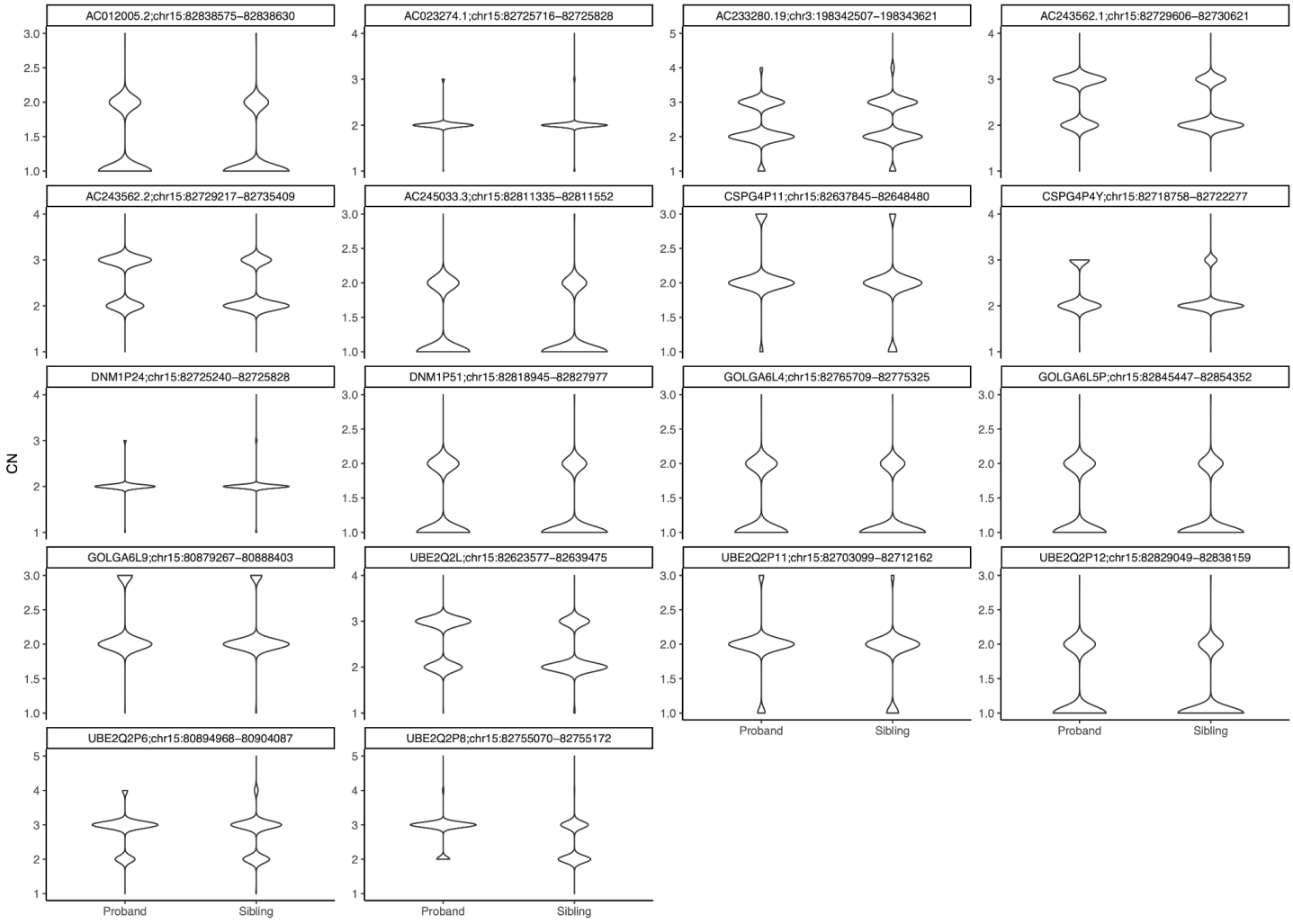

Figure S6. Human duplicated genes with significant copy-number differences between autistic probands and unaffected siblings from the Simons Simplex Collection.

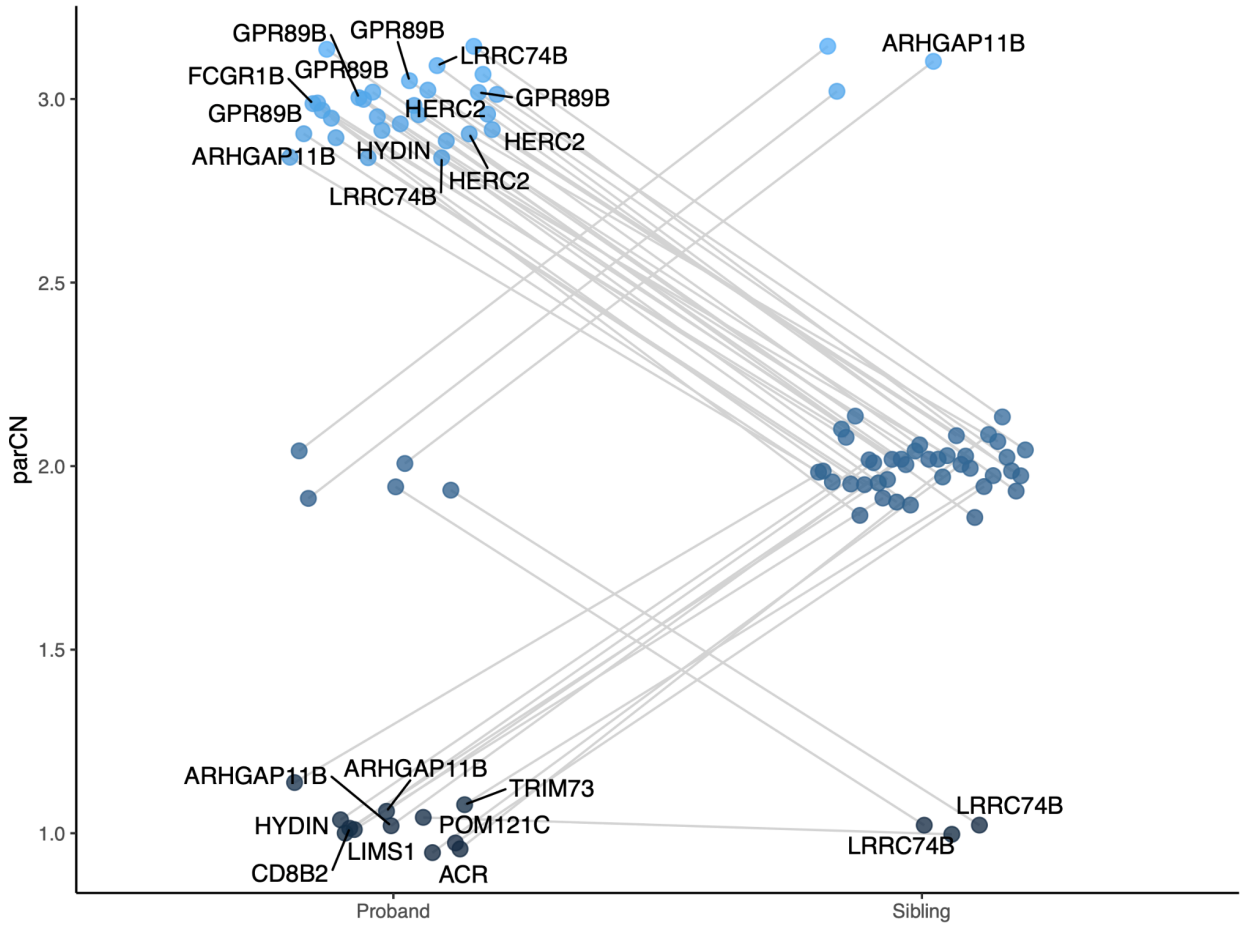

Figure S7. *De novo* deletions (copy number [CN]=1) and duplications (CN=3) identified in autistic probands and unaffected siblings from the Simon Simplex Collection. CN fixed human duplicated genes were considered in the analysis. Only protein coding genes displaying a *de novo* event are highlighted.

