Genetic and environmental contributions to ancestry differences in gene expression in the human brain

Kynon J.M. Benjamin[1-3]*, Qiang Chen[1], Nicholas J. Eagles[1], Louise A. Huuki-Myers[1], Leonardo Collado-Torres[1,4], Joshua M. Stolz[1], Geo Pertea[1], Joo Heon Shin[1], Apuã C.M. Paquola[1,2], Thomas M. Hyde[1-3], Joel E. Kleinman[1,3], Andrew E. Jaffe[3,5,6], Shizhong Han[1,3,7]*, and Daniel R. Weinberger[1-3,5,7]*

Affiliations:

[1]Lieber Institute for Brain Development, Baltimore, MD, USA

[2]Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[3]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[4]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA

[5]Department of Neuroscience, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[6]Neumora Therapeutics, Watertown, MA, USA

[7]Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA

*Corresponding authors

Summary: We examine the impact of genetic ancestry on gene expression and DNA methylation of admixed African/Black Americans, highlighting how genetic and environmental background affect risk for brain illness.

Abstract:

Ancestral differences in genomic variation are determining factors in gene regulation; however, most gene expression studies have been limited to European ancestry samples or adjusted for ancestry to identify ancestry-independent associations. We instead examined the impact of genetic ancestry on gene expression and DNA methylation (DNAm) in admixed African/Black American neurotypical individuals to untangle effects of genetic and environmental factors. Ancestry-associated differentially expressed genes (DEGs), transcripts, and gene networks, while notably not implicating neurons, are enriched for genes related to immune response and vascular tissue and explain up to 26% of heritability for ischemic stroke, 27% of heritability for Parkinson's disease, and 30% of heritability for Alzhemier's disease. Ancestry-associated DEGs also show general enrichment for heritability of diverse immune-related traits but depletion for psychiatric-related traits. The cell-type enrichments

and direction of effects vary by brain region. These DEGs are less evolutionarily constrained and are largely explained by genetic variations; roughly 15% are predicted by DNAm variation implicating environmental exposures. We also compared Black and White Americans, confirming most of these ancestry-associated DEGs. Our results highlight how environment and genetic background affect genetic ancestry differences in gene expression in the human brain and affect risk for brain illness.

Introduction

Health disparities have endured for centuries (*1*). In neuroscience and genomics, individuals with recent African genetic ancestry (AA) account for less than 5% of large-scale research cohorts for brain disorders but are 20% more likely to experience a major mental health crisis (*2*, *3*). Insights gained from genome-wide association studies (GWAS) about disease risk are promising for clinical applications (e.g., drug targets for novel therapeutics and polygenic risk prediction). However, the majority of GWAS of brain-related illness lack diversity with regards to inclusion of AA individuals, who account for less than 5% of GWAS participants (*4*), despite AA individuals having more extensive genetic variation than any other population. This lack of diversity limits the accuracy of genetic risk prediction, hinders the development of effective personalized neurotherapeutics for non-European genetic ancestry (EA) individuals (*5*), and limits our potential for novel discovery. While diversity in large-scale GWAS has increased in recent years (e.g., 1000 Genomes Project (*6*), All of Us research program, Trans-Omics for Precision Medicine [TOPMed] (*7*), and Human Heredity and Health in Africa [H3Africa] Consortium (*8*)), population-based genetic association studies do not directly elucidate potential biological mechanisms of risk variants.

To bridge this gap, we need studies of the biological impact of genetic variation on molecular traits (e.g., mRNA and DNA methylation [DNAm]) in disease-relevant tissues of diverse populations. Recent efforts to bridge this gap with cross-ancestry expression quantitative trait loci (eQTL) have focused on improved fine mapping while leaving unanswered the question of how gene expression and epigenetic regulation are parsed specifically by ancestry (*9*). Despite a clear urgent need, no large-scale studies examine the biological impact of genetic ancestry on gene expression in the human brain focused on the differences between AA and EA.

An obvious impediment to undertaking this task is the limited availability of brain tissue from AA individuals. Currently, the most widely used resource for human postmortem tissue is the Gene-Tissue Expression Project (GTEx), which has publicly available RNA-sequencing and single nucleotide polymorphisms (SNP) genotyping for nearly 1,000 mostly elderly individuals, including data from 13 brain regions (114 to 209 individuals per region). However, the majority of GTEx brain samples are of EA, and for some brain regions, GTEx has no non-EA individuals. In comparison, the BrainSeq Consortium, a collaboration between seven pharmaceutical companies and the Lieber Institute for Brain Development (LIBD), has one of the largest postmortem brain collections of psychiatric disorders, including 784 Black American samples across 587 unique individuals, with a mean age of 44. While reports from this consortium and other large-scale analyses in the brain – including from the hippocampus, caudate nucleus ("caudate"), dorsolateral prefrontal cortex (DLPFC), and granule cells of the dentate gyrus ("dentate gyrus") – have samples of diverse genetic ancestry (*10–16*), they have typically been "adjusted" for ancestry status, which limits our understanding of ancestry-specific effects in the brain.

To address these gaps, here we use the LIBD RNA-sequencing, SNP genotype, and whole genome bisulfite sequencing (WGBS) datasets to evaluate genetic and environmental contributions to genetic ancestry differences in gene expression in the human brain (**Fig. 1**). We identify transcriptional features associated with genetic ancestry (African or European) in admixed neurotypical Black American donors (n=151). We quantify the contributions of common genetic variations to genetic ancestry differences using a total of 425 samples, including the caudate (n=122), dentate gyrus (n=47), DLPFC (n=123), and hippocampus (n=133). Additionally, we examine the influence of genetic ancestry on DNAm using WGBS data of the admixed Black American donors from the

caudate (n=89), DLPFC (n=69), and hippocampus (n=69). To confirm the genetic ancestry-associated differences in gene expression and to highlight the effect of environment by genetic ancestry differences, we further examine transcriptional and DNAm differences in individuals of limited admixture (Black Americans ≥ 0.8 AA and White Americans > 0.99 EA).

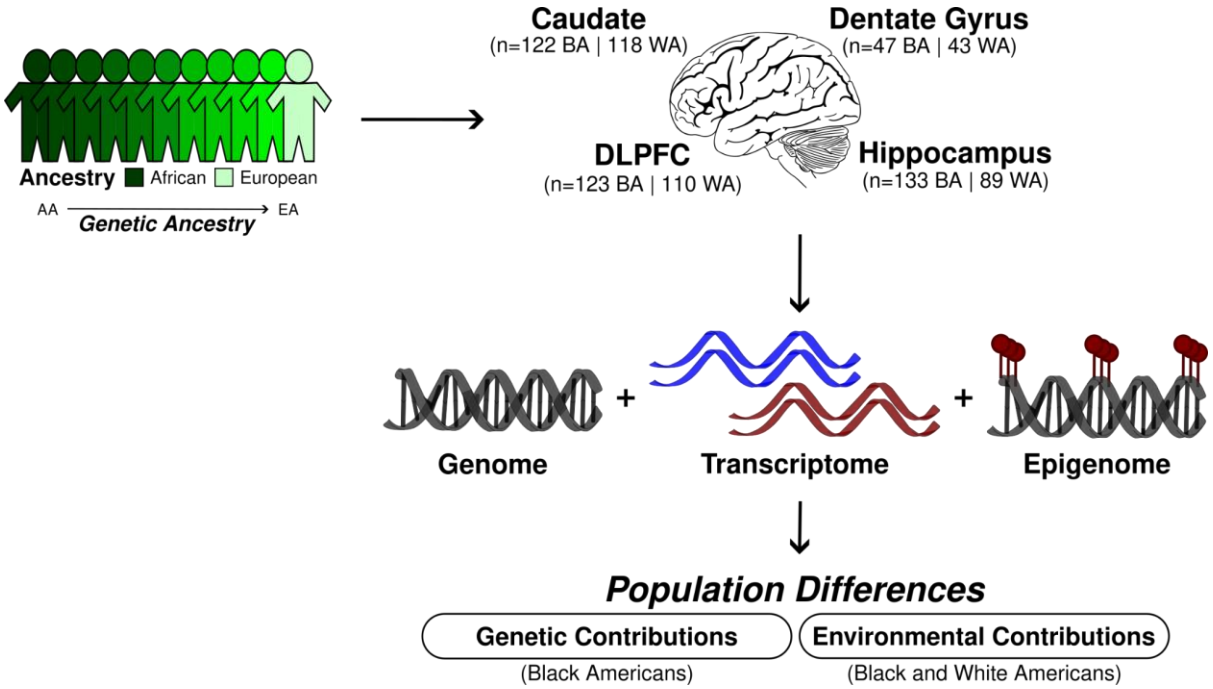

Fig. 1: Study design for the examination of the genetic and environmental contributions to genetic ancestry-associated expression differences. BA stands for Black Americans and WA for White Americans.

Results

Significant enrichment of immune response for differential expression associated with genetic ancestry across the brain

We selectively examined our admixed Black American population (151 unique individuals; **Table S1**) to 1) characterize transcriptional changes associated with African or European genetic ancestry in neurotypical adults (age > 17) and 2) limit potential confounding effects of systematic environmental factors that may differ between Black and White American samples. These analyses included RNA sequencing data from caudate (n=122), dentate gyrus (n=47), DLPFC (n=123), and hippocampus (n=133). Our admixed Black American population showed a varied proportion of EA (STRUCTURE (*17*); EA mean = 0.21, range = 0-0.62; **Fig. S1**) consistent with previous reports (*18*, *19*). As such, we used these continuous genetic ancestry estimates to identify differentially expressed features (genes, transcripts, exons, and junctions) that were linearly correlated with ancestry levels and adjusted for sex, age, and RNA quality. This RNA quality adjustment includes experiment-based RNA degradation metrics (obtained with the qSVA methodology) that account for batch effect and cell composition (*12*, *20*). To increase our power of detection and improve effect size estimates, we applied the multivariate adaptive shrinkage ("mash" (*21*)) method, which leverages the correlation structure of genetic ancestry effects across brain regions (see **Methods** for details).

Of the 16,820 genes tested, we identified 2,570 (15%; 1,437 of which are protein coding) unique differentially expressed genes (DEGs) based on ancestry variation (local false sign rate [lfsr] < 0.05; **Fig. 2A**, **Table S2**, and **Data S1**) across the caudate (n=1,273 DEGs), dentate gyrus (n=997), DLPFC (n=1,075), and hippocampus (n=1,025). While this number increased when we examined differential expression based on local ancestry (9,906 [62% of genes tested]; 6,982 protein coding; **Table S3**) across the caudate (n=6,657 DEGs), dentate gyrus (n=4,154), DLPFC (n=6,148), and hippocampus (n=7,006), effect sizes between global- and local-ancestry DEGs showed significant positive correlations (all Spearman; rho > 0.57, p-value < 0.01; **Fig. S3**) across all brain regions. When examining isoform-level associations (transcripts, exons, and junctions), we found an additional 8,012 unique global ancestry-associated DEGs (lfsr < 0.05; **Fig. S2**, **Table S2**, and **Data S1**) and 6,629 unique local ancestry-associated DEGs (lfsr < 0.05; **Table S3** and **Data S2**) in these Black Americans. Similarly, we found that isoform-level local ancestry DE features showed significant positive correlation in effect sizes compared with global ancestry DE features (**Fig. S3**).

To evaluate the functional aspects of these genetic ancestry-associated DEGs (global and local ancestry), we performed gene set enrichment analysis with the Gene Ontology (GO) and Disease Gene Network (DisGeNET (*22*)) databases for each brain region. It is noteworthy that while there was no enrichment of neuronal gene sets, we observed significant enrichment (GSEA and hypergeometric, q-value < 0.05) for GO and DisGeNET terms primarily related to immune response, including innate, adaptive, and virus responses (**Data S3, Fig. 2B,** and **Fig. S4**). Interestingly, the caudate showed an opposite direction of effect compared with the DLPFC and hippocampus. Specifically, the caudate showed enrichment of immune response associated with DEGs upregulated in relation to AA proportion, while dentate gyrus, DLPFC, and hippocampus showed enrichment for immune-related pathways associated with DEGs upregulated in EA proportion (**Fig. 2B** and **Fig. S5**). While not significant, we observed the same pattern of opposite directionality of effect for immune-related pathways with local ancestry-associated DEGs (**Fig. S6**).

When we expanded this analysis to the isoform level (transcripts, exons, and junctions), we also found significant association with immune-related pathways and similar directions of effect (upregulated for AA proportion in the caudate and upregulated for EA proportion in dentate gyrus, DLPFC, and hippocampus). Furthermore, we also found significant analogous enrichment of these DEGs for genes with population differences in macrophages (*18*) associated with innate immune response to infection (Fisher's exact test, false discovery rate [FDR] < 0.05; **Fig. S7**). Additionally, we found significant enrichment (Fisher's exact test, FDR < 0.01) for ancestry-associated DEGs (global ancestry) in gene coexpression network modules generated using WGCNA (Weighted Gene Co-expression Network Analysis (*23*); **Fig. S8**). Consistent with our DEG analysis, the immune response pathway enrichment in these modules showed analogous opposite direction of effects based on region (**Fig. S9**).

Observing an enrichment of the immune response pathway in bulk tissue, we performed cell-type (*24, 25*) enrichment analysis to evaluate the cellular context of these ancestry-associated DEGs (global ancestry). We found significant enrichment (Fisher's exact test, FDR < 0.05; **Fig. S10** and **Fig. S11A**) for genes specifically expressed in brain immune cells (i.e., glia and microglia cell types) and neurovasculature (i.e., pericyte, endothelial, and vascular tissue cells), but not peripheral immune cells. Additionally, we observed enrichment for distinct subtypes of glial cells (*26*) (**Fig. S12**). Interestingly, local ancestry-associated DEGs showed significant enrichment for brain and non-brain immune cells (Fisher's exact test, FDR < 0.05; **Fig. S13** and **Fig. S11B**) potentially due to the larger number of detected DEGs. Even so, we found the level of enrichment of non-brain immune cells (global and local) on average smaller than brain immune cells. Remarkably, we again found primarily significant depletion of DEGs (global and local) for any genes specific to neuronal cell types.

Consistently, we observed those immune-related pathways and associated cell types (i.e., microglia and perivascular macrophage) for DEGs upregulated with increasing AA proportion in the caudate and upregulated with increasing EA proportion in the dentate gyrus, DLPFC, and hippocampus. Although we found some glial cell subtype (*26*) composition differences (ANOVA, FDR < 0.05; **Fig. S14**) using publically available single cell data from brain regions with similar composition (*27*), no glial subtype (*26*) showed specificity for a specific direction of ancestry effect (**Fig. S12**). Altogether, these results suggest that ancestry-associated DEGs in the human brain are strongly associated with the brain-specific immune response, and specific direction of effects vary according to brain region.

Sharing of genetic ancestry-associated expression differences across the brain

To understand the regional specificity of global ancestry-associated differentially expressed features, we compared DEGs from each brain region and observed extensive sharing across regions. Specifically, we observed 1,210 DEGs (47.1%) shared between at least two brain regions, where all pairwise overlaps demonstrated significant enrichment (Fisher's exact test, p-value < 0.01; **Fig. 2C**). Moreover, 478 DEGs (18.6%) were shared among at least three brain regions with a significant overlap of 112 of these DEGs (4.4%; Monte Carlo simulation, p-value < 1e-5) across all four brain regions.

Interestingly, 27 of the 112 shared DEGs (24%) showed discordant direction of effect in at least one of the four brain regions. This correlated well with the pairwise correlation of shared DEGs that shared direction of effect (70% to 82%; **Fig. 2D).** While shared direction of effect across brain regions was relatively high, this proportion of sharing dropped substantially when effect size was taken into account (0.22 to 0.44; **Fig. 2D**). Corresponding with the large proportion of discordant DEGs, we also found a large number of brain region-specific DEGs (1360 [52.9%]; **Fig. 2E**), which increased when considering isoform-level analysis (transcript [63.6%], exon [67.6%], and junction [69.7%]). This is consistent with other studies that show isoform-level brain region specificity (*28*).