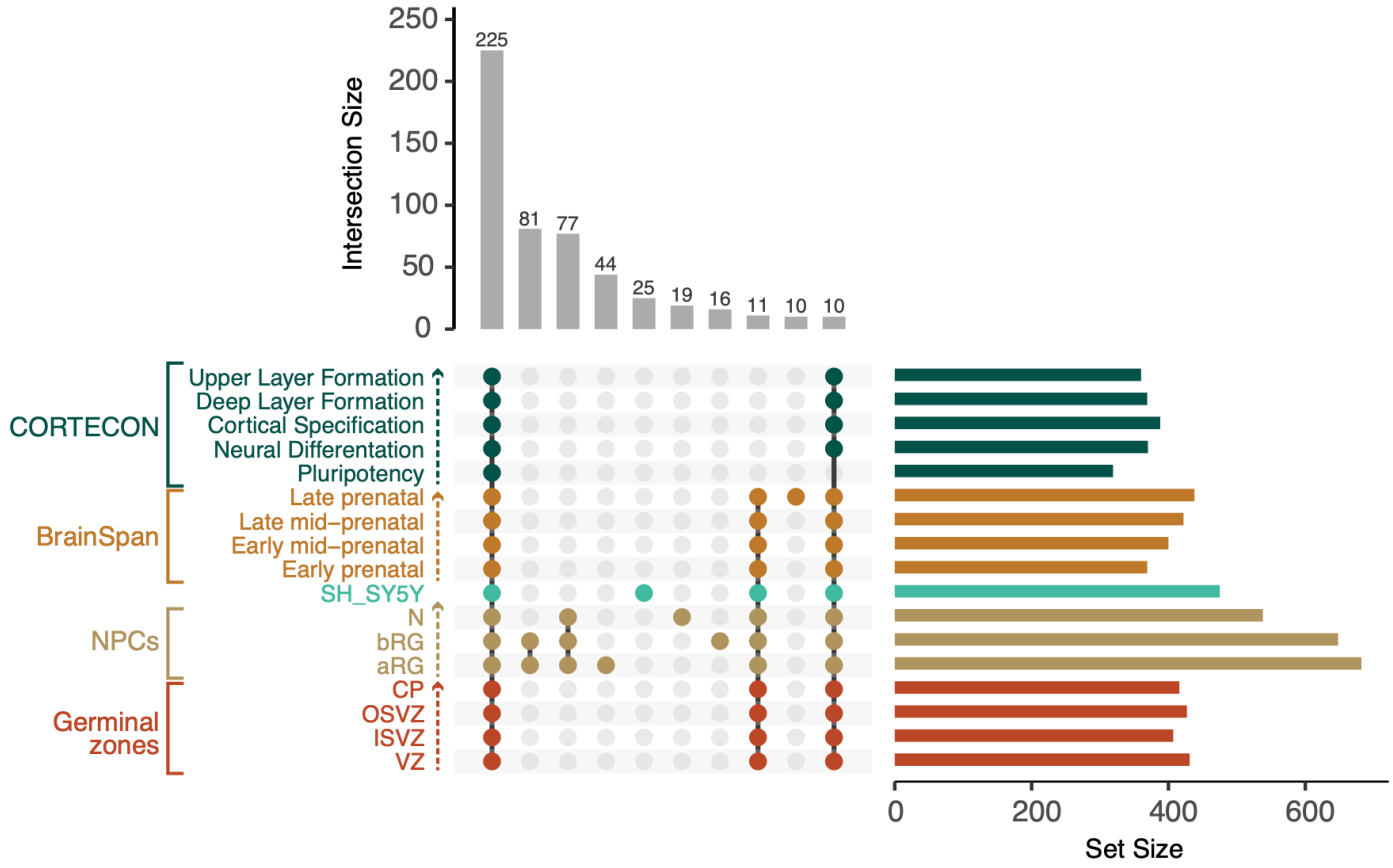

Figure S8. Intersection between human duplicated genes expressed (TPM≥1) across prenatal datasets.

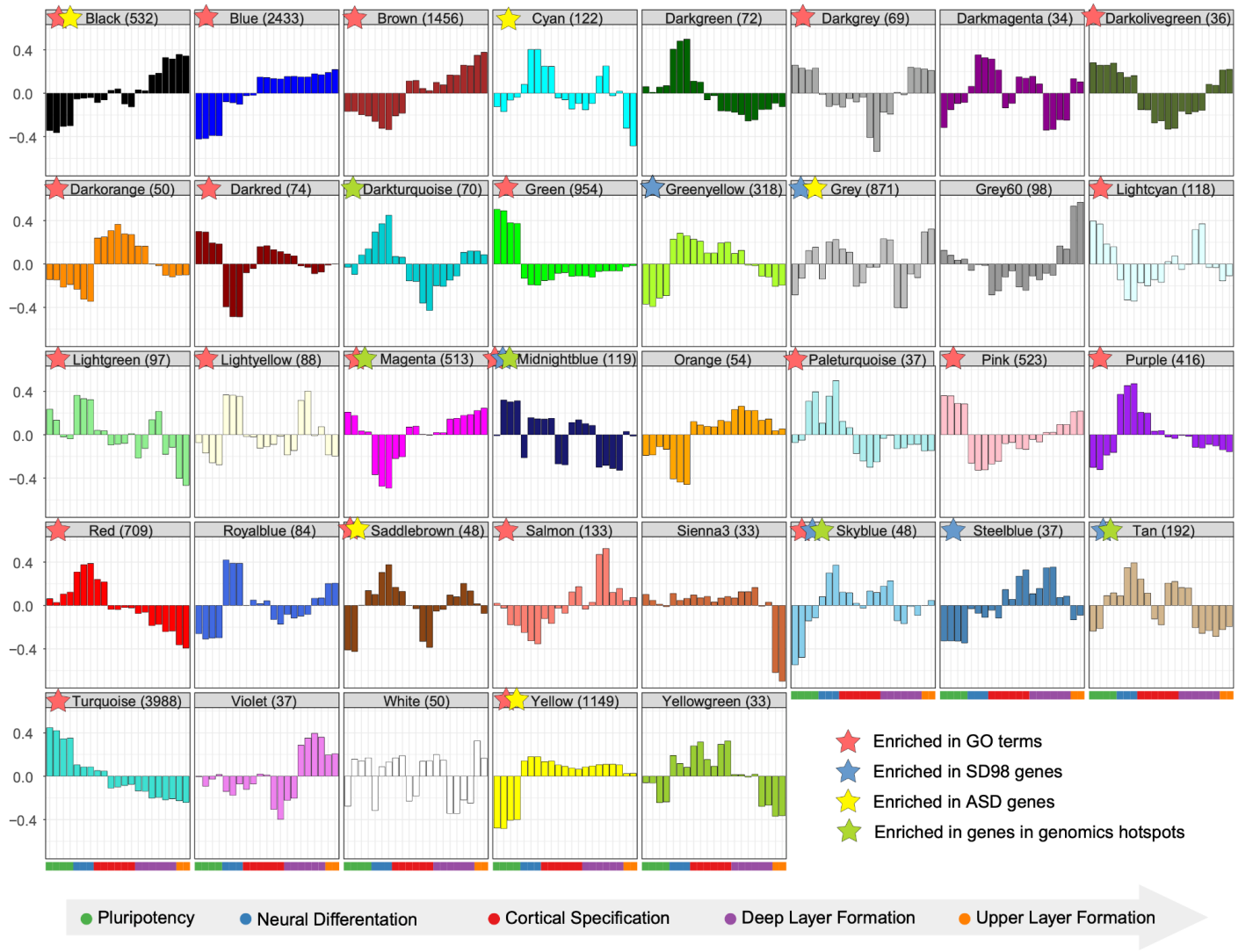

Figure S9. Module eigengenes for 37 modules obtained from WGCNA constructed from 15,695 genes after expression and quality control filters. Each module is represented by a randomly assigned color stated above each plot. Numbers in parentheses represent the total number of genes assigned to the module. Stars represent modules enriched on different gene categories, including GO terms (red), SD98 genes (blue), autism-associated (ASD) genes (yellow) and genomic hotspots from Sattertrom et al. [11] (green). Colored bars at the bottom indicate different developmental stages as described in van der Leemput [12].

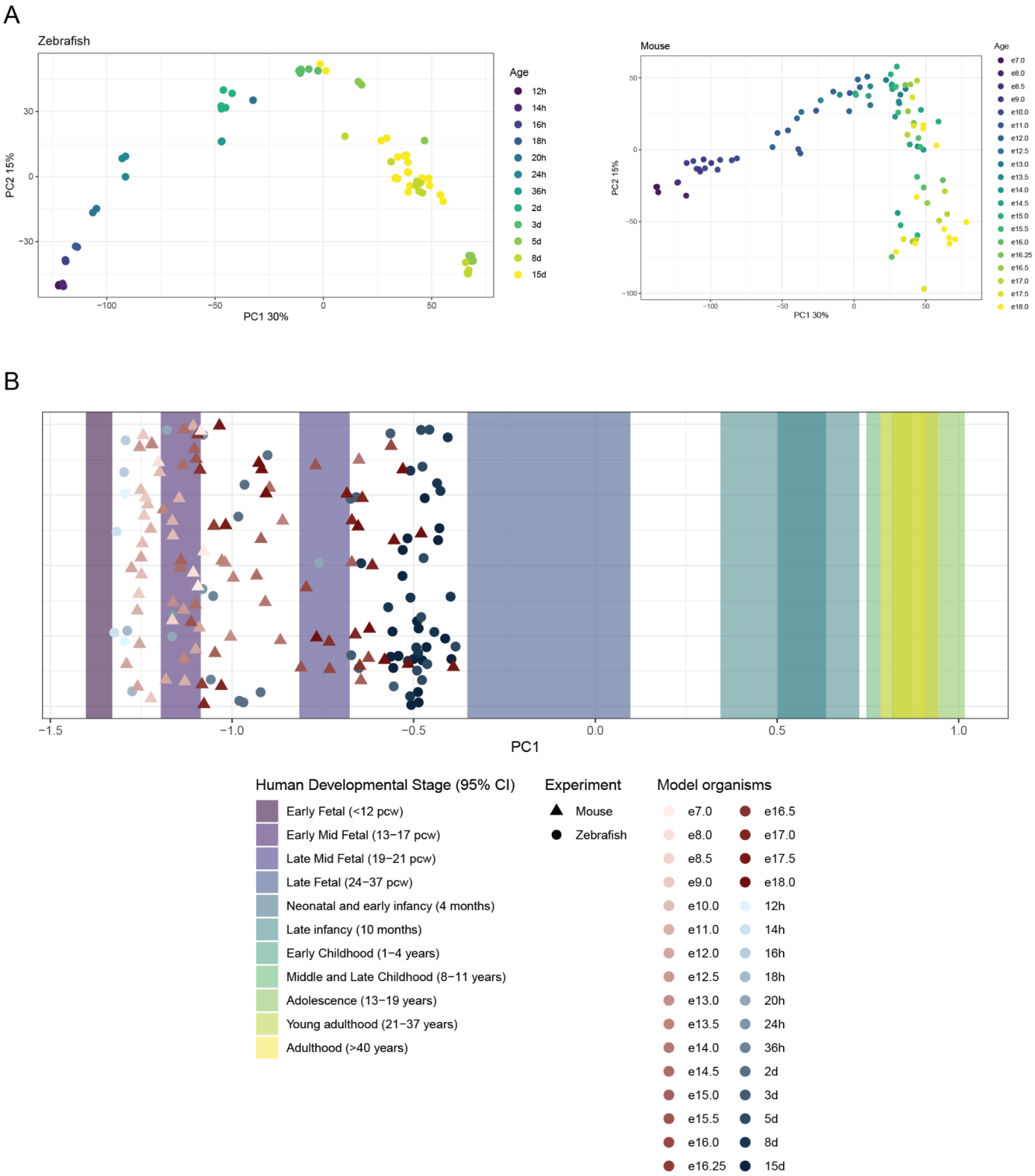

Figure S10. Matched neurodevelopment staging of human, mouse, and zebrafish using single-cell transcriptomic data.
(**A**) Principal component analysis of brain single-cell RNA-sequencing samples from zebrafish [13] and mouse [14] across development.
(**B**) Matching of zebrafish and mouse samples to human developmental stages from the BrainSpan data.

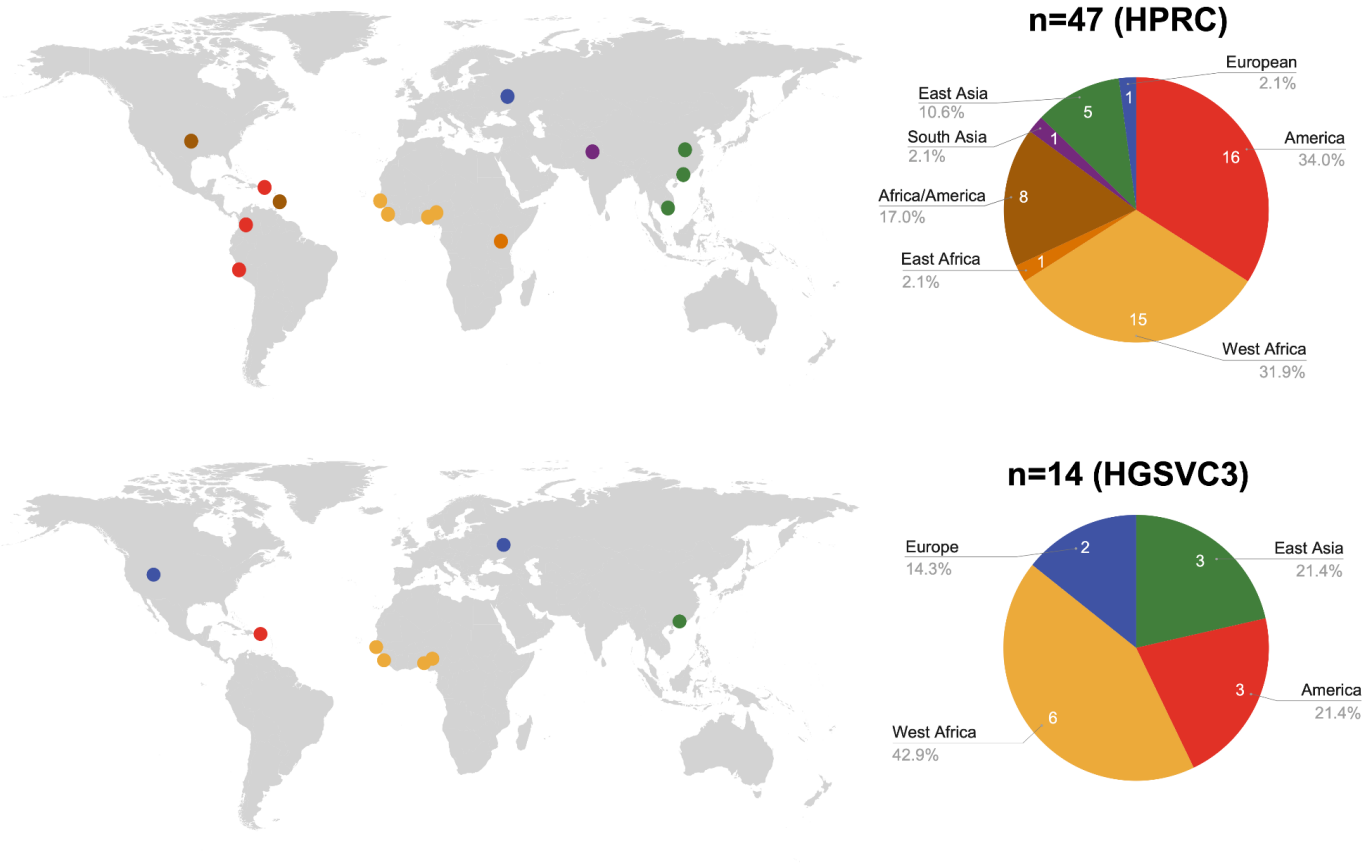

Figure S11. Counts of draft diploid assemblies of the Human Pangenome Reference Consortium (HPRC) and Human Genome Structural Variation Consortium (HGSVC) per super-population. World maps represent sample sites for each ancestry.