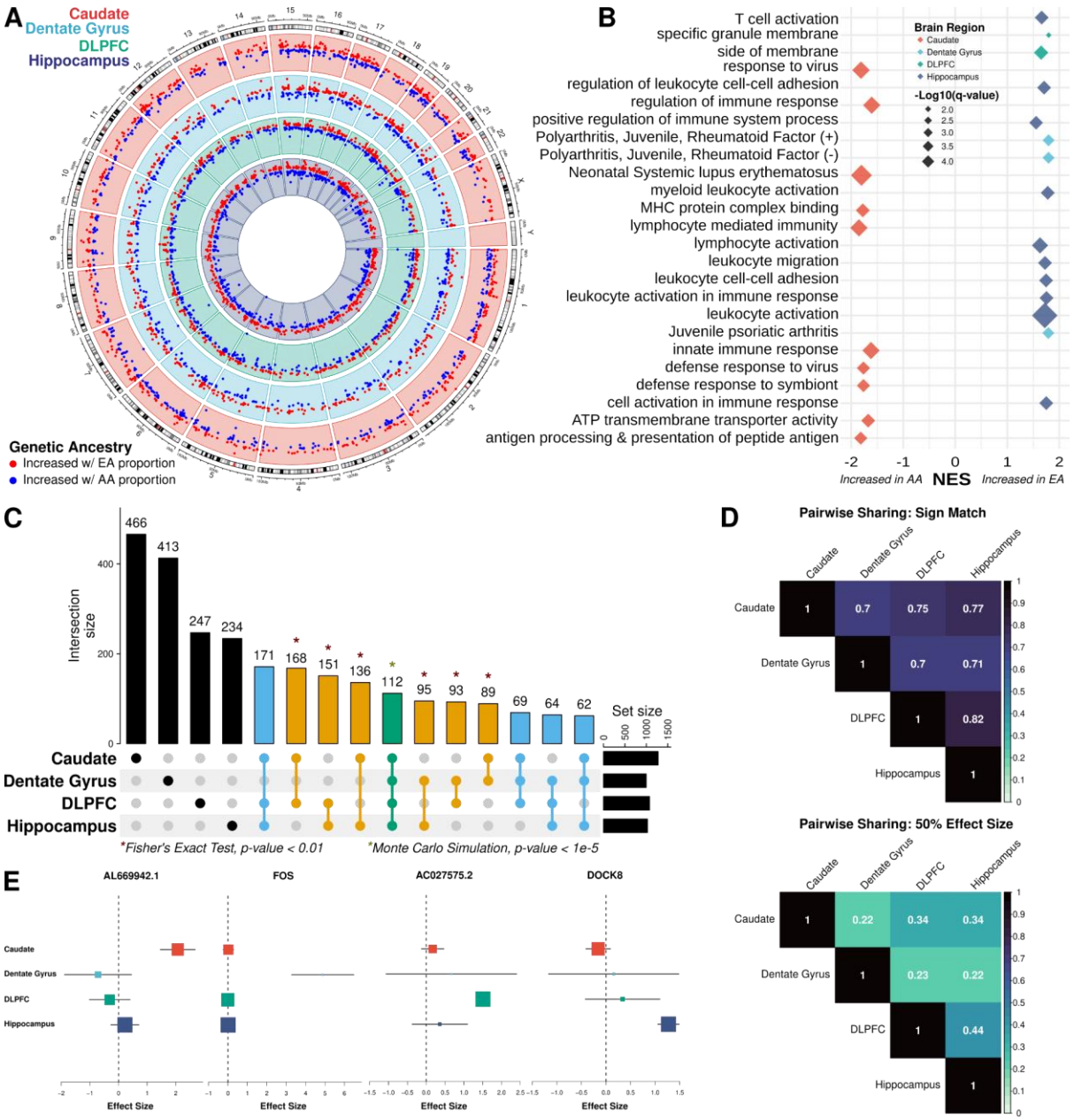

Fig. 2: Extensive ancestry-associated expression changes across the brain region. A. Circos plot showing ancestry DEGs across the caudate (red), dentate gyrus (blue), DLPFC (green), and hippocampus (purple). **B.** Gene set enrichment analysis (GSEA) of differential expression analysis across brain regions, highlighting terms associated with increased AA (African ancestry) or EA (European ancestry) proportions. **C.** UpSet plot showing large overlap between brain regions. Green is shared across the four brain regions; blue, shared across three brain regions; orange, shared between two brain regions; and black, unique to a specific brain region. * Indicating significant pairwise enrichment (Fisher's exact test) or significant overlap between all four brain regions (Monte Carlo simulation). **D.** Heatmaps of the proportion of ancestry DEG sharing with concordant direction (sign match; top) and within a factor 0.5 effect size (bottom) **E.** Metaplot showing examples of brain region-specific ancestry effects.

HLA region and immune cell composition play a limited role on ancestry-associated expression differences across the brain

Given the primary enrichment signal for immune-related pathways and cell types, we next investigated if immune variation was driving the observed transcriptional changes. Initially, we examined enrichment of ancestry-associated DEGs for the major histocompatibility complex (MHC) region. Here, we found global ancestry-associated DEGs of the caudate, DLPFC, and hippocampus enriched for HLA class II, while dentate gyrus enriched for Zinc finger proteins associated with the extended class I MHC region (Fisher's exact test, FDR < 0.05; **Fig. S15**). While we found limited enrichment of local ancestry-associated DEGs for gene clusters of the MHC region across brain regions, we still observed significant enrichment of HLA class II genes for the caudate similar to global ancestry DEGs (Fisher's exact test, FDR < 0.05; **Fig. S16**). Altogether, these results suggest that ancestry-associated DEGs (global and local) within the MHC region are primarily enriched for HLA class II genes.

Next, we re-examined functional enrichment of ancestry-associated DEGs after removing the MHC region (i.e., HLA-specific genes, MHC region, and extended MHC region) to determine if the limited MHC enrichment drove functional enrichment of immune-related pathways of our ancestry-associated DEGs. After excluding the HLA genes, we still observed strong enrichment for immune-related pathways (**Fig. S17**). Furthermore, we observed similar immune-related enrichment (i.e., response to virus, interleukin-12 production, macrophage activation, leukocyte migration, and innate immune response) after excluding the MHC region (**Fig. S18**) or the extended MHC region (**Fig. S19**) across brain regions. This was also the case with local ancestry DEGs (**Fig. S20**), suggesting that the extended MHC region does not drive ancestry-associated DEG enrichment of immune-related pathways.

Although the MHC region did not appear to drive our immune response enrichment, immune variation either from HLA gene diversity or glial cell composition could still contribute to the transcriptional changes observed in our ancestry-associated DEGs. As such, we next assessed to what degree HLA variation or glial cell composition contributed to the expression changes. To assess glial cell composition, we added glial cells composition (astrocytes, microglia, macrophage, oligodendrocytes, oligodendrocyte progenitor cells, and T cells) as covariates in our DE model. When we compared effect sizes with the original model, we found a high degree of correlation (Spearman; rho from 0.81 to 0.92; **Fig. S21A**), suggesting glial cell composition had a minimal effect. For HLA variation, we added the first five PCs of imputed HLA alleles (accounting for 66% of variance explained) as covariates and compared effect sizes with our original model. Similar to glial cell composition, we found HLA genetic variation only minimally changed effect sizes (Spearman; rho from 0.83 to 0.87; **Fig. S21B**). Altogether, these sensitivity analyses suggest that immune variation contributes only minimally to transcriptional changes for ancestry-associated DEGs.

Ancestry-associated DEGs are evolutionarily less constrained

With consistent significant enrichment of DEGs and co-expression modules for immune response, we hypothesized that this functional connection for the DEGs with a cellular biology that is uniquely adaptable would render them more likely to be tolerant of phenotypic consequences of gene disruption and would therefore be evolutionarily less constrained. To test this hypothesis, we examined the gene and transcript constraint scores (*29*) for the global ancestry-associated DEGs. Unsurprisingly, we found significant depletion of DEGs for highly constrained genes (Fisher's exact

test, FDR < 0.0001; **Fig. 3A**). On the transcript level, we found a similar trend (**Fig. 3B**) with differentially expressed (DE) transcripts associated with less constrained genes. Additionally, we observed a significant negative correlation with DEGs signal (lfsr) and gene and transcript constraint scores (Pearson, p-value < 0.0001; **Fig. 3C**). Unsurprisingly, these results suggest that ancestry-associated DE features are associated with the more rapidly evolving genes as previously seen in immunity related genes (*30*, *31*).

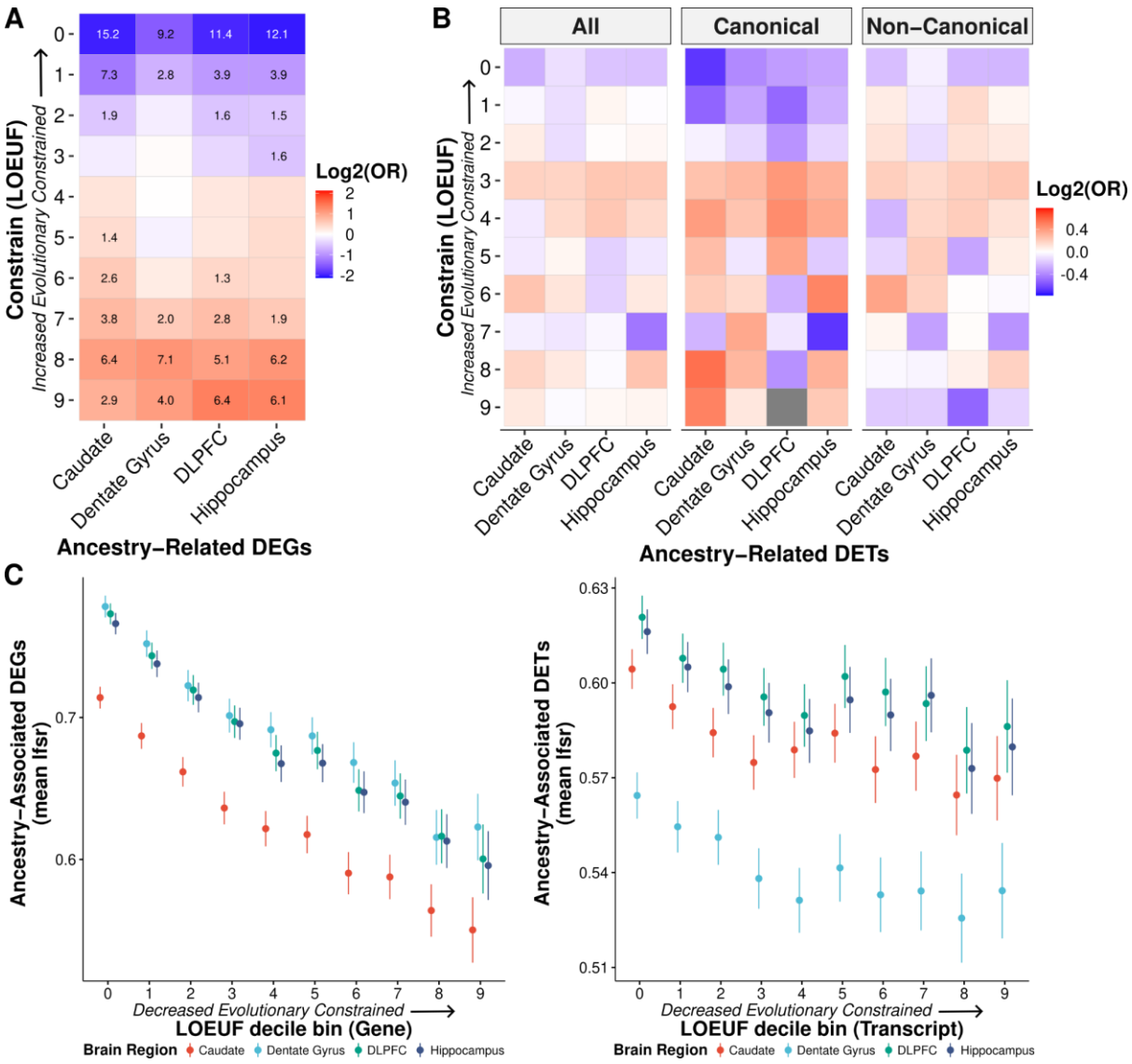

Fig. 3: **Ancestry-associated genes and canonical transcripts are evolutionarily less constrained.** **A.** Significant depletion of ancestry DEGs for evolutionarily constrained genes (canonical transcripts) across brain regions. Significant depletion/enrichments (two-sided, Fisher's exact test, FDR corrected p-values, -log10 transformed) are annotated within tiles. Odds ratios (OR) are $\log_2$ transformed to highlight depletion (blue) and enrichment (red). **B.** Similar trend of depletion of ancestry DE transcripts (DETs; all, canonical, and non-canonical) for evolutionarily constrained transcripts across brain regions. Odds ratios are $\log_2$ transformed to highlight depletion (blue) and enrichment (red). **C.** The mean of ancestry-associated DE feature (i.e., gene and transcript) lfsr as a function of LOEUF (loss-of-function observed/expected upper bound fraction) decile shows a significant negative correlation for genes (left; for the caudate, dentate gyrus, DLPFC, and hippocampus: two-sided, Pearson, r = -0.20, -0.20, -0.21, and -0.21; p-value = 3.0×10^{-122}, 7.6×10^{-113}, 8.6×10^{-126}, and 1.2×10^{-122}) and transcripts (right; for the caudate, dentate gyrus, DLPFC, and hippocampus: two-sided, Pearson, r = -0.05, -0.05, -0.04, and -0.04; p-value = 8.6×10^{-13}, 1.7×10^{-11}, 9.0×10^{-11}, and 3.2×10^{-10}). Error bars correspond to 95% confidence intervals.

The role of genetic variants on ancestry-associated expression differences in the brain

To assess the contribution of genetic variation to genetic ancestry-associated DEGs, we first mapped main effect cis-eQTL in Black American individuals (n=120, 45, 121, and 131 for the caudate, dentate gyrus, DLPFC, and hippocampus, respectively) examining genetic variants within +/- 500 kb of each feature (gene, transcript, exon, and junction). To improve detection of eQTL, we applied mash and identified at least one cis-eQTL for 13,857 genes ("eGenes") across brain regions (lfsr < 0.05; n=10,867 for the caudate; n=11,664 for the dentate gyrus; n=11,173 for the DLPFC; and n=10,408 for the hippocampus; **Table S4** and **Data S6**). Of these 13,857 eGenes, the majority (64.1%; **Fig. 4A**) were shared across all brain regions with only about 0.25 to 14.5% showing brain region specificity. When we examined the direction of effect, however, this number dramatically increased with more than 96% sign matching (**Fig. 4B**).

We also examined eQTL whose effects may vary as a function of genetic ancestry. Our examination followed a similar model to the main effect analysis but with an interaction term between SNP and ancestry proportion. We identified at least one ancestry-dependent cis-eQTL (within +/- 500 kb of each feature) for 943 unique genes across brain regions (lfsr < 0.05, n=531, 942, 573, and 531 for the caudate, dentate gyrus, DLPFC, and hippocampus, respectively; **Fig. S22**, **Table S5**, and **Data S7**) with 54.1% (510 eGenes) shared across the four brain regions (**Fig. S23**). This relatively limited detection of ancestry-dependent eQTL supports other work showing high correlation of causal effects across local ancestry of admixed individuals (*32*).

We next tested whether these eGenes (main effect and ancestry-dependent) were likely to be differentially expressed by genetic ancestry. Across brain regions, we found significant enrichment (Fisher's exact test, FDR < 0.05) of these eGenes (lfsr < 0.05) with ancestry-associated DEGs (lfsr < 0.05; **Fig. 4C** and **Fig. S23C**). Given the potential correlation of genotypes with eGenes and ancestry inference, we also examined allele frequency differences between DEGs and non-DEGs. We found a significant increase in allele frequency differences for DEGs compared with non-DEGs (Mann-Whitney U, p-value < 0.05; **Fig. 4D** and **Fig. S24**) across brain regions. These results suggest that a genetic component is likely influencing these expression differences, potentially due to divergence in allele frequencies.

To test this possibility, we imputed gene expression levels from genotypes using an elastic net model, and then examined the correlation between the observed genetic ancestry effect from our ancestry DE analysis and the predicted genetic ancestry effect computed from the predicted expression levels across samples. Unsurprisingly, eGenes showed higher prediction accuracy than non-eGenes; interestingly, however, eGenes with an ancestry difference in gene expression had a stronger genetic component (higher R^2) than eGenes without an ancestry difference across the four brain regions (**Fig. S25**). Furthermore, the imputed gene expression levels explained an average of 59.5%, 58.7%, 56.8%, and 56.8% of the variance in genetic ancestry effect sizes across the caudate, dentate gyrus, DLPFC, and hippocampus, respectively (**Fig. 4E**). This variance explained generally increased on the isoform level (transcript [R^2 = 50.8%±7.0%], exon [R^2 = 61.6%±4.1%], and junction [R^2 = 62.6%±5.1%]; **Fig. S26**) across brain regions. In contrast, the genetic variant for the top main effect eQTL associated with these genes explained on average ~20% of the variance in genetic ancestry effect sizes with a similar proportion for the isoform level (**Fig. S27**). Thus, genetic variants contributed to nearly 60% of the observed genetic ancestry in gene expression – and variant effects on alternative splicing were even greater.