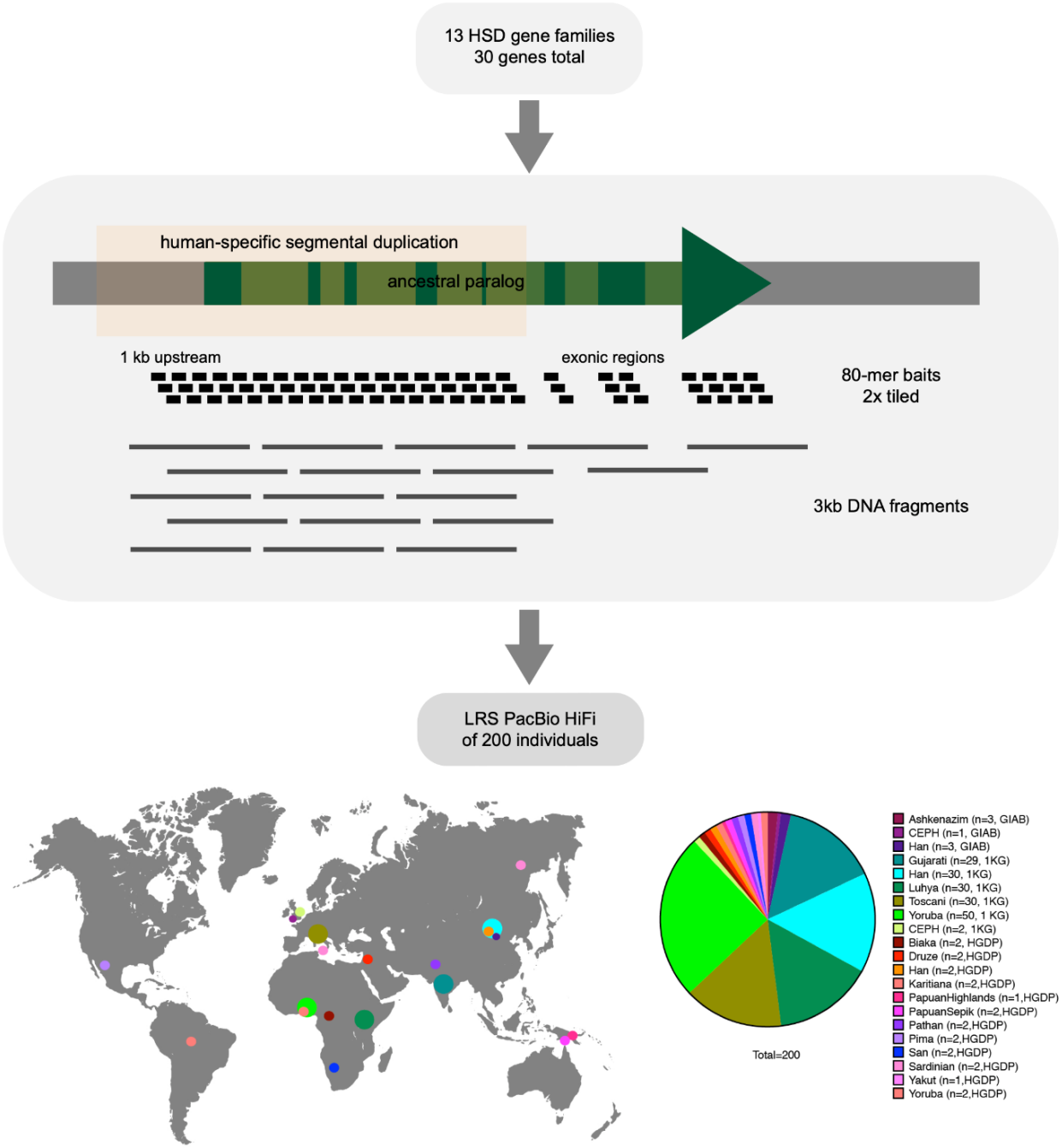

Figure S12. Capture strategy (Ren-Seq) followed by PacBio high-fidelity long-read sequencing. World map shows representative sample sites for each ancestry with legend and total sample counts on the right (n=200; n=144 unrelated).

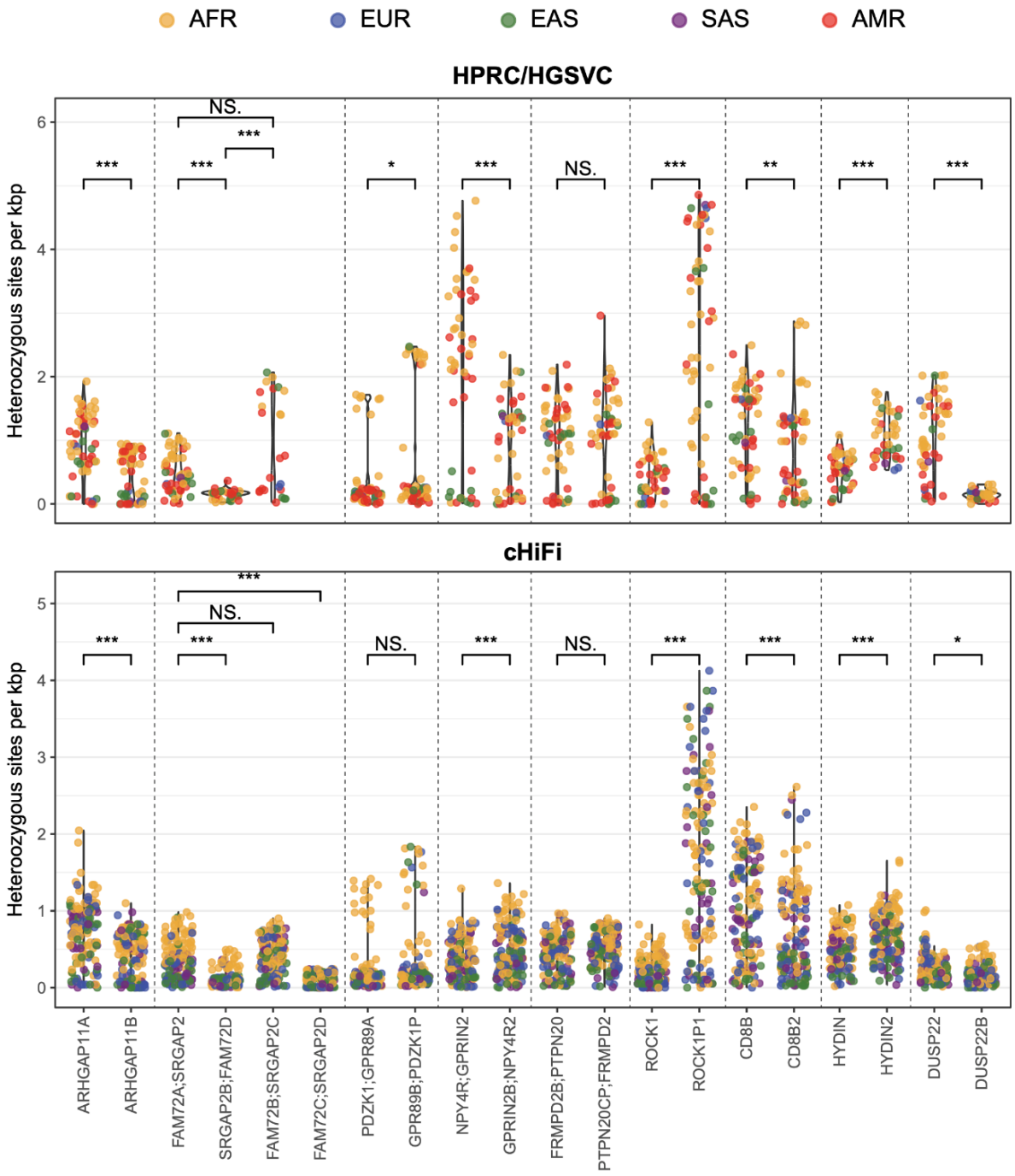

Figure S13. Heterozygous-site densities across duplicated portions of pHSD captured loci for HPRC and HGSVC samples (n=47) and non-redundant unrelated cHiFi individuals (n=144).

Figure S14. Human genetic variation landscape across *SRGAP2C* locus. The outlier Tajima's D value was derived from 1KGP SNVs, and Tajima's D plots derived from HPRC and HGSVC SNVs using 6-kbp windows and 500-bp steps for African (AFR), American (AMR) and all samples (green).

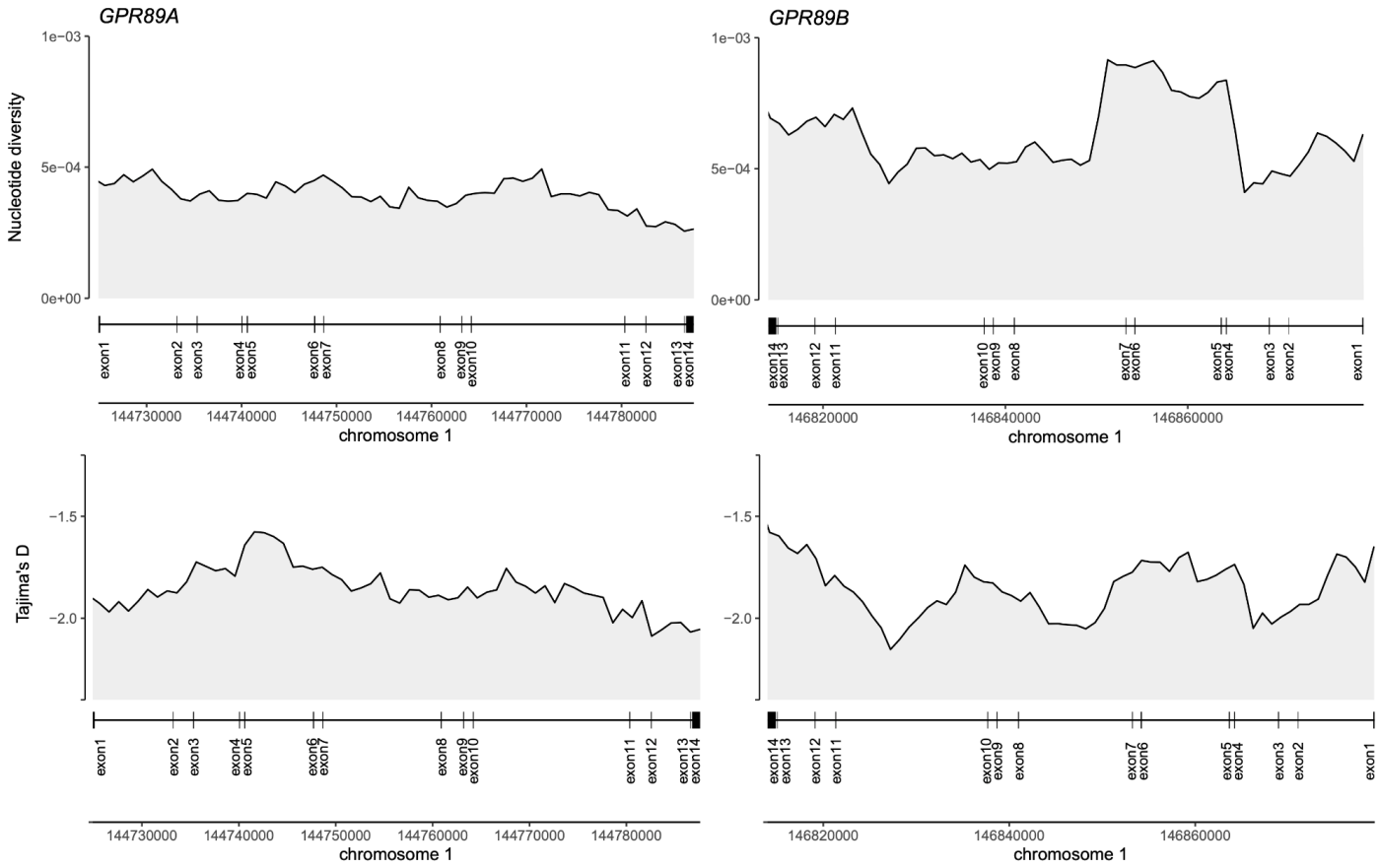

Figure S15. Human genetic variation landscape across *GPR89* paralogs. Nucleotide diversity (top) and Tajima's D (bottom) across corresponding duplicated exons between paralogs *GPR89A* and *GPR89B*, calculated in 15-kbp sliding windows with 1-kbp steps using HPRC/HGSVC variants.

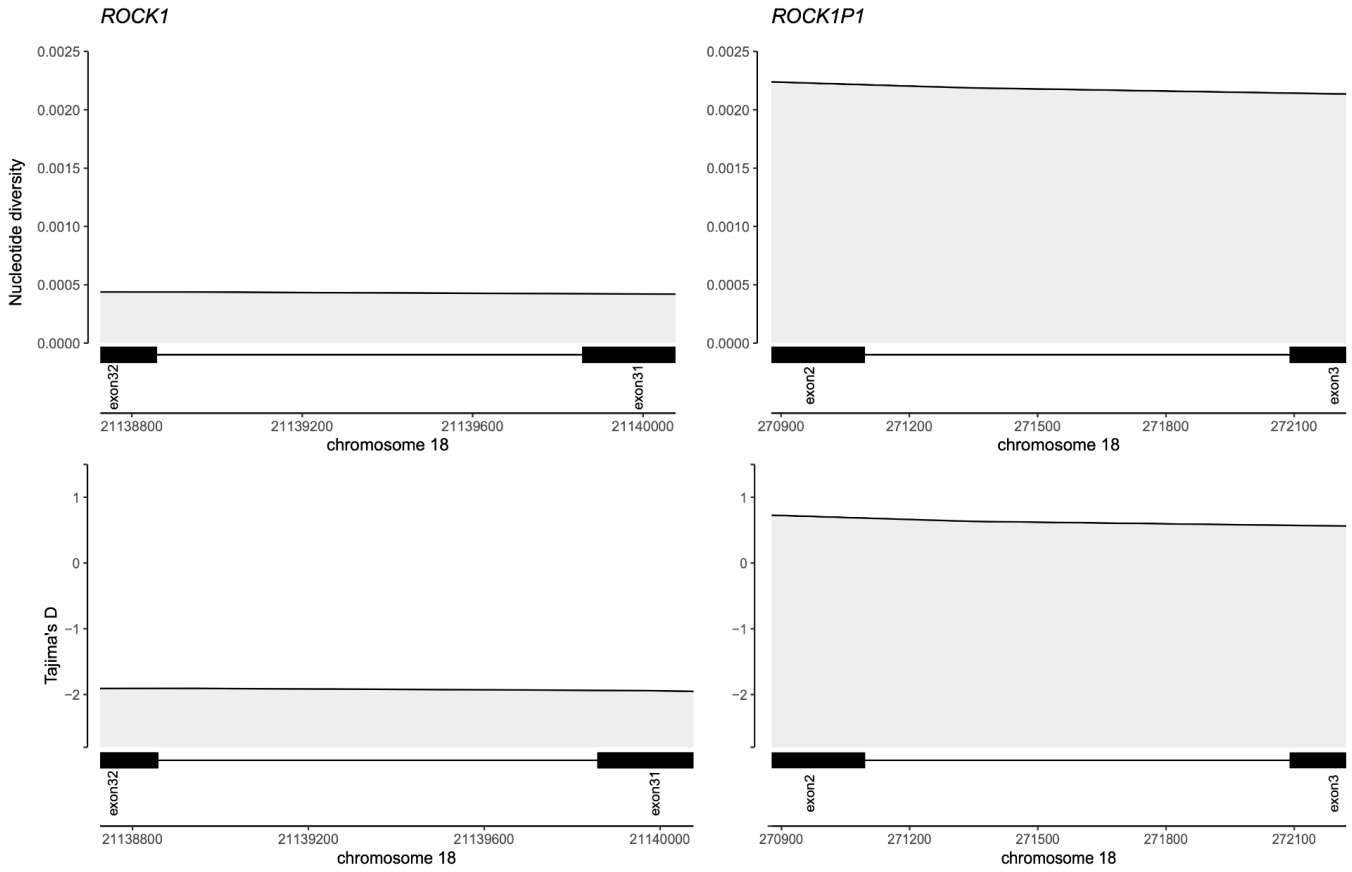

Figure S16. Human genetic variation landscape of *ROCK1* paralogs. Nucleotide diversity (top) and Tajima's D (bottom) across corresponding duplicated exons between paralogs *ROCK1* and *ROCK1P1*, calculated in 15-kbp sliding windows with 1-kbp steps using HPRC/HGSVC variants.

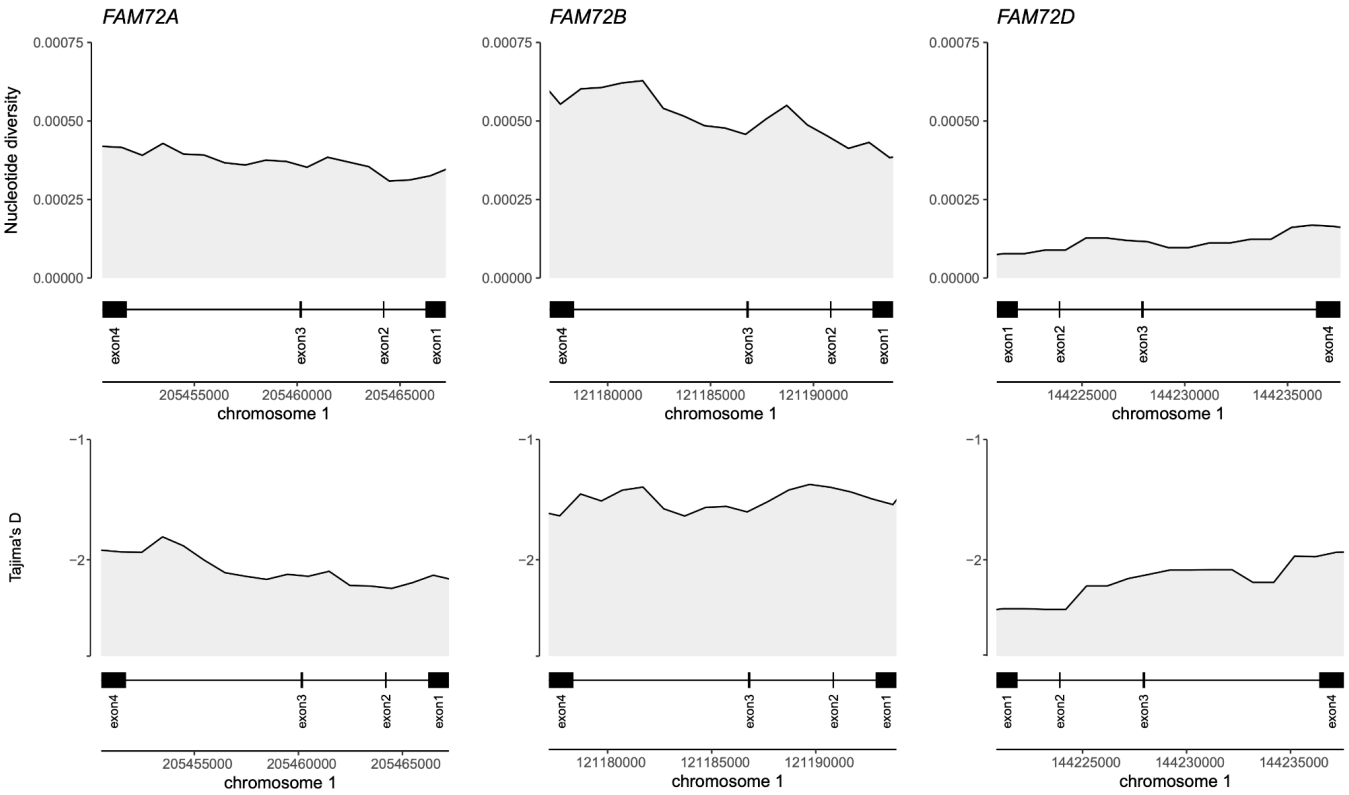

Figure S17. Human genetic variation landscape of *FAM72* paralogs. Nucleotide diversity (top) and Tajima's D (bottom) across corresponding duplicated exons between paralogs *FAM72A*, *FAM72B* and *FAM72D*, calculated in 15-kbp sliding windows with 1-kbp steps using HPRC/HGSVC variants.