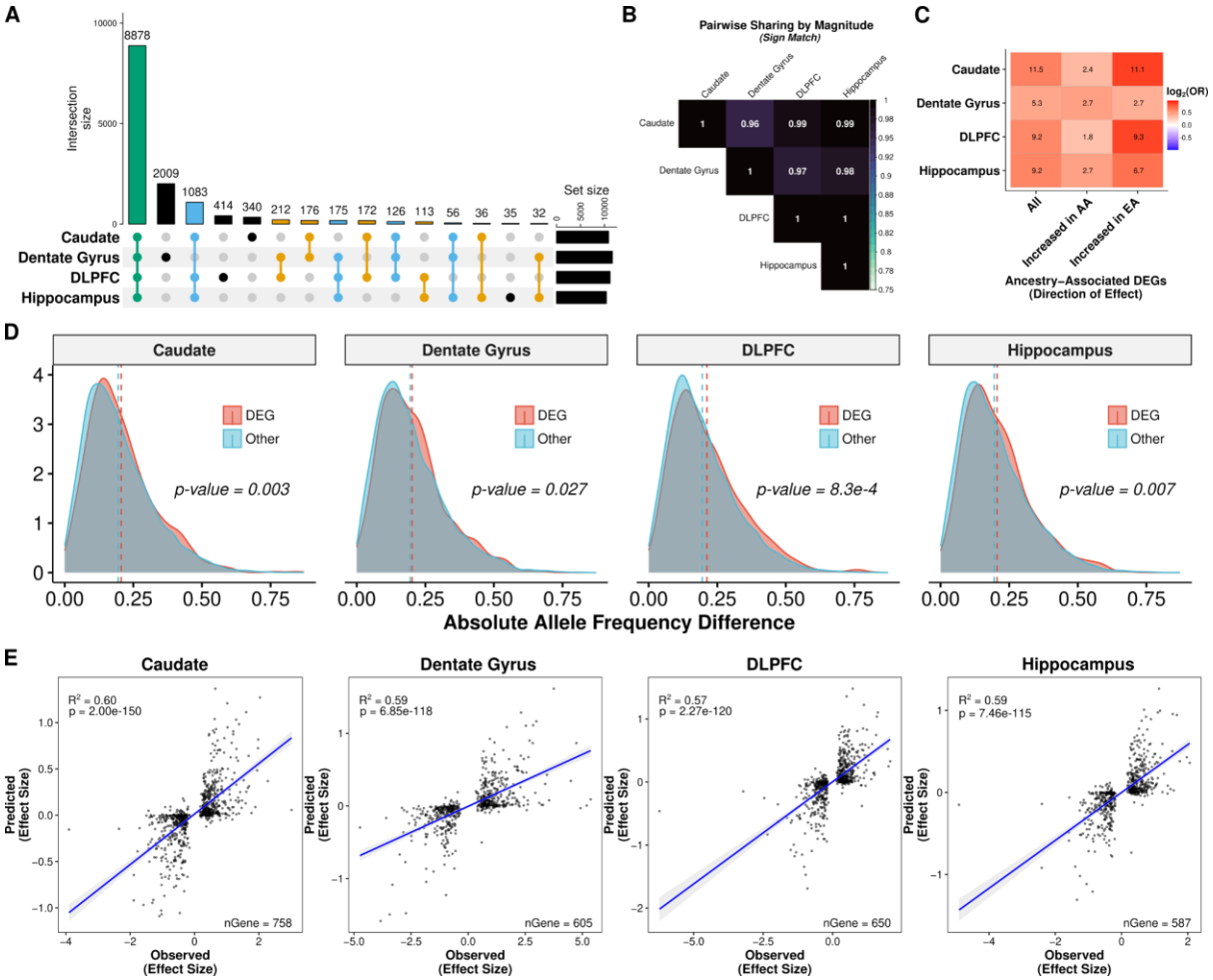

Fig. 4: Genetic contribution of genetic ancestry differences in expression across the brain. A. UpSet plot showing large overlap between brain regions of eGenes. **B.** Heatmap of the proportion of ancestry DEG sharing with concordant direction (sign match). **C.** Significant enrichment of ancestry-associated DE genes for eGenes (unique gene associated with an eQTL) across brain regions separated by direction of effect (increased in AA or EA proportion). **D.** Density plot showing significant increase in absolute allele frequency differences (AFD; one-sided, Mann-Whitney U, p-value < 0.05) for global ancestry-associated DEGs (red) compared with non-DEGs (blue) across brain regions. A dashed line marks the mean absolute AFD. Absolute AFD calculated as the average absolute AFD across a gene using significant eQTL (lfsr < 0.05). **E.** Correlation (two-sided, Spearman) of elastic net predicted (y-axis) versus observed (x-axis) ancestry-associated differences in expression among ancestry-associated DEGs with an eQTL across brain regions. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

Differential gene expression in a binary contrast of Black and White Americans

To extend our analysis of DEGs driven by genetic ancestry, we performed a binary analysis of combined Black and White American samples (**Table S6**) – the latter showing very little admixture of African ancestry (STRUCTURE; African ancestry mean = 0.03, range = 0-0.16; **Fig. S1**). Using these American samples, we selected individuals with relatively limited admixture (Black Americans ≥ 0.8 African genetic ancestry and White Americans > 0.99 European genetic ancestry) across the caudate, dentate gyrus, DLPFC, and hippocampus. To limit the influence of the larger sample size for this

binary analysis (Black American vs White American), we randomly sampled ten times without replacement to approximate the admixed Black American-only analysis sample size (caudate, n=122 [61 each]; dentate gyrus, n=46 [23 each]; DLPFC, n=124 [62 each]; and hippocampus, n=134 [67 each]). We identified more than double as many ancestry-associated DEGs (5,324 unique genes, median lfsr < 0.05; **Fig. S28A**, **Table S7**, and **Data S8**) representing 28% of all genes tested across the caudate (n=2,877), dentate gyrus (n=2,219), DLPFC (n=3,318), and hippocampus (n=2,818) with similar immune system enrichment patterns (**Fig. S28B** and **Data S9**).

We next compared the binary analysis DE results (genes, transcripts, exons, and junctions) with the admixed Black American-only results. While we found a significant overlap of ancestry associated DE features (Fisher's exact test, p-value < 0.0001), approximately 72% of features (3847 unique genes) were unique to the binary DE results (**Fig. S29**). Even so, effect sizes from binary analysis were significantly correlated (Spearman, rho = 0.43 to 0.49, p-value < 0.0001; **Fig. S30**) with effect sizes from admixed Black American-only analysis across features and brain regions, which increased when we examined only shared features (Spearman, rho = 0.60 to 0.66, p-value < 0.0001; **Fig. S31**). While these results confirm most of the ancestry-associated DEGs in the Black American sample alone, they also highlight additional ancestry-related factors that influence gene expression presumably including environmental events (i.e., epigenetic).

Environmental contributions to global ancestry-associated differential expression

Our binary DE analysis of Black and White Americans suggests that environmental factors may also contribute to global ancestry-associated DEGs. To identify DEGs driven by environmental factors, we used DNAm as an environmental proxy in Black Americans. We began by identifying the top 1% of variable CpGs that are likely driven by unknown environmental factors. We identified these CpGs by removing variation attributable to batch and to unknown technical and biological factors as captured by the top five DNAm principal components, while preserving variation due to global ancestry. We then grouped those top variable CpGs into variable methylated regions (VMRs) for the caudate (89 samples; 12,051 VMRs), DLPFC (69 samples; 9,701 VMRs), and hippocampus (69 samples; 9,924 VMRs). In contrast to our DE analysis results, we identified fewer VMRs that were differentially methylated regions (DMRs) for global ancestry (FDR < 0.05; n=3, 1, and 8 for the caudate, DLPFC, and hippocampus, respectively). However, we identified a larger number of local ancestry-associated DMRs (FDR < 0.05; n=494, 260, and 265 for the caudate, DLPFC, and hippocampus, respectively; **Fig. 5A**).

We reasoned that the difference in DMRs linked to global and local ancestry can be explained both biologically and statistically. Biologically, DNAm tends to be more influenced by local genetic variations. Statistically, local ancestry is more variable than global ancestry, which results in a higher power to detect DNAm differences and a smaller standard deviation in estimated effect size. This is demonstrated in **Fig. S32** and **Data S12**, where we compared DNAm levels against local ancestry and global ancestry levels for VMRs associated with local ancestry. Even so, we find significant correlation between local and global ancestry-associated DMRs (**Fig. S33**). Functional enrichment analysis of local ancestry-associated DMRs suggested that these DMRs were enriched for gene sets related to immune functions across all three brain regions (**Fig. 5B**), consistent with the functional enrichment results of ancestry-associated DEGs.

We next regressed out known biological factors (local ancestry, age, sex), as well as the potential batch effects and other unknown biological factors captured by the top five principal components of DNAm levels for each VMR. We used P_{ST} estimates (*18*) to provide a measure of proportion of

overall gene expression variance explained by between-population differences. P_{ST} values range from 0 to 1, where values close to 1 imply the majority of expression variance is due to differences between populations. We defined deltaP$_{ST}$ (ΔP_{ST}) as the difference between P_{ST} values before and after regressing out the effect of VMRs associated with each gene. Therefore, ΔP_{ST} quantifies the proportion of ancestry-associated DEGs that are likely due to environmental exposures. Using this method, we found that across brain regions the average ΔP_{ST} was 15% (12.2%, 14.4%, and 18.3% for the caudate, DLPFC, and hippocampus, respectively; **Fig. 5C**). Altogether, these results imply that unknown environmental exposures measured by DNAm provide a minor contribution to the observed, primarily immune-related expression differences in our Black American neurotypical sample.

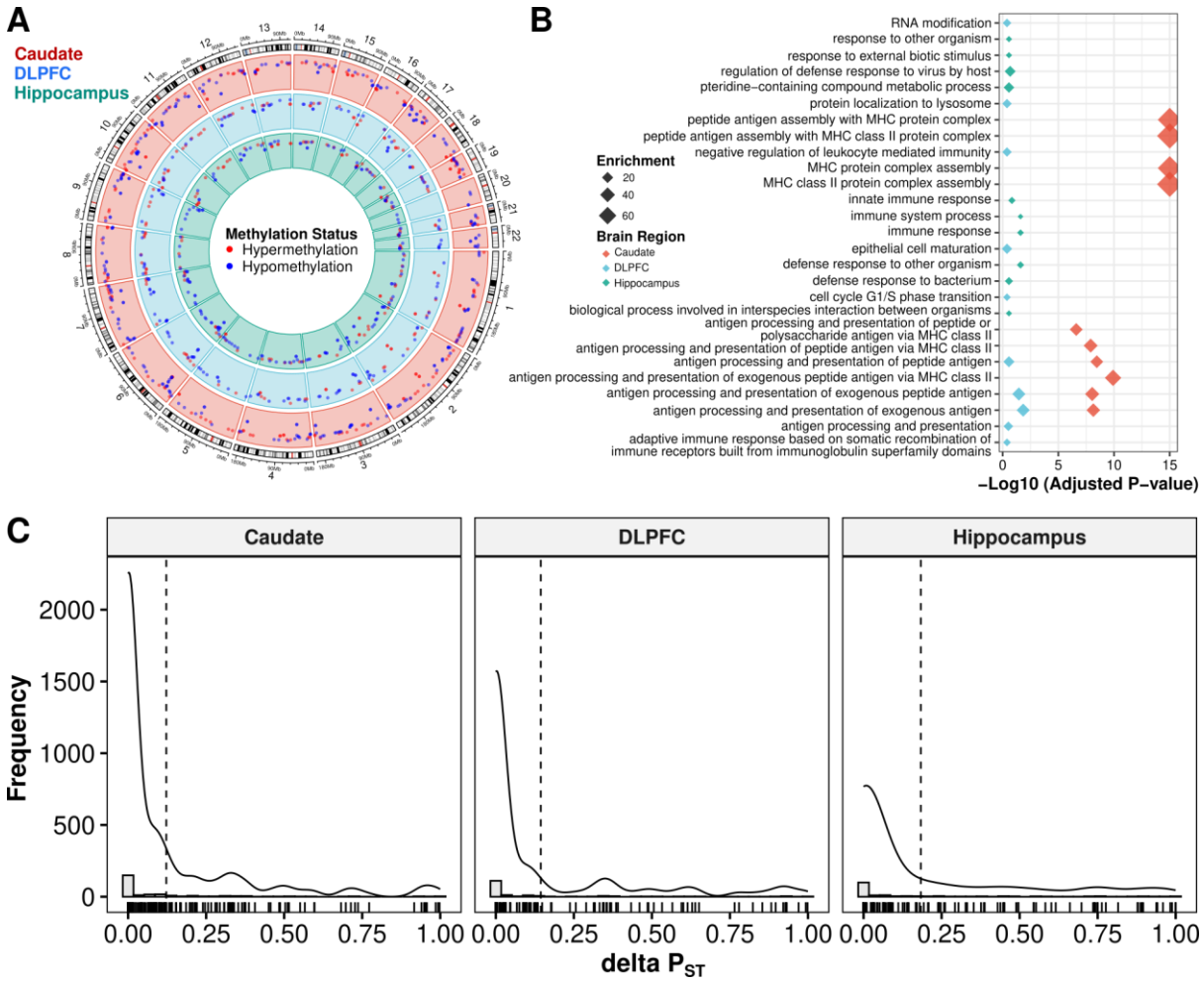

Fig. 5: Unknown environmental factors are primary drivers of nearby global ancestry-associated DEGs. A. Circos plot showing local ancestry-associated DMRs across the caudate (red), DLPFC (blue), and hippocampus (green). Methylation status is annotated in red for hypermethylation and blue for hypomethylation. **B.** Gene term enrichment of DMRs across brain regions. **C.** Histograms showing distribution of ΔP_{ST} associated with the impact of unknown environmental factors as captured by residualized VMR (corrected by local ancestry, age, sex, and unknown biological factors captured by PCA) for nearby global ancestry-associated DEGs. A dashed line marks the mean ΔP_{ST}. A solid line shows the density overlay.

Association of ancestry-associated expression differences with immune- and brain-related traits

We reasoned that ancestry-associated DEGs may contain risk genes that explain susceptibility of brain-related illnesses based on ancestry. To explore this hypothesis, we conducted stratified LD score regression (S-LDSC) to assess the polygenic contributions of global ancestry-associated DEGs to 17 brain-related traits (e.g., ADHD, autism, BMI, depression, and schizophrenia) (*33*). As our ancestry-associated DEGs were enriched for gene sets related to immune functions, we included five immune-related traits as a positive control in our S-DLSC analysis. Overall, we observed that ancestry-associated DEGs were enriched for heritability of neurological disorders and immune-related traits but not for psychiatric disorders and behavioral traits (**Fig. 6**, **Fig. S34**, and **Data S10**). This also included limited enrichment of peripheral immune function (*34–36*) (Fisher's exact test, FDR < 0.05; **Fig. S35**), which is consistent with our previous enrichment showing a greater association with brain immune cell types compared to non-brain immune cell types (**Fig. S12**).

Specifically, we found enrichment for heritability of ischemic stroke (enrichment fold = 1.5, FDR = 0.009) for ancestry-associated DEGs in the DLPFC, accounting for 26% of total heritability (**Fig. S34**). This enrichment was mainly driven by DEGs associated with an increase in AA proportion (enrichment fold = 1.7, FDR = 0.013), but not to EA (enrichment fold = 1.2, p-value = 0.2). Furthermore, stratified analysis by protein-coding and non-coding DEGs showed that enrichment was primarily driven by protein-coding DEGs, but not non-coding DEGs (**Fig. 6**). We observed stronger enrichment of ischemic stroke for protein-coding DEGs in the DLPFC (increased AA proportion; enrichment fold = 2.1, FDR = 0.011). This finding is consistent with epidemiological data that Black Americans are up to 50% more likely to experience ischemic stroke, and Black men are up to 70% more likely to die from stroke compared to non-Hispanic White men (*37*, *38*). Moreover, our cell-type enrichment analysis showed that the DEGs associated with increased AA proportion were enriched for vascular smooth muscle cells, endothelial cells, and pericytes (**Fig. S10**), all of which may contribute to vascular pathology implicated in stroke.