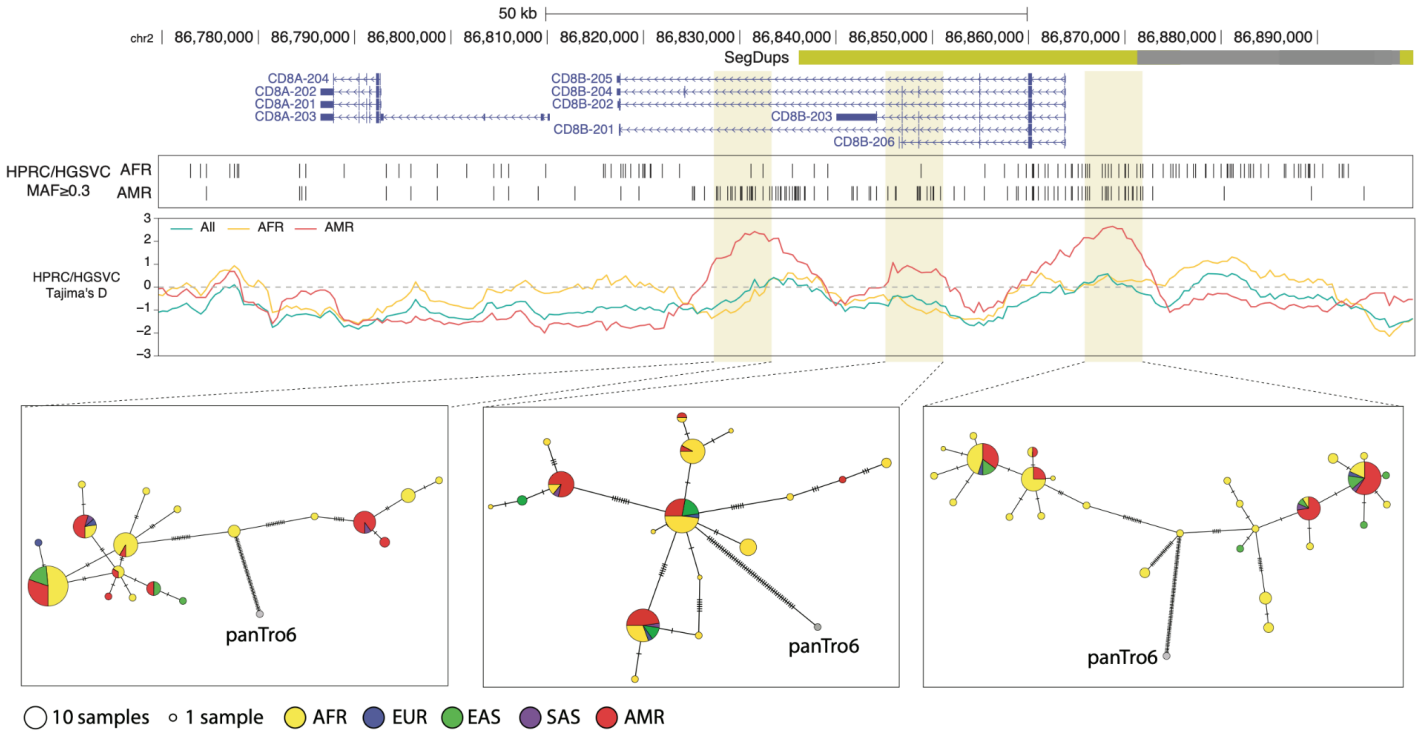

Figure S18. Human genetic variation landscape of *CD8B*. The *CD8B* locus in T2T-CHM13v1.0 reference genome in the UCSC browser, with HPRC/HGSVC intermediate allele frequency variants for African (AFR) and American (AMR) individuals, and derived Tajima's D values calculated in 6-kbp windows with 500-bp 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, with populations colors and scale explained in legend.

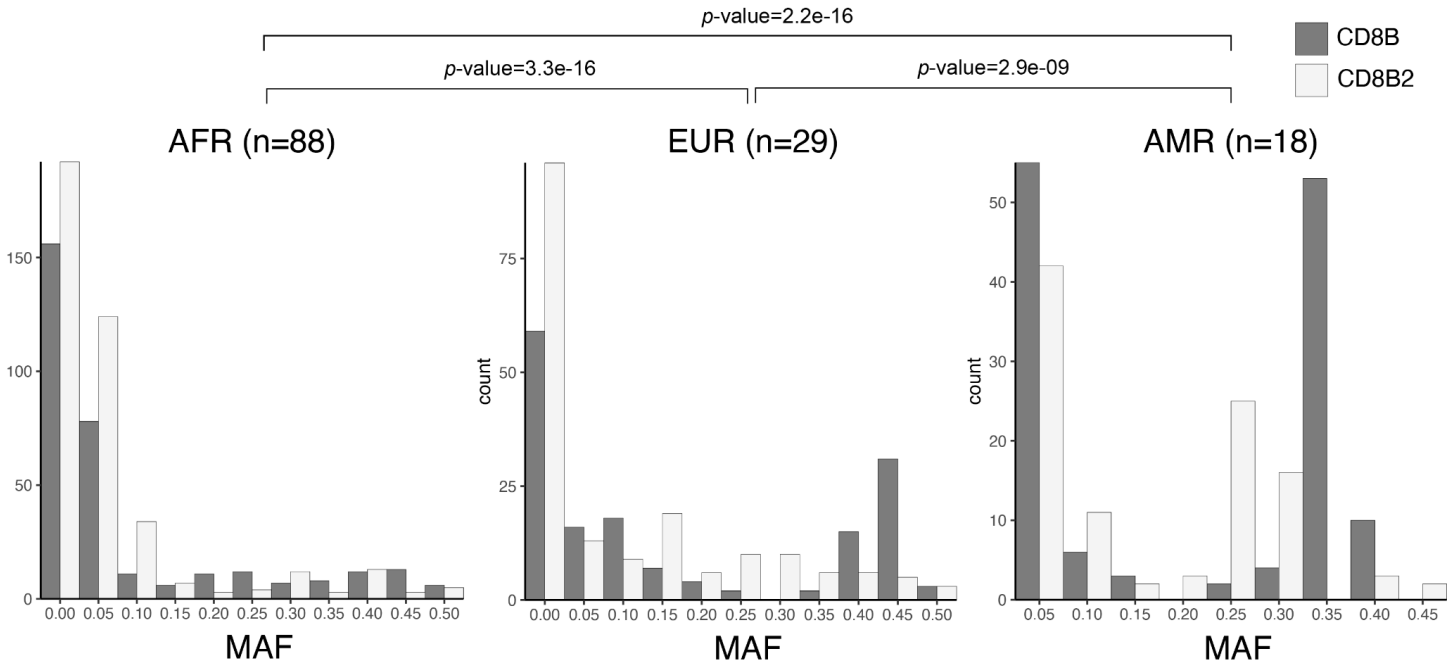

Figure S19. Folded Site frequency spectrum of the *CD8B* and *CD8B2* loci. Minor allele frequency (MAF) calculated across 35-kbp regions overlapping CD8B (light gray) and CD8B2 (dark gray) from variants detected in the combined dataset including long-read assemblies and capture PacBio HiFi sequencing from individuals of African ancestry (AFR, n=88 individuals), European ancestry (EUR, n=29 individuals), and American ancestry (AMR, n=18 individuals). Three individuals of European ancestry (NA20582, NA20525, NA20542) were excluded from this analysis. *P*-values were obtained comparing *CD8B* MAF distribution between populations using Kolmogorov-Smirnov test.

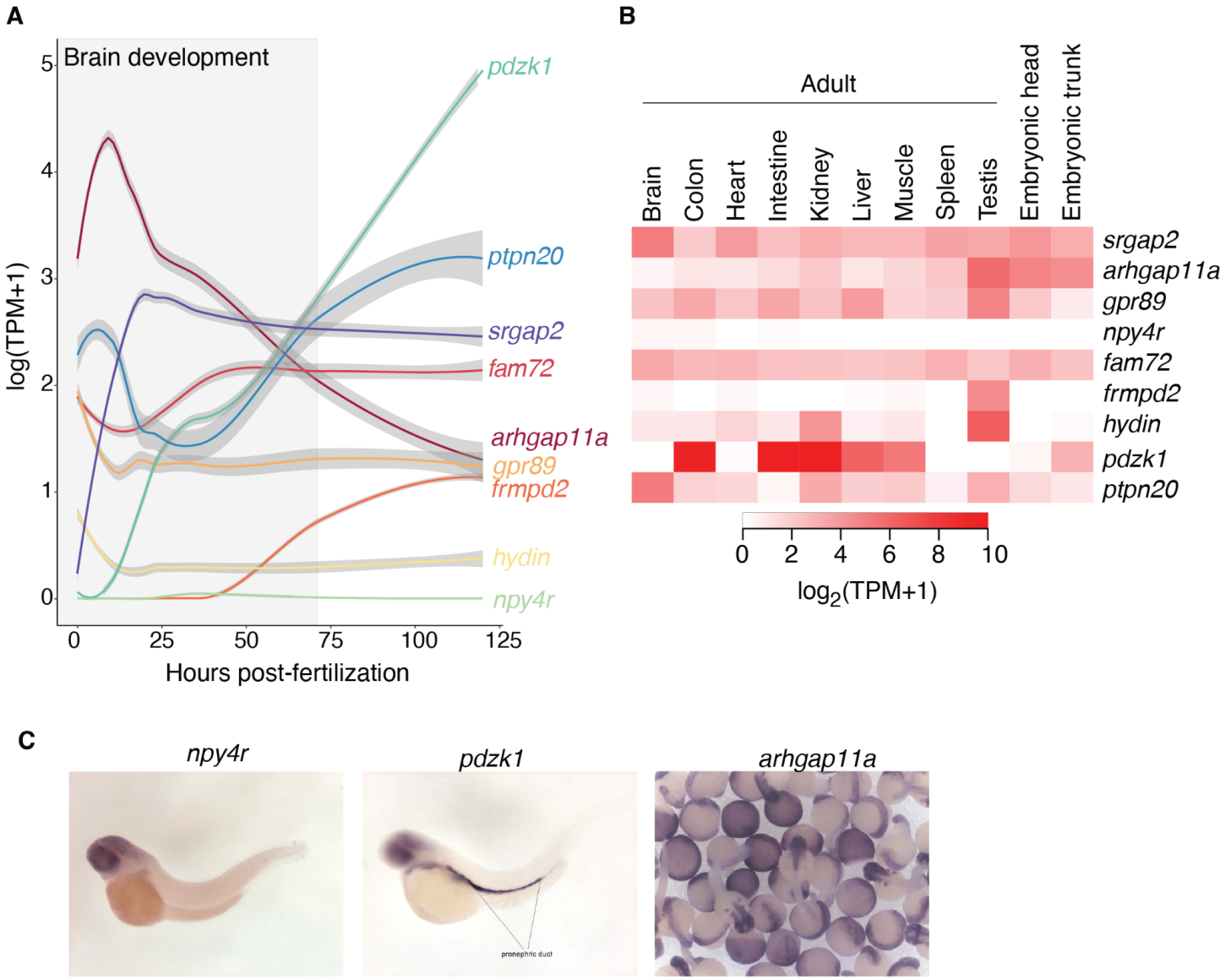

Figure S20. Endogenous gene expression of pHSD zebrafish orthologs during development. (A) Temporal expression between 0 and 120 hours post-fertilization using published data [15] of the zebrafish orthologs of the selected pHSDs. Shaded area corresponds to the brain development period in zebrafish embryos [16] of the zebrafish orthologs of the selected pHSDs. **(B)** Expression of the selected genes in embryonic or adult tissues (data from [17]). **(C)** Available expression patterns via *in situ* hybridization in the Zebrafish Information Network (ZFIN) [18].

***frmpd2* stable knockout allele (5 bp deletion):**

ATGAGCACGTTTGTGACCCTGGCCGAGGTGCTGGAGTCACGAGCAGCACCGCTGGAGGAGGA[CGAGG]TCT
GGGCTCTGCTCTTGGGTGCCACAGAAGCTCTGATAGACATCTCCAGTAAAGATGATGGGAACATGTGCTGT
GTGATCAGCCCCGGCTCCATGCTCCTCTCTGCAGTCGGCAGTATTGCCTTCAAGACCTGCAGCCGATCTGAA
GACGTGGGTTCATTCACTTGCCCAGAAATGTCTCAGAGCAACACTTCCTCCAGGAGACTGGCCTCAGAGAA
AATGGTGGTTTACTCACTAGGCATGACTCTCTATTGGGCAGTGGATTATAATCTTCCACAAAATCAGCCTGT
CCAGGTCAGCGACACCCTGAACAGTCTTCTACTGAGCATGTGCGAGGATGTGGCTCACAGACGGGTGAACC
TCCTGACGGTGCTGGAGACGGCTGAAAATCACCATCAGACTGCAGTCCTGCCGCGACCAGAGAAAACTATC
AGACAGATGGCAGAGGACGTACTTCAGACTGAAAAACCCTCAAATGATTCTGCTCATCTCATAGACCGAAG
CCAAATGGTTAGAGAAAGACTTCGGGGTGCTTCCAGTCAGAACTCAGTTTGGACACAGAAAGGGAACAGCA
CACCGTCAGCAGCAGACTTCAGCAGATTTCACAGTTCACACCAGAGAAGCAGAAACAGCACATATATGAGC
CAAAGTCCTGTTCTTGACCGACGCCGTCAGTTCGCCTTGAGGCCAGATCAGCTGTTTGCCTCTTCTTTTAGCC
TTAATGAACGGAAAATGAAGGATACGGGCCCAGAGTTCATCCGGATGCTGGAGGAGCCTCTTGTCGTTTTA
GAGCTGCCTAATTCTATTGTGTCAAACAAAGGGAAGTCAAGTTCAAACCAAAGGGACTTGACGGTTTTGAT
GCCAAACGGACAGAACGTTCTGCTCAAGTGCGATGTAAAGTGCAGAGGACGCGACGCATTCGATATGATCG
TCGCTCACACAAATCTCGTCGAGCATTTCTACTTCGGCCTTGCATATAGTTATGATAATGAGTTTTTCTTCCT
CGACAATGACACAAAAATATCCAAAGTTGCCCCAGGAGACTGGAAAAAAGTGCCCACTGCCACATTCGTGC
TTCACTTCCGCATTAAATATTTTGTGAGCGACGTATCTCTCCTTCTAAACAAGTTCACCCGTCATCAGTTCTA
CCTGCAGCTCAGACGTGATCTTCTGGACGATCGGCTGCAGTGTAATGAGGAGACCTGCCTGTTTCTCGGTGC
CTTGGCTCTTCAGG

CTGAGTTTGGAGACAGTTTGCCTGAGGTCTATGGGAAGAACTACTATCAGCCTGAGCATTATGTTTCCAAGA
GTGTTGTACAGAAAATGGCAATGCCATGCCTGAAAGAAGAGCTGCTGCGTTTACACGTCAACAATGCCAAC
ATGAGCGCAGAGGAGGCCGAACTCGAGTTCCTCAAGAGCGTTCAGCAGCTGCCGGAGTATGGTGTGCTGTT
TCACCGCGTGGCCCGAGAGAAGAAGCCTGTGTTTGGAGAGCTTCTGCTGGGAGTCTGTGCCAAGTGCATCA
CTGTGTATGAAGTTATTAATAACTGCCGCACCGCAAGCCTTACATTCCATTGGAGAGACACGAGCAGCATC
ACCTCAACTCGGCGCAAGTTCATCATTGAATGCAGCACCAGCAAAAAGAAGCACGTCTTTCTGACCGAAAA
GTCTAAAGTTGCCAAATACCTGTGCGACCTCTGCTCCGCACAGCACAAGTTTCACAAAGAGATGAGCTCTCG
G

CAGCTCACGCACAGCCTGGCTTCAGAGGAGAGTATTGTTCAGTACGCAGCCCTGTGCAGAGCTCAGAATAA
TGAGGTGAACGACAGCAGCGCTGACGAATCCATGAGCAAACTCTGTGAAGACATCGCCACCCGGATCGAG
GCCAAGATCAAACTACAGCGAGACTTACTGGACAACACCGGATCCCAAAGTCCAATTCTACAGAGAAGCCT
TTGTAGTACTCCAAAGCGCGGCTCTGACGTCCCTTCAGTGTCTTCGCCTCTAAGAGACACGCCAACAGGTGC
CCGTACTCCAGAGAGGGAAATCATCTGCGTTTCCCTTAAAAAAGACCCTAAAGTCGGACTTGGTATTGTCAT
AGTTGGAGAAGACACCGCTGGGAAATTGGACCTCGGGATCTTCATCGCTTCAGTTGTACCTGGAAGTCCAG
CAGATCGAGATGGACGCATCAAACCAGGTGGTCGTCTGATCTCTCTGAATCAGCTCAGTCTGGAGGGCATG
TCGTTTAGTGAAGCCGCTGACATCATGCAAAGCAGCTCCAATGACGTGGAGCTCATCATTTCACAGCCCAA
AGCTCAGTGCAAGCGTGGAGGATCTGTTTCTCTGAACGAGCGCAGCTACGAGTCTCAGAGCACAATCCTGG
CAGACGGCAGAACCGGAGACGAGTTTCTGGACGAGCTGGTCAGTGTCATGATGACCCCGAAAGCCTCGAAC
AGACTGCACGTTCCTGAGGTCCGCATCATCAACGCACAGGATGATTATTCCAGATCAGCGTCTCTGATCAGC
CTGAGGCCAGAAGAGTTTACAGTGACGCTCATGAAGTCTGGAGGCAGTCTGGGCATCAGCATTGCCGGAGG
AGTGAACACAGGTCTCCGCTACGGTGGAATTTACATCAAGAGTCTGGTGTCAGGTGGCGTGGCGGAACAAG
ACGGACGAATACAGACAGGAGACAGACTGTTGGAAGTGGACGGCATCAGACTTCAGGGATTCACAGATCA
GCAGGCGGCTGAATGTTTGGCCAGAACAGGCGAGGTTGTGGGTCTGGTTCTGGAGCGAGATGGCGGCTCTA
TGCTTCAGCAAGGGCCTGGAAGCCCTCAACTACGGAACACTCTCTCCATCTGCACACCCGGTCCTGCTGGAA
CACAGGCCAGAA