In addition to ischemic stroke, we also found enrichment for heritability of Parkinson's disease (enrichment fold = 1.6, FDR = 0.025) for ancestry-associated DEGs in the DLPFC, accounting for 27% of disease heritability (**Fig. S34**). Interestingly, this enrichment was primarily driven by DEGs that were increased with EA proportion (enrichment fold = 1.9, FDR = 0.032), but not to AA proportion (enrichment fold = 1.3, p-value = 0.23). Again, this enrichment for Parkinson's disease in the DLPFC was driven by protein-coding DEGs (increased EA proportion; enrichment fold = 2.3, FDR = 0.038; **Fig. 6**). This finding echos epidemiological studies suggesting that the prevalence of Parkinson's disease is greater in White Americans compared with Black Americans (*39*). Additionally, cell-type enrichment analysis for DEGs associated with increased EA proportion showed enrichment for cell-type-specific genes related to the microglia, astrocytes, and oligodendrocyte progenitor cells (**Fig. S10**). Interestingly, we also found ancestry-associated glial cell subtypes (i.e., astrocyte [AST7] and oligodendrocyte lineage [OPC1]) significantly enriched for Parkinson's disease heritability (enrichment fold > 2.0, FDR < 0.01; **Fig. S36**), suggesting a potential role for specific glial subtypes in the pathogenesis of Parkinson's disease.

We also observed enrichment for heritability of Alzheimer's disease for ancestry-associated DEGs across DLPFC, hippocampus and caudate accounting for 26%, 23% and 30% of total heritability, respectively (**Fig. S34**). These enrichments were mainly driven by protein-coding DEGs associated with an increase in AA proportion for the DLPFC (enrichment fold = 2.0, FDR = 0.013; **Fig. 6**) and

hippocampus (enrichment fold = 1.9, FDR = 0.02; **Fig. 6**). Surprisingly, we found the opposite effect with an increase in EA proportion for the caudate when considering all DEGs (**Fig. S34**), which disappeared when considering only protein coding or non-protein coding DEGs (**Fig. 6**). Cell-type enrichment analysis of astrocytes, however, shows ancestry-specific effects consistent with this finding for the caudate (increased EA proportion; **Fig. S10**). Moreover, we found ancestry-associated glial cell subtypes (i.e., microglia [MG0] and astrocyte [AST1 and AST7]) significantly enriched for Alzheimer's disease heritability (enrichment fold > 2.2, FDR < 0.01; **Fig. S36**) and ancestry-associated DEGs enrichment for multiple activated microglia states (*40*) (**Fig. S37A**). Interestingly, these microglia states were associated with mouse Alzhiemer's disease-associated microglial genes and Alzhiemer's disease GWAS signals (**Fig. S37B**). We also observed significant enrichment of ancestry-associated DEGs primarily with Alzheimer's disease-related DEGs between early Alzheimer's and late Alzheimer's (late response; **Fig. S38**).

In contrast, we observed significant depletion in heritability for several brain-related traits (e.g., education years, smoking initiation, age of smoking, schizophrenia, and depression; enrichment fold < 1, FDR < 0.05; **Fig. 6**, **Fig. S34**, and **Data S10**) of our ancestry-associated DEGs across brain regions. These results are consistent with our observations that ancestry-associated DEGs are depleted for gene sets related to the neuronal functions that are believed to play major roles in psychiatric disorders and behavior traits.

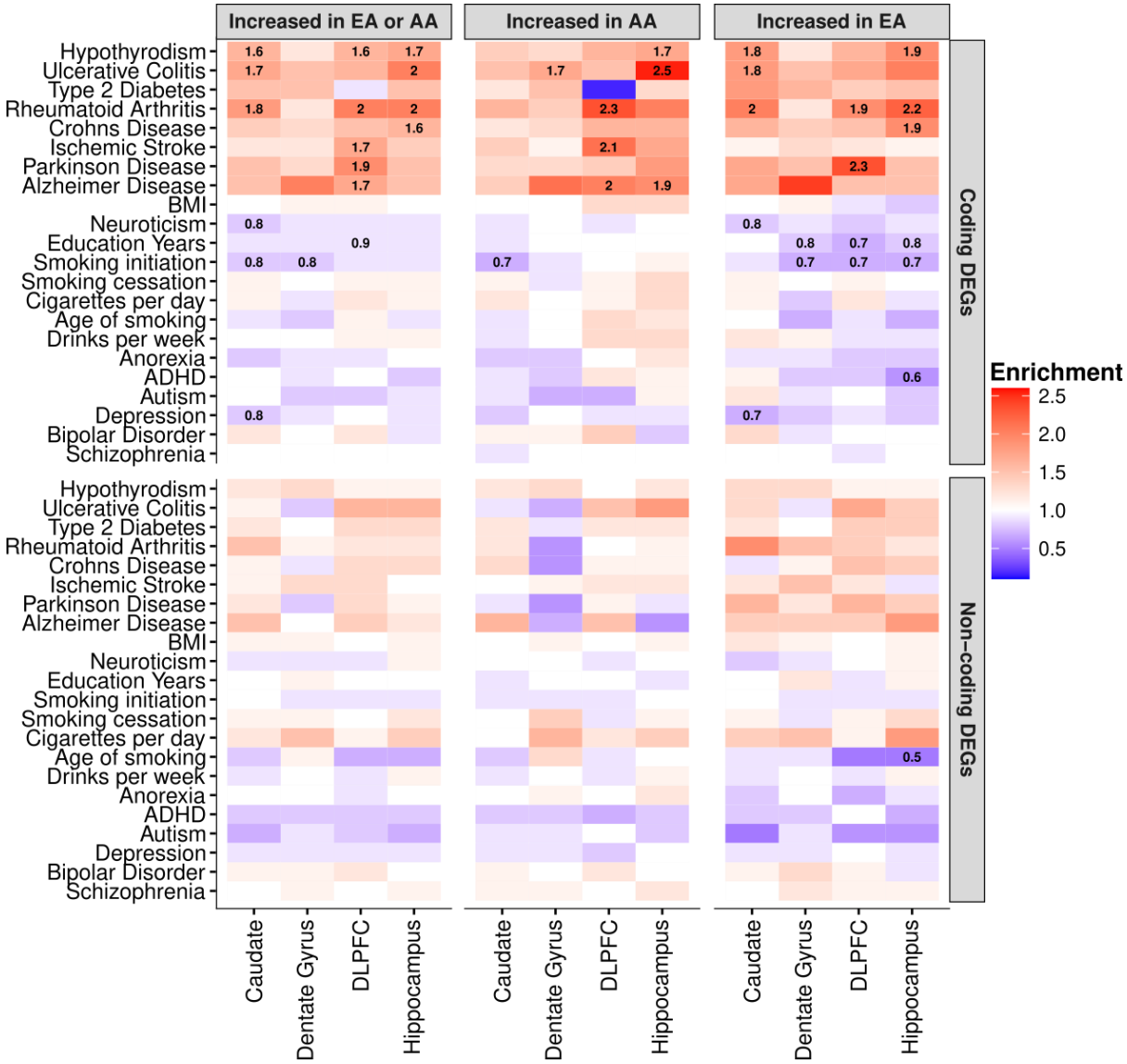

Fig. 6: Global ancestry-associated DEGs stratified by coding or non-coding DEGs show general enrichment for heritability of several neurological and immune-related traits, but depleted for brain-related behavioral traits. Heatmap for ancestry-associated DEGs that show enrichment (red) or depletion (blue) for heritability of brain- and immune-related traits from S-LDSC analysis. Significant enrichment for heritability traits disappears when limited to non-coding DEGs. Numbers within tiles are levels of enrichment (> 1) or depletion (< 1) that are significant after multiple testing correction (FDR < 0.05). The left panel shows results for all DEG in each brain region. The middle and right panels show results for DEG increased with AA or EA proportions for each brain region, respectively.

Discussion

Here we provide the first detailed characterization of the impact of genetic ancestry on expression and DNA methylation in the human brain. Using admixed Black American donors, we have identified thousands of genomic features (i.e., genes, transcripts, exons, and junctions) associated with genetic ancestry and demonstrated that these features are evolutionarily less constrained. Approximately 60% of these ancestry-associated DEGs are associated with genetic variations. Our data show consistent enrichment for immune response pathways for genetic ancestry-associated DEGs and consistent absence of ancestry associations with neuronal functions. Furthermore, we found similar trends when we examined local genetic ancestry. Even so, given expression heritability is dominated (i.e., about 70%) by many small trans effects (*41*, *42*), we have chosen to focus primarily on global genetic ancestry.

Interestingly, our findings show the direction of enrichment varies by brain region for immune-related pathways, increasing in relation to AA proportion in caudate and increasing in relation to EA proportion in the other regions. Because the specific genes in these immune function sets vary somewhat across regions, it is tempting to speculate on 1) how genetics and the environment sculpt variation in this regional biology and 2) whether the functional and behavioral impact of these ancestry-associated DEGs depends on the biology of particular brain regions. However, there is no simple "up or down" bias to the functional associations independent of brain region. For example, if AA proportion is a risk factor to immune response in the caudate, then by the same reasoning AA proportion would be a protector factor for immune response in the hippocampus and prefrontal cortex. We considered that differences in directionality across regions may reflect variation in cell composition as the caudate was the only brain region without a laminar architecture. However, laminar architecture in the brain has generally implicated neuronal biology (*43*), which was not the case here (i.e., enrichment of immune-related pathways). Notably, virtually all of our findings are more significant at the isoform level, implicating gene splicing and processing as a more incisive method for explaining the effect of ancestry on gene expression.

Among the more striking findings of our data is the enrichment of heritability for neurological brain illness among ancestry-associated DEGs. Small vessel stroke and ischemic stroke are up to 50% more frequent in Black Americans (*37*, *38*), and here we show that heritability for stroke was enriched among DEGs that were increased in proportion to AA in our admixed Black population. In contrast, heritability for Parkinson's disease, which is more prevalent in White Americans (*39*), was enriched among DEGs in proportion to EA. Interestingly, we observed a nearly two-fold enrichment for heritability of Alzheimer's disease among DEGs that were increased with AA proportion in DLPFC and hippocampus, regions cardinally involved in Alzheimer's disease. This observation echoes the fact that Alzheimer's disease is twice as prevalent in Black Americans (*44*, *45*). However, general enrichment of DEGs for Alzheimer's disease in the caudate associated with an increase in EA proportion highlights the potential regional complexity of the disorder in the brain as caudate is not generally considered a site of Alzheimer's disease pathology. Ancestral DEGs increase heritability for several immune disorders and traits but not specifically in relation to either ancestry across the brain. It is noteworthy that the DEGs are not linked with heritability of psychiatric disorders and related behavioral traits, perhaps consistent with genes associated with these traits being especially enriched in neurons, which were again, conspicuously lacking in DEGs based on ancestry.

In addition to our analysis of the admixed Black American population, we also performed a combined analysis with White Americans. As an internal validation, we found significant overlap between this

and our Black American-only analyses (i.e., DE), but a dramatic increase in the extent of differentially expressed features. Additionally, this combined analysis (Black and White Americans) revealed similar enrichment of the immune response, again in analogous alternating directionality depending on brain region. While these results implicate environmental exposures that might reflect systematic differences between the two ancestral groups, disambiguating genetic from environmental factors in this context is challenging. We, therefore, chose to examine the environmental impact on our Black American-only global ancestry-associated DEGs. To this end, we identified thousands of VMRs across the brain in this context.

To highlight those VMRs likely enriched for environmental influence, we focused on the top 1% of VMRs and looked for ancestry-associated DMRs within these genomic regions. Consistent with DE analysis, we found that local ancestry DMRs were enriched for genomic regions related to immune functions. When we used VMRs as an environmental proxy to examine the effect of environmental exposures on the DEGs, we found they explained, on average, roughly 15% of population differences in gene expression. Although we used local ancestry to correct for genetic background, we cannot be sure that the variation captured via methylation is solely attributed to environmental factors or that methylation can capture all environmental factors. A limitation of this study is the lack of social determinants of health information, which could have directly measured specific environmental exposures instead of using DNAm as a proxy. Nevertheless, our analyses demonstrate the potential to limit the impact of potentially systematic environmental factors by leveraging admixture populations for genetic ancestry analyses.

This enrichment in immune-related pathways is not altogether unexpected: a previous study showed population differences in macrophages associated with the innate immune response to infection (*18*). Furthermore, it is well documented that genetic variation is an important contributor to immune variation (*46–48*) and immune cell function (*34–36*). This research is particularly important for neuropsychiatric disorders (including schizophrenia, autism spectrum disorder, and Alzheimer's disease) where the immune system has been implicated (*49–51*). Many of these neuropsychiatric disorders also show a racial health disparity (*44*, *52–54*). As a result, we examined our enrichment of immune function in more detail. Interestingly, we found little evidence that the MHC region, HLA variation, or glial cell composition drove our identified immune-response pathway enrichment. Additionally, we found stronger enrichment of brain immune compared with peripheral immune cell types, suggesting the potential involvement of a brain-specific immune response of these DEGs. Altogether, our results provide a starting point for further investigation for potential therapeutic interventions involving the immune response – therapeutic interventions that could address these health disparities.

In summary, we provide a detailed examination of the genetic and environmental contributions to genetic ancestry transcriptional changes in the brain. We leveraged genetic diversity within admixture populations to limit environmental confounders, resulting in converging evidence of the immune response in genetic ancestry-associated transcriptional changes in the brain. The research we have provided here substantively furthers our understanding of the contribution of genetic ancestry in the brain, opening new avenues to the development of ancestry-aware therapeutics and paving the way for equitable, personalized medicine.

References

1. Z. D. Bailey, N. Krieger, M. Agénor, J. Graves, N. Linos, M. T. Bassett, Structural racism and health inequities in the USA: evidence and interventions. *Lancet*. **389**, 1453–1463 (2017).

2. D. Gurdasani, I. Barroso, E. Zeggini, M. S. Sandhu, Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* **20**, 520–535 (2019).

3. G. Sirugo, S. M. Williams, S. A. Tishkoff, The missing diversity in human genetic studies. *Cell*. **177**, 26–31 (2019).

4. D. R. Weinberger, K. Dzirasa, L. L. Crumpton-Young, Missing in action: african ancestry brain research. *Neuron*. **107**, 407–411 (2020).

5. A. R. Bentley, S. L. Callier, C. N. Rotimi, Evaluating the promise of inclusion of African ancestry populations in genomics. *NPJ Genom. Med.* **5**, 5 (2020).

6. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

7. D. Taliun, D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech, R. Torres, S. A. G. Taliun, A. Corvelo, S. M. Gogarten, H. M. Kang, A. N. Pitsillides, J. LeFaive, S.-B. Lee, X. Tian, B. L. Browning, S. Das, A.-K. Emde, W. E. Clarke, D. P. Loesch, A. C. Shetty, G. R. Abecasis, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. **590**, 290–299 (2021).

8. H3Africa Consortium, C. Rotimi, A. Abayomi, A. Abimiku, V. M. Adabayeri, C. Adebamowo, E. Adebiyi, A. D. Ademola, A. Adeyemo, D. Adu, D. Affolabi, G. Agongo, S. Ajayi, S. Akarolo-Anthony, R. Akinyemi, A. Akpalu, M. Alberts, O. Alonso Betancourt, A. M. Alzohairy, G. Ameni, et al., Enabling the genomic revolution in Africa. *Science*. **344**, 1346–1348 (2014).

9. B. Zeng, J. Bendl, R. Kosoy, J. F. Fullard, G. E. Hoffman, P. Roussos, Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat. Genet.* **54**, 161–169 (2022).

10. L. Collado-Torres, E. E. Burke, A. Peterson, J. Shin, R. E. Straub, A. Rajpurohit, S. A. Semick, W. S. Ulrich, BrainSeq Consortium, A. J. Price, C. Valencia, R. Tao, A. Deep-Soboslay, T. M. Hyde, J. E. Kleinman, D. R. Weinberger, A. E. Jaffe, Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron.* **103**, 203-216.e8 (2019).

11. A. E. Jaffe, D. J. Hoeppner, T. Saito, L. Blanpain, J. Ukaigwe, E. E. Burke, L. Collado-Torres, R. Tao, K. Tajinda, K. R. Maynard, M. N. Tran, K. Martinowich, A. Deep-Soboslay, J. H. Shin, J. E. Kleinman, D. R. Weinberger, M. Matsumoto, T. M. Hyde, Profiling gene expression in the human dentate gyrus granule cell layer reveals insights into schizophrenia and its genetic risk. *Nat. Neurosci.* **23**, 510–519 (2020).

12. K. J. M. Benjamin, Q. Chen, A. E. Jaffe, J. M. Stolz, L. Collado-Torres, L. A. Huuki-Myers, E. E. Burke, R. Arora, A. S. Feltrin, A. R. Barbosa, E. Radulescu, G. Pergola, J. H. Shin, W. S. Ulrich, A. Deep-Soboslay, R. Tao, BrainSeq Consortium, T. M. Hyde, J. E. Kleinman, J. A. Erwin, A. C. M. Paquola, Analysis of the caudate nucleus transcriptome in

individuals with schizophrenia highlights effects of antipsychotics and new risk genes. *Nat. Neurosci.* **25**, 1559–1568 (2022).

13. A. E. Jaffe, R. E. Straub, J. H. Shin, R. Tao, Y. Gao, L. Collado-Torres, T. Kam-Thong, H. S. Xi, J. Quan, Q. Chen, C. Colantuoni, W. S. Ulrich, B. J. Maher, A. Deep-Soboslay, BrainSeq Consortium, A. J. Cross, N. J. Brandon, J. T. Leek, T. M. Hyde, J. E. Kleinman, D. R. Weinberger, Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).

14. K. A. Perzel Mandell, N. J. Eagles, R. Wilton, A. J. Price, S. A. Semick, L. Collado-Torres, W. S. Ulrich, R. Tao, S. Han, A. S. Szalay, T. M. Hyde, J. E. Kleinman, D. R. Weinberger, A. E. Jaffe, Genome-wide sequencing-based identification of methylation quantitative trait loci and their role in schizophrenia risk. *Nat. Commun.* **12**, 5251 (2021).

15. M. Fromer, P. Roussos, S. K. Sieberts, J. S. Johnson, D. H. Kavanagh, T. M. Perumal, D. M. Ruderfer, E. C. Oh, A. Topol, H. R. Shah, L. L. Klei, R. Kramer, D. Pinto, Z. H. Gümüş, A. E. Cicek, K. K. Dang, A. Browne, C. Lu, L. Xie, B. Readhead, P. Sklar, Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).

16. M. J. Gandal, P. Zhang, E. Hadjimichael, R. L. Walker, C. Chen, S. Liu, H. Won, H. van Bakel, M. Varghese, Y. Wang, A. W. Shieh, J. Haney, S. Parhami, J. Belmont, M. Kim, P. Moran Losada, Z. Khan, J. Mleczko, Y. Xia, R. Dai, D. H. Geschwind, Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*. **362** (2018), doi:10.1126/science.aat8127.

17. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics*. **155**, 945–959 (2000).

18. Y. Nédélec, J. Sanz, G. Baharian, Z. A. Szpiech, A. Pacis, A. Dumaine, J.-C. Grenier, A. Freiman, A. J. Sams, S. Hebert, A. Pagé Sabourin, F. Luca, R. Blekhman, R. D. Hernandez, R. Pique-Regi, J. Tung, V. Yotova, L. B. Barreiro, Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*. **167**, 657-669.e21 (2016).

19. S. A. Tishkoff, F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J.-M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, S. M. Williams, The genetic structure and history of Africans and African Americans. *Science*. **324**, 1035–1044 (2009).

20. A. E. Jaffe, R. Tao, A. L. Norris, M. Kealhofer, A. Nellore, J. H. Shin, D. Kim, Y. Jia, T. M. Hyde, J. E. Kleinman, R. E. Straub, J. T. Leek, D. R. Weinberger, qSVA framework for RNA quality correction in differential expression analysis. *Proc Natl Acad Sci USA*. **114**, 7130–7135 (2017).

21. S. M. Urbut, G. Wang, P. Carbonetto, M. Stephens, Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).

22. J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, L. I. Furlong, DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)*. **2015**, bav028 (2015).

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

24. A. Zeisel, H. Hochgerner, P. Lönnerberg, A. Johnsson, F. Memic, J. van der Zwan, M. Häring, E. Braun, L. E. Borm, G. La Manno, S. Codeluppi, A. Furlan, K. Lee, N. Skene, K. D. Harris, J. Hjerling-Leffler, E. Arenas, P. Ernfors, U. Marklund, S. Linnarsson, Molecular architecture of the mouse nervous system. *Cell*. **174**, 999-1014.e22 (2018).

25. H. E. Randolph, J. K. Fiege, B. K. Thielen, C. K. Mickelson, M. Shiratori, J. Barroso-Batista, R. A. Langlois, L. B. Barreiro, Genetic ancestry effects on the response to viral infection are pervasive but cell type specific. *Science*. **374**, 1127–1133 (2021).

26. Y. Su, Y. Zhou, M. L. Bennett, S. Li, M. Carceles-Cordon, L. Lu, S. Huh, D. Jimenez-Cyrus, B. C. Kennedy, S. K. Kessler, A. N. Viaene, I. Helbig, X. Gu, J. E. Kleinman, T. M. Hyde, D. R. Weinberger, D. W. Nauen, H. Song, G.-L. Ming, A single-cell transcriptome atlas of glial diversity in the human hippocampus across the postnatal lifespan. *Cell Stem Cell*. **29**, 1594-1610.e8 (2022).

27. M. N. Tran, K. R. Maynard, A. Spangler, L. A. Huuki, K. D. Montgomery, V. Sadashivaiah, M. Tippani, B. K. Barry, D. B. Hancock, S. C. Hicks, J. E. Kleinman, T. M. Hyde, L. Collado-Torres, A. E. Jaffe, K. Martinowich, Single-nucleus transcriptome analysis reveals cell-type-specific molecular signatures across reward circuitry in the human brain. *Neuron*. **109**, 3088-3103.e5 (2021).

28. H. J. Kang, Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. M. Sousa, M. Pletikos, K. A. Meyer, G. Sedmak, T. Guennel, Y. Shin, M. B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S. N. Lisgo, A. Vortmeyer, D. R. Weinberger, N. Sestan, Spatio-temporal transcriptome of the human brain. *Nature*. **478**, 483–489 (2011).

29. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. **581**, 434–443 (2020).

30. S. De, N. Lopez-Bigas, S. A. Teichmann, Patterns of evolutionary constraints on genes in humans. *BMC Evol. Biol.* **8**, 275 (2008).

31. L. Quintana-Murci, A. G. Clark, Population genetic tools for dissecting innate immunity in humans. *Nat. Rev. Immunol.* **13**, 280–293 (2013).

32. K. Hou, Y. Ding, Z. Xu, Y. Wu, A. Bhattacharya, R. Mester, G. M. Belbin, S. Buyske, D. V. Conti, B. F. Darst, M. Fornage, C. Gignoux, X. Guo, C. Haiman, E. E. Kenny, M. Kim, C. Kooperberg, L. Lange, A. Manichaikul, K. E. North, B. Pasaniuc, Causal effects on complex traits are similar for common variants across segments of different continental ancestries within admixed individuals. *Nat. Genet.* **55**, 549–558 (2023).

33. S. Gazal, H. K. Finucane, N. A. Furlotte, P.-R. Loh, P. F. Palamara, X. Liu, A. Schoech, B. Bulik-Sullivan, B. M. Neale, A. Gusev, A. L. Price, Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

34. V. Orrù, M. Steri, G. Sole, C. Sidore, F. Virdis, M. Dei, S. Lai, M. Zoledziewska, F. Busonero, A. Mulas, M. Floris, W. I. Mentzen, S. A. M. Urru, S. Olla, M. Marongiu, M. G.

Piras, M. Lobina, A. Maschio, M. Pitzalis, M. F. Urru, F. Cucca, Genetic variants regulating immune cell levels in health and disease. *Cell*. **155**, 242–256 (2013).

35. V. Orrù, M. Steri, C. Sidore, M. Marongiu, V. Serra, S. Olla, G. Sole, S. Lai, M. Dei, A. Mulas, F. Virdis, M. G. Piras, M. Lobina, M. Marongiu, M. Pitzalis, F. Deidda, A. Loizedda, S. Onano, M. Zoledziewska, S. Sawcer, F. Cucca, Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. *Nat. Genet.* **52**, 1036–1045 (2020).

36. E. Patin, M. Hasan, J. Bergstedt, V. Rouilly, V. Libri, A. Urrutia, C. Alanio, P. Scepanovic, C. Hammer, F. Jönsson, B. Beitz, H. Quach, Y. W. Lim, J. Hunkapiller, M. Zepeda, C. Green, B. Piasecka, C. Leloup, L. Rogge, F. Huetz, Milieu Intérieur Consortium, Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. *Nat. Immunol.* **19**, 302–314 (2018).

37. S. S. Virani, A. Alonso, H. J. Aparicio, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, S. Cheng, F. N. Delling, M. S. V. Elkind, K. R. Evenson, J. F. Ferguson, D. K. Gupta, S. S. Khan, B. M. Kissela, K. L. Knutson, C. D. Lee, T. T. Lewis, J. Liu, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee., Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association. *Circulation*. **143**, e254–e743 (2021).

38. S. Prapiadou, S. L. Demel, H. I. Hyacinth, Genetic and genomic epidemiology of stroke in people of african ancestry. *Genes (Basel)*. **12** (2021), doi:10.3390/genes12111825.

39. I. I. Kessler, Epidemiologic studies of Parkinson's disease. II. A hospital-based survey. *Am. J. Epidemiol.* **95**, 308–318 (1972).

40. N. Sun, M. B. Victor, Y. P. Park, X. Xiong, A. N. Scannail, N. Leary, S. Prosper, S. Viswanathan, X. Luna, C. A. Boix, B. T. James, Y. Tanigawa, K. Galani, H. Mathys, X. Jiang, A. P. Ng, D. A. Bennett, L.-H. Tsai, M. Kellis, Human microglial state dynamics in Alzheimer's disease progression. *Cell*. **186**, 4386-4403.e29 (2023).

41. X. Liu, Y. I. Li, J. K. Pritchard, Trans effects on gene expression can drive omnigenic inheritance. *Cell*. **177**, 1022-1034.e6 (2019).

42. F. W. Albert, J. S. Bloom, J. Siegel, L. Day, L. Kruglyak, Genetics of trans-regulatory variation in gene expression. *eLife*. **7** (2018), doi:10.7554/eLife.35471.

43. K. R. Maynard, L. Collado-Torres, L. M. Weber, C. Uytingco, B. K. Barry, S. R. Williams, J. L. Catallini, M. N. Tran, Z. Besich, M. Tippani, J. Chew, Y. Yin, J. E. Kleinman, T. M. Hyde, N. Rao, S. C. Hicks, K. Martinowich, A. E. Jaffe, Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021).

44. Alzheimer's Association, 2010 Alzheimer's disease facts and figures. *Alzheimers Dement*. **6**, 158–194 (2010).

45. M. C. Power, E. E. Bennett, R. W. Turner, N. M. Dowling, A. Ciarleglio, M. M. Glymour, K. Z. Gianattasio, Trends in Relative Incidence and Prevalence of Dementia Across Non-Hispanic Black and White Individuals in the United States, 2000-2016. *JAMA Neurol.* **78**, 275–284 (2021).

46. L. L. Colbran, E. R. Gamazon, D. Zhou, P. Evans, N. J. Cox, J. A. Capra, Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nat.*

Ecol. Evol. **3**, 1598–1606 (2019).

47. A. Liston, E. J. Carr, M. A. Linterman, Shaping variation in the human immune system. *Trends Immunol.* **37**, 637–646 (2016).

48. M. Mangino, M. Roederer, M. H. Beddall, F. O. Nestle, T. D. Spector, Innate and adaptive immune traits are differentially affected by genetic and environmental factors. *Nat. Commun.* **8**, 13850 (2017).

49. M. Debnath, Adaptive immunity in schizophrenia: functional implications of T cells in the etiology, course and treatment. *J. Neuroimmune Pharmacol.* **10**, 610–619 (2015).

50. X. Li, A. Chauhan, A. M. Sheikh, S. Patil, V. Chauhan, X.-M. Li, L. Ji, T. Brown, M. Malik, Elevated immune response in the brain of autistic patients. *J. Neuroimmunol.* **207**, 111–116 (2009).

51. S. Jevtic, A. S. Sengar, M. W. Salter, J. McLaurin, The role of the immune system in Alzheimer disease: Etiology and treatment. *Ageing Res. Rev.* **40**, 84–94 (2017).

52. H. Heun-Johnson, M. Menchine, S. Axeen, K. Lung, I. Claudius, T. Wright, S. A. Seabury, Association Between Race/Ethnicity and Disparities in Health Care Use Before First-Episode Psychosis Among Privately Insured Young Patients. *JAMA Psychiatry*. **78**, 311–319 (2021).

53. J. P. Hemming, A. L. Gruber-Baldini, K. E. Anderson, P. S. Fishman, S. G. Reich, W. J. Weiner, L. M. Shulman, Racial and socioeconomic disparities in parkinsonism. *Arch. Neurol.* **68**, 498–503 (2011).

54. A. Roman-Urrestarazu, R. van Kessel, C. Allison, F. E. Matthews, C. Brayne, S. Baron-Cohen, Association of race/ethnicity and social disadvantage with autism prevalence in 7 million school children in england. *JAMA Pediatr.* **175**, e210054 (2021).

55. K. J. Benjamin, LieberInstitute/aanri_phase1. *Zenodo* (2023), doi:10.5281/zenodo.8403713.

Acknowledgments

We would like to extend our appreciation to the Offices of the Chief Medical Examiner of Washington DC, Northern Virginia, Kalamazoo Michigan, Santa Clara County, University of North Dakota, and Maryland for the provision of brain tissue used in this work. Additionally, we also extend our appreciation to the late Dr. Llewellyn B. Bigelow and members of the LIBD Neuropathology Section for their work in assembling and curating the clinical and demographic information and organizing the Human Brain Tissue Repository of the LIBD. Finally, we gratefully acknowledge the families that have donated this tissue to advance our understanding of psychiatric disorders.

The authors are indebted to many colleagues whose advice and suggestions were critical in this work, including Hailiang Huang, PhD, Kafui Dzirasa, MD, PhD, Ambroise Wonkam, MD, PhD, Yasmin Hurd, PhD, Amanda Brown, PhD, and select leaders of Black In Neuro. The African Ancestry Neuroscience Research Initiative is a collaboration between the LIBD, Morgan State University, Duke University, and members of the community led by Rev. Dr. Alvin C. Hathaway. We are also grateful to the administrative support of Jean Dubose, Yanet Raesu, and Gwenaëlle E. Thomas, PhD.

Funding

Research reported in this work was supported by LIBD, Brown Capital Management, the Abell Foundation, the State of Maryland, and the Chan-Zuckerberg Initiative. Additional support for this work was provided by the National Institute on Minority Health and Health Disparities and the National Institute of Mental Health of the National Institutes of Health under award numbers K99MD016964 to KJMB and R01MH123183 to LC-T, respectively.

Author Contributions

Conceptualization: KJMB, SH, and DRW; Methodology: KJMB, QC, NJE, LAH-M, JMS, ACMP, AEJ, SH, and DRW; Software: KJMB, QC, and SH; Formal Analysis: KJMB, QC, and SH; Investigation: JHS and TMH; Data Curation: KJMB, NJE, LC-T, GP, and JEK; Writing – Original Draft: KJMB, QC, SH, and DRW; Writing – Review & Editing: KJMB, QC, LAH-M, LC-T, ACMP, TMH, JEK, AEJ, SH, and DRW; Visualization: KJMB and SH; Supervision: KJMB, SH, and DRW; Project Administration: KJMB and SH; Funding Acquisition: KJMB, LC-T, SH, and DRW.

Conflict of interest

AEJ is currently an employee and shareholder of Neumora Therapeutics, which is unrelated to the contents of this manuscript. DRW serves on the Scientific Advisory Boards of Sage Therapeutics and Pasithea Therapeutics. All other authors declare no competing interests.