AGAGCTGTCCTGCCATCACCATGACCAGACCCTTCAACATCAAGCCCAGAGACTACAGCTTTGTGTCGGAC
GGTAATAAAAAACAACATTCAGGAGGTGACGCTTAA

***gpr89* stable knockout allele (8 bp deletion):**

ATGTCTTTCTTCGTAGATTCAGTGATTATGTTCACTTCTCAGGTGCTGTTTTTTGGATTTGGGTGGCTGTTTTT
CATGCGTCAGTTGTTTAAAGATTATGAGGTTCGGCAGTATGTGGTGCAGGTGGTGTTCTCCATCACGTTCGC
CTTCTCTTGCACAATGTTTGAGCTCATTATCTTTGAGATCCTCGGTGCATTGAGCAGCTCGTCCAGGTATTTC
CACTGGAAGCTGAATTTGTATGTAATATTGCTGGTTCTGATATTTGTGGTGCCTTTCTACATTGGCTACTTTG
TGGTCAGTAATATACGTTTATTGCAGAGACAGAGGTTGCTTTTTTCATGTGTGGTCTGGTTTACATTCATGTA
TTTCTTTTGGAAACTGGGCGATCCTTTTCCTATACTCAGTCCCAAACATGGTATTCTGTCCATTGAGCAGCTG
ATCAGTCGCGTGGGGGTCATTGGGGTCACTCTAATGGCTCTACTGTCCGGTTTTGGTGCTGTGAACTGTCCG
TACACATACATGTCATATTTCCTAAGGAATGTGACAGACAGTGATATCTTGGCTCTTGAGAGAAGACTCCTA
CAAACTATGGATATGATTGTCAGTAAAAAGAAAAGGATTGCCATGACAAGAAGGCAGAT[GTACCAGC]GA
GGAGAAGAGCAGAATAAACAGACAGGATTCTGGGGGATGATCAAGAGTGTGACCTCTTCACCATCAGGCA
GTGAGAATCTGTCTCTGATCCAGCAGGAAGTGGATGCTCTGGAGGAACTCAGCAGACAGCTTTTCTTGGAG
ACTGTAGATCTGCAGGCAACCAAGGAGCGGATAGAATACTCAAAAACATTTCAAGGGAAATACTTTAACTT
CTTAGGGTATTTCTTCTCCATCTACTGTGTGTGGAAAATATTCATGGCCACTATAAACATAGTGTTTGACCGT
GTGGGGAAGACTGACCCGGTGACGAGGGGAATCGAGATCACCGTCAACTATCTTGGAATTCAGTTTGATGT
CAAGTTCTGGTCTCAGCACATTTCCTTTATTCTGGTGGGAATCATTATAGTCACATCCATACGAGGCCTTTTA
ATCACACTCACCAAGTTTTT
CTACGCCATCTCCAGCAGCAAGTCCTCCAATGTTATTGTGCTCGTCTTGGCTCAGATCATGGGCATGTATTTT
GTGTCGTCTGTTCTGCTGATGCGGATGAGCATGCCGCTGGAGTACCGCAGTATTGTGTCAGAGGTGTTGGGT
GAACTGCAGTTTAACTTCTACCACCGCTGGTTCGACGTGATCTTTTTGGTCAGCGCTCTCTCCAGCATCCTCT
TCCTCTACCTGGCACATAAACAGGCACCCGAGAAGCACATGGCCCTGTGA

Figure S21. Alleles in the stable knockout lines for *frmpd2* and *gpr89* with the deleted bases highlighted.

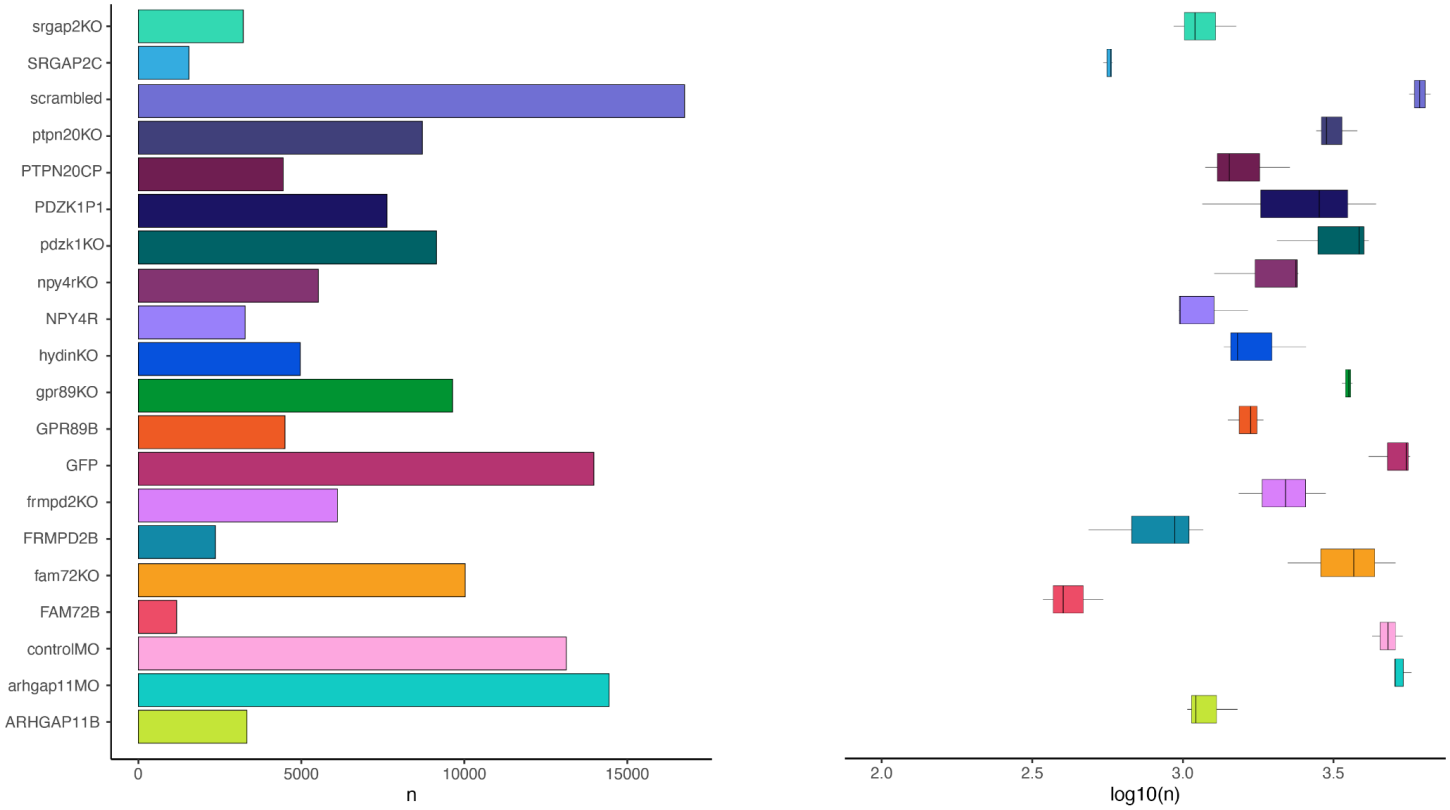

Figure S22. Description of the number of cells per zebrafish mutant model used for single-cell transcriptomic analysis.

Supplemental References

1. 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.

2. 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.

3. 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.

4. 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. (2021). A complete reference genome improves analysis of human genetic variation. bioRxiv, 2021.07.12.452063. https://doi.org/10.1101/2021.07.12.452063.

5. 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.

6. 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.

7. 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.

8. 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.

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

10. 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.

11. 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.

12. 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.

13. Raj, B., Farrell, J.A., Liu, J., El Kholtei, J., Carte, A.N., Navajas Acedo, J., Du, L.Y., McKenna, A., Relić, Đ., Leslie, J.M., et al. (2020). Emergence of Neuronal Diversity during Vertebrate Brain Development. Neuron *108*, 1058–1074.e6.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.