Data availability

Publicly available BrainSeq Consortium total RNA DLPFC and hippocampus RangedSummarizedExperiment R Objects with processed counts are available at http://eqtl.brainseq.org/phase2/. Publicly available BrainSeq Consortium total RNA caudate RangedSummarizedExperiment R Objects with processed counts are available at http://erwinpaquolalab.libd.org/caudate_eqtl/. Publicly available dentate gyrus RangedSummarizedExperiment R Objects with processed counts and phenotype information are

available at http://research.libd.org/dg_hippo_paper/data.html. Analysis-ready genotype data will be shared with researchers that obtain dbGaP access. FASTQ files are available for total RNA dentate gyrus, DLPFC, and hippocampus via Globus collections jhpce#bsp2-dlpfc and jhpce#bsp2-hippo available at https://research.libd.org/globus/. For the caudate, FASTQ files are available via dbGaP.

We used publicly available single cell datasets. Glial subpopulation single-cell data from the human postmortem hippocampus astrocyte, microglia, and oligodendrocyte lineage is available from UCSC cell browser (*"Human Hippocampus Lifespan"* collection). The human PBMCs single-cell data is available from Zenodo (10.5281/zenodo.4273999). Multiple human brain region single-cell datasets (i.e., DLPFC, hippocampus, nucleus accumbens, amygdala, and subgenual anterior cingulate cortex) are available by brain region from GitHub (https://github.com/LieberInstitute/10xPilot_snRNAseq-human). Human microglial state dynamics in Alzhiemer's disease single-cell data is available from http://compbio.mit.edu/microglia_states/.

Code availability

All code and Jupyter Notebooks are available through GitHub at https://github.com/LieberInstitute/aanri_phase1 with more detail (*55*).

List of Supplementary Materials

Figure captions

Fig. 1: Study design for the examination of the genetic and environmental contributions to genetic ancestry-associated expression differences. BA stands for Black Americans and WA for White Americans.

Fig. 2: Extensive ancestry-associated expression changes across the brain region. A. Circos plot showing ancestry DEGs across the caudate (red), dentate gyrus (blue), DLPFC (green), and hippocampus (purple). **B.** Gene set enrichment analysis (GSEA) of differential expression analysis across brain regions, highlighting terms associated with increased AA (African ancestry) or EA (European ancestry) proportions. **C.** UpSet plot showing large overlap between brain regions. Green is shared across the four brain regions; blue, shared across three brain regions; orange, shared between two brain regions; and black, unique to a specific brain region. * Indicating significant pairwise enrichment (Fisher's exact test) or significant overlap between all four brain regions (Monte Carlo simulation). **D.** Heatmaps of the proportion of ancestry DEG sharing with concordant direction (sign match; top) and within a factor 0.5 effect size (bottom) **E.** Metaplot showing examples of brain region-specific ancestry effects.

Fig. 3: **Ancestry-associated genes and canonical transcripts are evolutionarily less constrained. A.** Significant depletion of ancestry DEGs for evolutionarily constrained genes (canonical transcripts) across brain regions. Significant depletion/enrichments (two-sided, Fisher's exact test, FDR corrected p-values, -log10 transformed) are annotated within tiles. Odds ratios (OR) are $\log_2$ transformed to highlight depletion (blue) and enrichment (red). **B.** Similar trend of depletion of ancestry DE transcripts (DETs; all, canonical, and non-canonical) for evolutionarily constrained transcripts across brain regions. Odds ratios are $\log_2$ transformed to highlight depletion (blue) and enrichment (red). **C.** The mean of ancestry-associated DE feature (i.e., gene and transcript) lfsr as a function of LOEUF (loss-of-function observed/expected upper bound fraction) decile shows a significant negative correlation for genes (left; for the caudate, dentate gyrus, DLPFC, and hippocampus: two-sided, Pearson, r = -0.20, -0.20, -0.21, and -0.21; p-value = 3.0×10^{-122}, 7.6×10^{-113}, 8.6×10^{-126}, and 1.2×10^{-122}) and transcripts (right; for the caudate, dentate gyrus, DLPFC, and hippocampus: two-sided, Pearson, r = -0.05, -0.05, -0.04, and -0.04; p-value = 8.6×10^{-13}, 1.7×10^{-11}, 9.0×10^{-11}, and 3.2×10^{-10}). Error bars correspond to 95% confidence intervals.

Fig. 4: Genetic contribution of genetic ancestry differences in expression across the brain. A. UpSet plot showing large overlap between brain regions of eGenes. **B.** Heatmap of the proportion of ancestry DEG sharing with concordant direction (sign match). **C.** Significant enrichment of ancestry-associated DE genes for eGenes (unique gene associated with an eQTL) across brain regions separated by direction of effect (increased in AA or EA proportion). **D.** Density plot showing significant increase in absolute allele frequency differences (AFD; one-sided, Mann-Whitney U, p-value < 0.05) for global ancestry-associated DEGs (red) compared with non-DEGs (blue) across brain regions. A dashed line marks the mean absolute AFD. Absolute AFD calculated as the average absolute AFD across a gene using significant eQTL (lfsr < 0.05). **E.** Correlation (two-sided, Spearman) of elastic net predicted (y-axis) versus observed (x-axis) ancestry-associated differences in expression among ancestry-associated DEGs with an eQTL across brain regions. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

Fig. 5: Unknown environmental factors are primary drivers of nearby global ancestry-associated DEGs. A. Circos plot showing local ancestry-associated DMRs across the caudate (red), DLPFC (blue), and hippocampus (green). Methylation status is annotated in red for hypermethylation and blue for hypomethylation. **B.** Gene term enrichment of DMRs across brain regions. **C.** Histograms showing distribution of ΔP_{ST} associated with the impact of unknown environmental factors as captured by residualized VMR (corrected by local ancestry, age, sex, and unknown biological factors captured by PCA) for nearby global ancestry-associated DEGs. A dashed line marks the mean ΔP_{ST}. A solid line shows the density overlay.

Fig. 6: Global ancestry-associated DEGs stratified by coding or non-coding DEGs show general enrichment for heritability of several neurological and immune-related traits, but depleted for brain-related behavioral traits. Heatmap for ancestry-associated DEGs that show enrichment (red) or depletion (blue) for heritability of brain- and immune-related traits from S-LDSC analysis. Significant enrichment for heritability traits disappears when limited to non-coding DEGs. Numbers within tiles are levels of enrichment (> 1) or depletion (< 1) that are significant after multiple testing correction (FDR < 0.05). The left panel shows results for all DEG in each brain region. The middle and right panels show results for DEG increased with AA or EA proportions for each brain region, respectively.

Supplementary materials

Material & Methods

The research described herein complies with all relevant ethical regulations. Additionally, we declare that all specimens used in this study were obtained with informed consent. We obtained informed consent from the next kin under protocols No. 12-24 (the Department of Health and Mental Hygiene for the Office of the Chief Medical Examiner for the State of Maryland) and No. 20111080 (the Western Institutional Review Board for the Offices of the Chief Medical Examiner for Kalamazoo Michigan, University of North Dakota in Grand Forks North Dakota, and Santa Clara County California). We obtained samples at the Clinical Brain Disorder Branch (CBDB) at the National Institute of Mental Health (NIMH) from the Northern Virginia and District of Columbia Medical Examiners' Office, according to NIH Institutional Review Board guidelines (Protocol #90-M-0142). The LIBD received the tissues by donation under the terms of a material transfer agreement. The Institutional Review Board of the University of Maryland at Baltimore and the State of Maryland approved the study protocols that collected these brain regions (*10–12*). Details of case selection, curation, diagnosis, and anatomical localization and dissection can be found in previous publications from our research group (*10–12*).

BrainSeq consortium RNA-sequencing data processing

We surveyed covariates, FASTQ files, SNP array genotypes, RNA degradation metrics obtained with the qSVA methodology (*20*), phenotype information, and raw counts (gene, transcript, exon, and exon-exon junction) for the caudate, dentate gyrus, DLPFC, and hippocampus from the BrainSeq Consortium (*10*, *12*) and http://research.libd.org/dg_hippo_paper/data.html (*11*).

BrainSeq consortium genotype imputation

General imputation

Samples were genotyped and imputed as part of the full LIBD cohort, using procedures as previously described (*10, 12, 13*). Briefly, samples were genotyped on four different types of Illumina microarrays over the years (HumanHap650, Human1M, HumanOmni2.5, or HumanOmni5-Quad BeadChips). We merged samples genotyped by the same type of microarray and followed standard pre-imputation quality control (QC) to remove low quality (Hardy-Weinberg equilibrium p-value < 1×10^{-6}) and low frequency (minor allele frequency [MAF] < 0.005) variants. We converted genotype positions from hg19 to hg38 with liftOver (*56*). Once converted, we imputed genotypes, separately by genotyping array, on the Trans-Omics for Precision Medicine (TOPMed) imputation server (*7*, *57*, *58*) using the Haplotype Reference Consortium (HRC) reference panels. We phased genotypes per chromosome using eagle (version 2.4; (*59*)). We performed post-imputation QC of each imputed dataset for Black and White American samples separately.

We filtered out variants with low-quality imputation scores (R^2 < 0.8) and removed variants: 1) MAF less than 0.05, 2) missing call frequencies greater than 0.1, or 3) Hardy-Weinberg equilibrium below p-value of 1e-10 using PLINK2 (version 2.00a3LM; (*60*)). We then merged imputed genotypes across four genotyping platforms based on overlapping filtered imputed variants. This resulted in 6,225,756 and 6,097,532 common variants for Black and White American donors, respectively.

HLA imputation

For HLA allele imputation, we extracted the extended MHC region on chromosome 6 from pre-imputed QC'd, genotypes (hg38) by genotype array (see **General imputation**) with PLINK2. We performed HLA imputation on the Michigan Imputation Server (*57*) using the four-digit, multi-ethnic HLA imputation reference panel (version 2) (*61*). Similar to general imputation, we phased genotypes using eagle on the server. Following imputation, we filtered low-quality imputation scores ($R^2 < 0.7$) per genotype array with BCFtools (version 1.13; (*62*)). We then merged imputed genotypes across the four genotyping arrays with BCFtools and extracted HLA alleles from the VCF file. This resulted in a total of 2,850 HLA alleles.

BrainSeq consortium DNA methylation data processing

We generated WGBS datasets in our previous studies for three adult brain regions (DLPFC, hippocampus and caudate). Details about study samples, data generation, and data processing have been described in our prior reports (*14*, *63*). Briefly, we assessed quality control with FastQC. Following assessment with FastQC, we removed adapter content with Trim Galore (*64*). We aligned trimmed reads with Arioc (*65*) to the hg38 genome build (GRCh38.p12) and removed duplicate alignments with SAMBLASTER (*66*). After removing duplicates, we filtered alignments with samtools (*67*) (v1.9) to include only primary alignments with a mapping quality (MAPQ) ≥ 5. From these filtered alignments, we extracted methylation data using Bismark methylation extractor (*68*). Following methylation extraction, we processed and combined DNA methylation proportions across samples using bsseq (version 1.18; (*69*)), an R/Bioconductor package. We locally smoothed methylation data with BSmooth using default parameters. We filtered the resulting CpG data to remove: 1) CpGs within the blacklist regions and 2) CpGs with coverage < 3.

Subject selection and details

We selected samples per brain region using five common inclusion criteria: 1) RiboZero RNA-sequencing library preparation, 2) recent African ancestry, 3) TOPMed imputed genotypes available, 4) adults (age > 17) and 5) diagnosis of neurotypical control. This resulted in a total of 425 samples from 151 unique individuals across the caudate (n=121), dentate gyrus (n=47), DLPFC (n=123), and hippocampus (n=133). Subject details are summarized in **Table S1**.

Estimation of genome-wide admixture levels

We estimated the admixture proportion for each individual based on SNPs that were informative with respect to ancestry using the STRUCTURE program. We selected 1,634 such SNPs based on genetic information downloaded from the 1000 Genomes CEU (Northern Europeans from Utah) and AFR (African ancestry superpopulation, including Esan, Gambian, Luhyu, Mende, and Yoruba populations) samples. Markers were chosen based on the following criteria: 1) absolute difference (δ) in allele frequency between the two ancestry populations > 0.5; 2) r^2 between each pair of SNPs < 0.1 within each population; 3) p-value > 0.01 for testing Hardy-Weinberg equilibrium within each population; and 4) successfully imputed in our brain samples (info > 0.8). The structure was run within a two-ancestry population model with 5,000 burn in and 10,000 iterations.

Estimation of local ancestry

We used RFMIX (*70*), a discriminative modeling approach for rapid and robust local-ancestry inferences, to infer local ancestry in our admixed samples using the European and African ancestry samples from the 1000 Genomes project (*71*) as reference. We extracted the posterior probability of African ancestry at each SNP per haplotype from the Forward–Backward output of RFMIX. Local ancestry for a genomic region was then estimated as the average African ancestry across all SNPs within the region. As RFMIX also computed and output a global ancestry estimate for each sample, we compared global ancestry estimates between STRUCTURE and RFMIX and observed a high correlation between estimates from the two programs (Spearman correlation, rho = 0.99).

Differential expression analysis

Cell-type deconvolution analysis

Deconvolution was performed with the ReferenceBasedDecomposition function from the R package BisqueRNA version 1.0.4 (*72*), using the *use.overlap = FALSE* option. The single cell reference data set used is single nucleus RNA-seq from the 10X protocol, which includes tissue from eight donors and five brain regions (*27*). The ten cell types considered in the deconvolution of the tissue were astrocytes (Astro), endothelial (Endo), microglia (Micro), macrophage (Macro), mural cells (Mural), oligodendrocytes (Oligo), oligodendrocyte progenitor cells (OPC), T cells (Tcell), excitatory neurons (Excit), and inhibitory neurons (Inhib). Marker genes were selected by first filtering for genes common between the bulk data and the reference data and then calculating the ratio of the mean expression of each gene in the target cell type over the highest mean expression of that gene in a non-target cell type. The 25 genes with the highest ratios for each cell type were selected as markers.

Quality control and identification of relevant confounders

To evaluate potential sources of confounding for expression and genetic ancestry, we first correlated technical and RNA quality variables available from the downloaded R variables and removed highly correlated variables (Pearson, $r > 0.95$) present in two or more brain regions. Following this, we retained variables common across the four brain regions. In addition to these variables, we also accounted for hidden variables using the downloaded qSVA (**Fig. S39** and **Equation 1**; k=13, 6, 9, and 14, for the caudate, dentate gyrus, DLPFC, and hippocampus, respectively). We have found qSVs also accurately correct for observed variables like batch effect and cell-type composition (*12*, *20*).

$$E(Y) = \beta_0 + \beta_1 Ancestry + \beta_2 Sex + \beta_3 Age + \beta_4 MitoRate + \beta_5 rRNArate + \beta_6 TotalAssignedGenes + \beta_7 OverallMappingRate + \sum_{i=1}^{k} \gamma_i qSV_i$$

Equation 1

Given the potential influence of cell composition on gene expression, we also examined cell-type proportion associated with genetic ancestry and any potential confounding effects on gene expression. To this end, we performed cell-type deconvolution (**Data S13**). When we examined the Black American population, we found that the majority of cell types across brain regions showed no correlation with genetic ancestry (**Fig. S40**), and only oligodendrocytes in the DLPFC showed a significant association (Spearman, p-value < 0.05) with genetic ancestry. In contrast, when we included White American donors, we found that seven of the ten cell types showed a significant association (Spearman, p-value < 0.05) with genetic ancestry in at least one brain region (**Fig. S41**).

These cell-type proportions also showed high correlation with confounders (**Fig. S42**). As such, our model also accounted for cell-type proportions for each brain region (**Fig. S43**).

Global ancestry-associated differential expression analysis

We performed differential expression analysis using mash modeling in R. Initially, we determined effect size and standard error of effect size using limma-voom modeling as previously described (*12*). Briefly, we filtered low-expressing genes using *filterByExpr* from edgeR (*73*, *74*) and normalized library size. Next, we applied voom normalization (*75*) as a model of genetic ancestry adjusted for age, RNA quality (mitochondria mapping rate, gene assignment rate, genome mapping rate, rRNA mapping rate, and hidden variance using qSVA; **Equation 1**). Following voom normalization, we fit the model using eBayes and extracted out effect size (log fold-change) and standard error of effect size from the model (**Equation 2**) by brain region for each feature (gene, transcript, exon, and junction).

$$SE = \frac{\sigma}{\sqrt{n}}$$

Equation 2

Next, we implemented mash modeling using mashr (version 0.2.57) for each feature using the limma-voom extracted effect sizes and standard errors across brain regions. We learned the correlation structure across the brain regions and used all features as an unbiased representation of the results to account for overlapping samples. Following this, we calculated the canonical covariances. A strong set of features was determined condition by condition using *mash_1by1*, and data-driven covariance was calculated with the strong set of features. Once calculated, we fit the mash model to the full set of features and computed the posterior summaries for all features. Features were considered significant if they had a lfsr less than 0.05.

Local ancestry-associated differential expression analysis

For local ancestry differential expression analysis, we first calculated a local African ancestry score per feature (i.e., gene, transcript, exon, and junction). Here, we averaged all haplotypes within a 200kbp window of each feature using the RFMIX results. Following this estimate of local African ancestry per feature, we applied a separate linear model per feature using **Equation 1** modified for local ancestry. We limited our analysis to features tested for global ancestry differential expression. As each model was per feature, we replaced voom normalized with CPM log-normalized counts. We fit our model with limma lmFit and extracted effect size and standard error for downstream mash modeling as described in "*Global ancestry-associated differential expression analysis*". We compared local and global ancestry DE results and found large overlap (**Fig. S44**).

Expression residualization

For residualized expression, we regressed out covariates from voom normalized expression using a null model (**Equation 3**) and applied a z-score normalization as previously described (*12*).

$$E(Y) = \beta_0 + \beta_1 Sex + \beta_2 Age + \beta_3 MitoRate + \beta_4 rRNArate + \beta_5 TotalAssignedGenes + \beta_6 OverallMappingRate + \sum_{i=1}^{k} \gamma_i qSV_i$$

Equation 3

MHC region enrichment

To examine the contribution of the MHC region for immune-related pathway enrichment, we extracted genes within the MHC from hg38 annotation (GENCODE v25). Specifically, we extracted genes from the MHC region (chr6:28,510,120-33,480,577) and the extended MHC region (chr6:25,726,063-33,400,644) using PyRanges (*76*) and gtfparse. We further subset the extended MHC region for any gene names that started with HLA. Following this, we assessed enrichment for the MHC regions (i.e., MHC region, extended MHC region, and HLA genes) using a two-sided, Fisher's exact test. We corrected for multiple testing with Benjamini-Hochberg.

Public data comparison and enrichment analysis

For public data comparison, we downloaded the ancestry-associated DEGs in immune cells (*18*) and immune function GWAS prioritized genes (*34–36*). We assessed enrichment with our ancestry-associated DEGs using a two-sided, Fisher's exact test and corrected for multiple testing with Benjamini-Hochberg.

Single-cell specificity and cell-type enrichment analysis

To understand the cellular context of ancestry-associated DEGs in the human brain, we performed cell-type enrichment analysis by leveraging existing gene expression data from 39 broad categories of cell types from the mice central and peripheral nervous system (*24*). Specifically, we examined the overlap between DEGs and cell-type-specific genes for each cell type defined in a previous study (*77*). We assessed enrichment for each brain cell type using a two-sided, Fisher's exact test. We corrected for multiple testing with Benjamini-Hochberg.

We next expanded our cell-type enrichment analysis to single-cell datasets with glial (i.e., astrocyte, microglia, and oligodendrocyte) subtype annotation and non-brain immune cells (i.e., peripheral blood mononuclear cells [PBMCs]). For glial subpopulations, we downloaded human postmortem hippocampus astrocyte, microglia, and oligodendrocyte lineage single-cell data (*26*) from UCSC cell browser (*78*). For PBMCs, we downloaded human PBMC single-cell data (*25*) from Zenodo (10.5281/zenodo.4273999).

To calculate cell-type specificity, we adapted cell-type specificity code from https://github.com/jbryois/scRNA_disease/blob/master/Code_Paper/Code_Zeisel/get_Zeisel_Lvl4_input.md (*77*) for these additional datasets. Briefly, we converted Seurat objects (*79*) into SingleCellExperiment (*80*) in R (version 4.3). Next, we aggregated mean counts across annotated cell types with scuttle (*81*). Following aggregation, we removed genes with zero expression and applied transcripts per million (TPM) normalization. Across all cell types, we calculated a specificity score for each gene defined as the proportion of total expression of a gene. To assign marker genes based on cell specificity, we filtered out genes with less than 1 TPM and selected the top 10% of genes based on specificity score for each cell type. We used these marker genes to assess enrichment of ancestry-associated DEGs using a two-sided, Fisher's exact test and corrected for multiple testing with Benjamini-Hochberg.

For disease single-cell enrichment, we downloaded marker genes and Alzhiemer's differential expression results for each microglial state (*40*) from http://compbio.mit.edu/microglia_states/. For

enrichment analysis, we applied two-sided, Fisher's exact test using all annotated genes as a universe. We corrected for multiple testing with Benjamini-Hochberg.

Glial cell composition across multiple brain region

To investigate glial cell composition across the caudate, DLPFC, and hippocampus, we downloaded single-cell datasets from multiple brain regions (*27*) similar to ours (i.e., nucleus accumbens, DLPFC, and hippocampus). To integrate three brain regions single-cell data, we modified the across regions analysis script from https://github.com/LieberInstitute/10xPilot_snRNAseq-human/blob/master/10x_across-regions-analyses_step02_MNT.R. Specifically, we cleaned the annotated datasets, removing pre-calculated metrics. Following this, we combined the data and normalized with multiBatchNorm from batchelor R package (*82*). Next, we subset the data set specifically for annotated glial cells (i.e., microglia, astrocyte, and oligodendrocyte lineage).

To annotate the glia subpopulation to the multiple brain region dataset, we first converted R objects to H5AD files using zellkonverter (https://github.com/theislab/zellkonverter). We integrated the multi-brain region combined dataset (*27*) with glia subpopulation dataset (*26*) using single-cell variational inference (scVI; (*83*)) from scvi-tools (*84*) per glia subpopulation. Following integration, we transferred the glia subpopulation annotations to the multi-brain region dataset with single-cell annotation using variational inference (scANVI; (*85*)) from scvi-tools. We visualized glia subpopulation clustering after removing batch effects from the PCA subspace with fastMNN from batchelor package and applying tSNE using scater package (*81*).

To test glial cell composition differences across brain regions, we applied the propeller function from the speckle package in R (version 4.3; (*86*)) with arcsin transformed counts. The propeller function corrected for multiple testing.

Binary contrast of Black and White Americans

For internal validation of global ancestry-associated DE features (i.e., gene, transcript, exon, and junction), we performed differential expression analysis with a combination of Black and White American individuals using mash. Similar to *"Global ancestry-associated differential expression analysis",* we determined effect size and standard error of effect size using limma-voom modeling. Here, we replaced the continuous variable genetic ancestry with the binary, self-reported race. Additionally, we selected individuals with limited admixture by including: 1) Black Americans with African genetic ancestry greater than or equal to 0.8 and 2) White American with European genetic ancestry greater than 0.99. To limit the influence of the larger sample size compared to *"Global ancestry-associated differential expression analysis"*, we randomly sampled ten times without replacement to approximately the admixed Black American-only analysis sample size. Following extraction of effect sizes and standard errors, we implemented mash modeling for each feature across brain regions as described in *"Global ancestry-associated differential expression analysis"*.

Immune variation modeling

To remove the potential effect of immune variation, we added HLA variation (**Equation 4**) or glial cell proportion (astrocytes [Astro], microglia [Micro], macrophage [Macro], oligodendrocytes [Oligo], oligodendrocyte progenitor cells [OPC], and T cells [Tcell]; **Equation 5**) to our DE model as covariates. Previously, we found only the oligodendrocytes in the DLPFC showed a significant association (Spearman, p-value < 0.05; **Fig. S40**) with genetic ancestry (see **Quality control and identification of relevant confounders**). Given the potential correlation between HLA variation and

global genetic ancestry, we first examined HLA variation association with global genetic ancestry. For this, we first generated HLA variation PCs by applying PCA on the 2,850 HLA imputed alleles. We found limited correlation between the ten PCs and global genetic ancestry (Spearman, p-value < 0.05; **Fig. S45**).

$$E(Y) = \beta_0 + \beta_1 Ancestry + \beta_2 Sex + \beta_3 Age + \beta_4 MitoRate + \beta_5 rRNArate + \beta_6 TotalAssignedGenes + \beta_7 OverallMappingRate + \sum_{i=1}^{k} \gamma_i qSV_i + \sum_{j=1}^{5} \sigma_j HLA_j$$

Equation 4

$$E(Y) = \beta_0 + \beta_1 Ancestry + \beta_2 Sex + \beta_3 Age + \beta_4 MitoRate + \beta_5 rRNArate + \beta_6 TotalAssignedGenes + \beta_7 OverallMappingRate + \sum_{i=1}^{k} \gamma_i qSV_i + \beta_8 Astro + \beta_9 Macro + \beta_{10} Micro + \beta_{11} Tcell + \beta_{12} Oligo + \beta_{13} OPC$$

Equation 5

Weighted correlation network analysis

We performed a signed-network WGCNA (*23*) analysis using residualized expression to generate the co-expression network with neurotypical control individuals (n=151 Black Americans) in a single block by brain region. For this analysis, we filtered genes and outlier individuals with the WGCNA function *goodSamplesGenes*. Following this, we applied additional sample filtering based on sample expression with a total Z-normalized distance of 2.5 or greater from other samples. After evaluating power and network connectivity for each brain region, we selected a soft power of 12.

For network construction, we used *bicor* correlation and the following parameters: 1) *mergeCutHeight* set to 0.3 for the dentate gyrus and default values for the caudate, DLPFC, and hippocampus and 2) *minModuleSize*set to 30 for the dentate gyrus and default values for the caudate, DLPFC, and hippocampus. We set all other parameters to default values. The co-expression network was made using Pearson correlation values for the caudate (117 samples; 19,883 genes), dentate gyrus (46 samples; 18,747 genes), DLPFC (121 samples; 20,070 genes), and hippocampus (128 samples; 19,794 genes). We determined significant associations with ancestry using a linear model that correlates ancestry proportions (see **Estimation of genome-wide admixture levels**) with module eigengenes.

For each module, we calculated overlap enrichment/depletion with ancestry-associated DEGs (at FDR < 0.05) separated by direction of effect (such as DEGs that are upregulated in AA, upregulated in EA, or upregulated in either ancestry) using the two-sided Fisher's exact test in Python with SciPy (*87*) stats module. The following p-values were corrected using statsmodels (*88*) stats module with the Benjamini-Hochberg method in Python.

When we examine the most significantly enriched modules for ancestry-associated DEGs upregulated in Black American individuals across brain regions, we found the cyan module (enriched for response to virus) for the caudate; the pink module (enriched for wound healing and cell migration) for the dentate gyrus; the saddlebrown module (enriched for cellular response to virus) for the DLPFC; and the yellow module (enriched for cilium movement and assembly) for the hippocampus (**Figure S7A** and **Data S4**). In contrast, when we examined the most significantly enriched modules for ancestry-

associated DEGs downregulated in proportion to Black American individuals across brain regions, we found the greenyellow module (enriched for inflammatory response) for the caudate; the saddlebrown module (enriched for immune response) for the dentate gyrus; the pink module (enriched for immune response) for the DLPFC; and the blue module (enriched for immune response) for the hippocampus (**Figure S7B** and **Data S4**). Although the caudate and DLPFC showed modules enriched for immune response for both directions of effect, the most significantly enriched non-gray module (two-sided, Fisher's exact test) was associated with a specific direction of effect consistent with DE analysis for the caudate (cyan module, DEGs upregulated in African ancestry) and DLPFC (pink module, DEGs downregulated in African ancestry).

Gene term enrichment analysis

Differential expression analysis: gene term enrichment and hypergeometric

We determined significant enrichment for gene sets using the gene set enrichment analysis (GSEA) (*89*, *90*), which is less susceptible to gene length bias because it uses permutation enrichment within gene sets. In this study, we performed GSEA with gseGO (GO gene set database) from the clusterProfiler package (*91*) and gseDGN (DisGeNET gene set database (*22*)) from the DOSE package (*92*). We defined the gene set "universe" as all unique genes tested for differential expression. When examining isoform-level enrichment (transcript, exon, or junction), we selected, for each unique gene, the feature with the largest absolute effect size. For gseGO, minimal gene set size (minGSSize) was set to 10, maximum gene set size (maxGSSize) set to 500, and p-value cutoff set to 0.05. For gseDGN, minGSSize was set to five and p-value cutoff to 0.05. We used the default settings for all other parameters.

For hypergeometric analysis, we used enrichGO and enrichDGN from the clusterProfiler and DOSE packages, respectively. Similar to GSEA analysis, we defined the gene set "universe" as all unique genes tested for differential expression.

Co-expression network analysis: gene term enrichment

For gene term enrichment analysis, we used GOATOOLS Python package (*93*) using hypergeometric tests with the GO database. Similar to *"Differential expression analysis: gene term enrichment"*, we defined the gene set universe as all unique genes tested from differential expression analysis.

Enrichment of evolutionary constraint

For evolutionary constraint enrichment analysis, we downloaded genome aggregation database (gnomAD; version 2) gene- and transcript-level loss-of-function metrics (*29*). We assessed enrichment with the observed/expected loss-of-function upper bound fraction (LOEUF) using the decile bins. Additionally, we tested correlation between ancestry-associated differentially expressed features (i.e., genes and transcripts) and the LOEUF with a two-sided Pearson's correlation. We corrected both statistical tests for multiple testing using Benjamini-Hochberg.

Expression quantitative trait loci analysis

We performed all cis-eQTL mapping for neurotypical controls (Black American individuals, age > 17; **Table S1**) using tensorQTL, which leverages GPUs to substantially increase computational speed (*94*). Initially, we filtered low expression as previously described (*12*) using the GTEx Python script

(i.e., eqtl_prepare_expression.py) with modifications for isoform-level genomic features (i.e., transcripts, exons, and junctions). This script retained features with expression estimates greater than 0.1 TPM in at least 20% of samples and aligned read counts of six or more. Additionally, this script used Python functions defined by rnaseqnorm.py to normalize counts with TMM, a Python port of edgeR function.

To generate the TPM files as input for the eqtl_prepare_expression.py, we used effective length (**Equation 6**). For genes and exons, we calculated effective length (**Equation 7**) using mean insert size from Picard tools CollectInsertSizeMetrics (http://broadinstitute.github.io/picard/). For junctions, we fixed effective length at 100. After calculating effective length, we dropped any feature with an effective length less than or equal to one.

$$TPM = 1e6 \times \frac{Count\ /\ Effective\ Length}{\Sigma(Count\ /\ Effective\ Length)} \qquad \textbf{Equation 6}$$

$$\boldsymbol{Effective\ Length = Length - [Mean\ Insert\ Size] + 1} \qquad \textbf{Equation 7}$$

Main effect analysis

For main effect cis-eQTL mapping, we quantified the effects of unobserved confounding variables on expression after adjusting for sex, population stratification (SNP PCs 1-5), and *k* unobserved confounding variables on expression. We determined these variables via *num.sv* function (*vfilter* set to 50,000) from sva, an R/Bioconductor package (*95*), and principal components analysis (PCA) of expression for each feature. To identify *cis*-eQTL, we implemented nominal mapping, adjusting for covariates with a mapping window within 0.5 Mb of the TSS of each feature and a minor allele frequency ≥ 0.01. The tensorQTL used a two-sided *t*-test to estimate the nominal p-value for each variant-gene pair. To generate a subset of "strong" signals for downstream mash modeling in R, we also performed adaptive permutations. Following this, the empirical p-values were corrected for multiple testing across features using Storey's q-value method (*96*, *97*). This resulted in a file with the top variant for each feature. In addition to this permutation analysis, we also performed conditional analysis. This resulted in additional feature-variant pairs to generate our set of "strong" associations for mash modeling.

Ancestry-dependent interaction analysis

For genetic ancestry-dependent cis-eQTL mapping, we used the confounders generated from main effect analysis but removed variables associated with population stratification (SNP PCs 1-5). To identify genetic ancestry-dependent *cis*-eQTL, we implemented nominal mapping, adjusting for covariates with a mapping window within 0.5 Mb of the TSS of each feature and a MAF greater than or equal to 0.05. To generate a subset of strong signals for downstream mash modeling, we performed eigenMT (*98*) by setting *run_eigenmt* to True. This resulted in a file with the top variant for each feature.

Integration with mash modeling in R

To assess sharing across brain regions and to increase our power to detect main and interacting eQTL effects within admixed Black American only individuals, we used the multivariate adaptive shrinkage framework as previously described (*12*). We extracted effect sizes and standard errors for these effect sizes from the nominal results for either main or interacting cis-eQTL. To specify a correlation structure across brain regions (i.e., overlapping sample donors), we used the

estimate_null_correlation_simple function prior to fitting the mash model. The mash model included both the canonical covariance matrices and the data-driven covariance matrices learned from our data.

We defined the data-driven covariance matrices as the top four PCs from the PCA performed on the "strong" signals. For gene level analysis, we defined a set of "strong" tests running a simple condition-by-condition (*mash_1by1*) analysis as described in "*Differential expression analysis"*. For isoform-level analysis (i.e., transcript, exon, and junction), we defined a set of "strong" tests using either the results from permutation or eigenMT analyses. Specifically, for main effect analysis, the set of "strong" tests were selected if a feature-variant pair was present in at least one brain region within permutation or conditional analyses. For interaction analysis, we selected the set of "strong" tests if a feature-variant pair was present in at least one brain region from the eigenMT top associations.

To learn the mixture weights and scaling for the main and interacting effects, we initially fit the mash model with a random set (i.e., unbiased representation of the results) of the nominal eQTL results (i.e., 5% for gene-variant pairs and 1% for transcript-, exon-, and junction-variant pairs). We next fitted these mixture weights and scaling to all of the main and interacting eQTL results in chunks. Following model fitting, we extracted posterior summaries and measures of significance (i.e., lfsr). We considered main and interacting eQTL significant if the lfsr was less than 0.05.

Absolute allele frequency difference

To calculate absolute allele frequency differences, we first calculated allele frequency within the 1000 Genome Project AFR (super population) and EUR (super population) reference genome using PLINK per chromosome. Prior to allele frequency calculation, we filtered SNPs based on a MAF of 0.01 for AFR and 0.005 for EUR. To calculate differences between the two super populations, we matched SNP and reference allele before calculating absolute frequency differences (**Equation 8**). We assessed absolute allele frequency differences for ancestry-associated DEGs compared with other eGenes using two methods: 1) top SNP per gene and 2) average SNPs across the gene.

$$AFD = |AFR - EUR| \qquad \textbf{Equation 8}$$

Genetic control of ancestry effects on expression

We estimated the predicted *cis*-genetic population differences in expression by first computing predicted expression from genotype dosage (0, 1, or 2; see below). With these predicted expression values, we performed differential expression for genetic ancestry using a model analogous to **Equation 1** (see **Global ancestry-associated differential expression analysis**) to obtain predicted genetic ancestry effects. We extracted the observed population differences in expression from the effect sizes estimated after applying mash as described in *"Ancestry-associated differential expression analysis"*.

Expression residualization for prediction models

To generate residualized expression for our prediction models, we fit a linear model with *lmFit* from limma to normalized expression (see **Expression quantitative trait loci analysis**) and covariates (see **Differential expression analysis**; **Equation 3**). Using this model, we regressed out covariates from normalized expression using the *residuals* function in R (version 4.0.3).

Calculating predicted expression using genetic variants in a linear model

For our linear model, we extracted the posterior effect size of the top genetic variant from the mash model for each feature (gene, transcript, exon, and junction). We imputed residualized expression using an individual's genotype dosage *(j)* and feature *(i)* posterior effect size (**Equation 9**) using PyTorch (*99*).

$$predicted\ expression_{i,j} = effect\ size\ (eQTL)_i * genotype_j \quad \textbf{Equation 9}$$

Calculating predicted expression using genetic variants in an elastic net model

We selected all genetic variants within ± 500kb of the gene body. We removed variants with missing genotypes and filtered variants based on a MAF threshold of 0.01 and a Hardy-Weinberg equilibrium below a p-value of 1e-5. We used an elastic net model, ideal for relatively smaller sample sizes. For our elastic net model, we fitted a sparse linear regression model using *big_spLinReg* from the bigstatsr R package (v1.5.12; (*100*)). We tuned the alpha parameter using a sequence of 20 alphas (i.e., 0.05 to 1 using 0.05 step size). Additionally, we used four sets for the "Cross-Model Selection and Averaging" procedure. We averaged feature weights for genetic variants across *k*-folds (five folds for each of the caudate, DLPFC, and hippocampus; and three folds for dentate gyrus). We imputed residualized expression with these feature weights *(i)* and an individual's genotype dosage *(j)* (**Equation 10**). We calculated the correlation coefficient *(r)* using Pearson's correlation on the test samples for each *k*-fold.

$$predicted\ expression_{i,j} = variant\ weight_i * genotype_j \quad \textbf{Equation 10}$$

Linkage disequilibrium score regression

We performed stratified LD score regression (S-LDSC) (*33*) to evaluate global ancestry-associated DEGs for their enrichment for heritability of complex traits, mainly focused on 17 brain and five immune-related traits as a positive control. We downloaded GWAS summary statistics of each trait from the sources listed in **Data S14**. Following recommendations from the LDSC resource website (https://alkesgroup.broadinstitute.org/LDSCORE/), we ran S-LDSC for each list of candidate genes. We used the baseline LD model (version 2.2) that included 97 annotations to control for the LD between variants with other functional annotations in the genome. To remove other potential confounding factors in our analysis,we also included one annotation of all protein-coding genes.

To capture the regulatory regions of each gene, we defined gene intervals as a region spanning 500 kb upstream of the gene's start position and 50 kb downstream of its end position. We used HapMap Project Phase 3 SNPs as regression SNPs and 1000 Genomes SNPs of European ancestry samples as reference SNPs. We downloaded all SNPs from the LDSC resource website.

We ran S-LDSC for all ancestry-associated DEGs and conducted separate runs for DEGs of protein-coding and non-coding genes. For cell type-specific enrichment, we used glia subpopulation specificity markers generated in *"Single-cell specificity and cell-type enrichment analysis"*.

Differential methylation and environmental control of genetic ancestry

Variably methylated region analysis

To identify environmental-driven VMRs, we used only our admixed Black American neurotypical individuals (caudate [n=89], DLPFC [n=69], and hippocampus [n=69]). We considered approximately 24 million CpGs that had sequencing coverage of > 5 reads in > 80% samples of each brain region. We also excluded CpGs within ENCODE "blacklist" regions from analysis. We selected the top one million variable CpGs to compute principal components (PC) based on smoothed DNAm levels while removing the variation due to global ancestry of our primary variable of interest. Specifically, we regressed out global ancestry from each variable CpG and the residual DNAm was used for PC analysis. To capture CpGs whose variation of DNAm level was potentially driven by unknown environmental factors, we computed standard deviation for residualized DNAm levels of each CpG after regressing out top five PCs to remove variations due to batch effects and biological factors. We then selected the top 1% variable CpGs for calling VMRs for each brain region using the *regionFinder3* function of bsseq and VMRs, retaining VMRs with more than five CpGs for further analysis. We estimated the DNAm level of each VMR by the total number of reads supporting methylated cytosine divided by the total number of reads supporting either methylated or unmethylated cytosine in the region.

Differentially methylated region analysis

For differentially methylated region analysis, we applied a linear model on VMRs (see **Variably methylated region analysis**) as a function of: 1) global genetic ancestry, 2) local genetic ancestry, 3) sex, 4) age, and 5) top five principal components of DNAm derived from top one million variable CpGs. We corrected both statistical tests for multiple testing using Benjamini-Hochberg.

Functional enrichment analysis

We associate biological functions to global ancestry-associated DMRs using rGREAT (version 2.0.2) (*101*), an R/bioconductor package. Specifically, we selected significant DMRs (FDR < 0.05) and converted them into genomic ranges format with the plyranges (version 1.18.0) (*102*), an R/Bioconductor package. Following this conversion and filtering, we applied the *great* function from rGREAT with the MSigDB Canonical Pathway C5 (*90*) gene ontology database with background set to human genome (hg18) autosomal chromosomes. We extracted enrichment results using the *getEnrichmentTable* function and plotted region-gene associations with *plotRegionGeneAssociation* function from the rGREAT package.

Evaluating environmental impact of global ancestry-associated DEGs

To evaluate the impact of unknown environmental factors on global ancestry-associated DEGs, we first annotated the VMRs using *annotate_regions* and basic genes hg38 annotation from the R/Bioconduction package annotatr (version 1.24.0) (*103*) after converting to genomic ranges with plyranges. Following annotation, we estimated P_{ST} [(18)]. P_{ST} is essentially the partial coefficient of determination. As such, we estimated the P_{ST} statistic for each gene with **Equation 11**. We calculated the P_{ST} statistics for ancestry before and after including the residualized VMRs annotated to an ancestry-associated DEG. The residual was derived from raw DNAm levels of each VMR by regressing out known biological factors (local ancestry, age, sex), as well as potential batch effects and other unknown biological factors captured by the top five principal components of DNAm levels.

Following this, we calculated ΔP_{ST} to extract the fraction of change associated with the environment (**Equation 12**).

$$R^2_{partial} = \frac{SSE(reduced) - SSE(full)}{SSE(reduced)}$$

Equation 11

$$\Delta P_{ST} = \frac{Pst - Pst_{VMR}}{Pst}$$

Equation 12

Graphics

We used R to generate all plots. We generated UpSet plots using ComplexHeatmap (version 2.10.0; (*104*)). To generate Circos plots, we used circlize (version 0.4.15; (*105*)). We generated enrichment heatmaps, gene term enrichment, error plots, box plots, distribution plots, and scatterplots using a combination of ggplot2 (version 3.3.6; (*106*)) and ggpubr (version 0.4.15; (*107*)). For pairwise comparison plots, we used corrplot (version 0.92; (*108*)). We generated meta plots using the mashr function *mash_plot_meta*. We generated venn diagrams with ggvenn.

Code availability

All code and Jupyter Notebooks are available through GitHub at https://github.com/LieberInstitute/aanri_phase1 with more detail (*55*).

Supplementary Figures

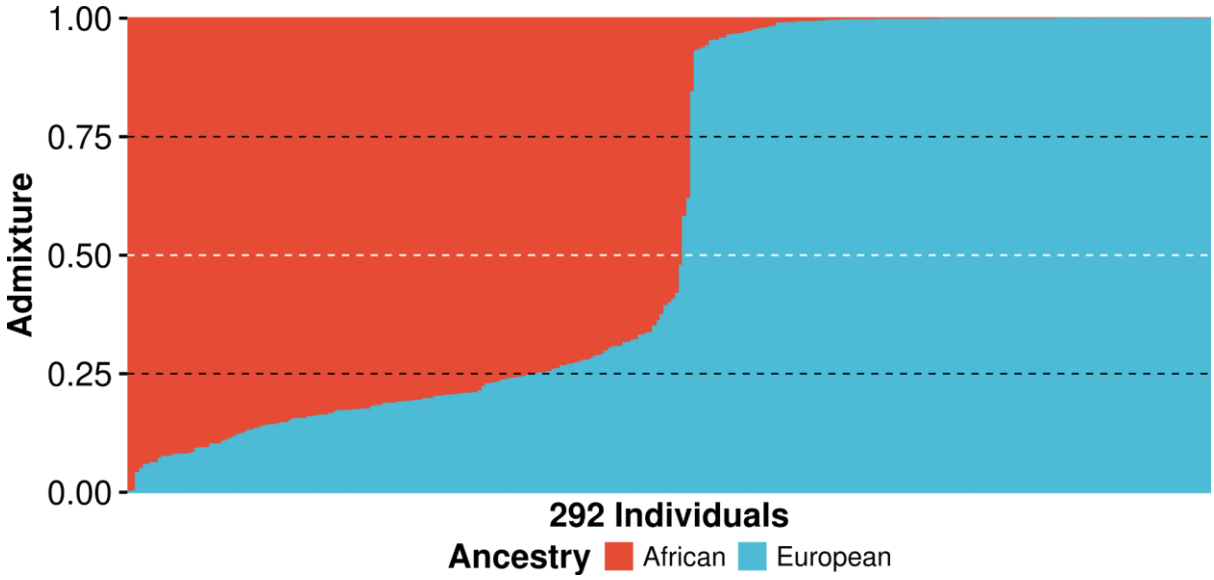

Fig. S1: Genetic ancestry estimates for Black and White American neurotypical individuals. Histogram showing estimated African and European ancestry for each unique neurotypical individual across the caudate, dentate gyrus, DLPFC, and hippocampus.

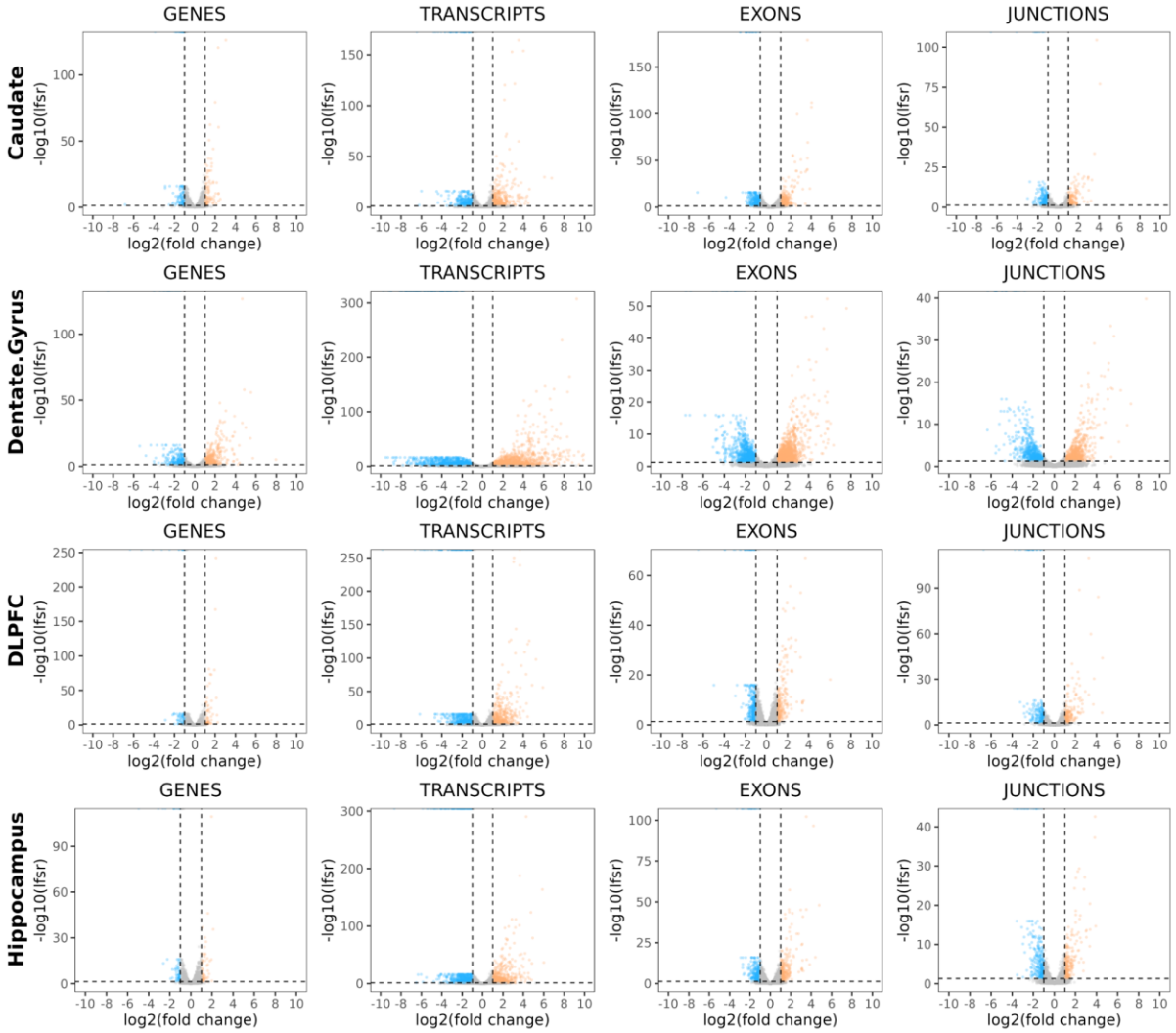

Fig. S2: Extensive global ancestry-associated differential expression across brain regions and features (i.e., gene, transcript, exon, and junction). Volcano plot of effect size ($\log_2$ of fold change) estimated from mash modeling and significance ($-\log_{10}$ of lfsr) with features associated with increased AA (blue) or EA (orange) proportions.

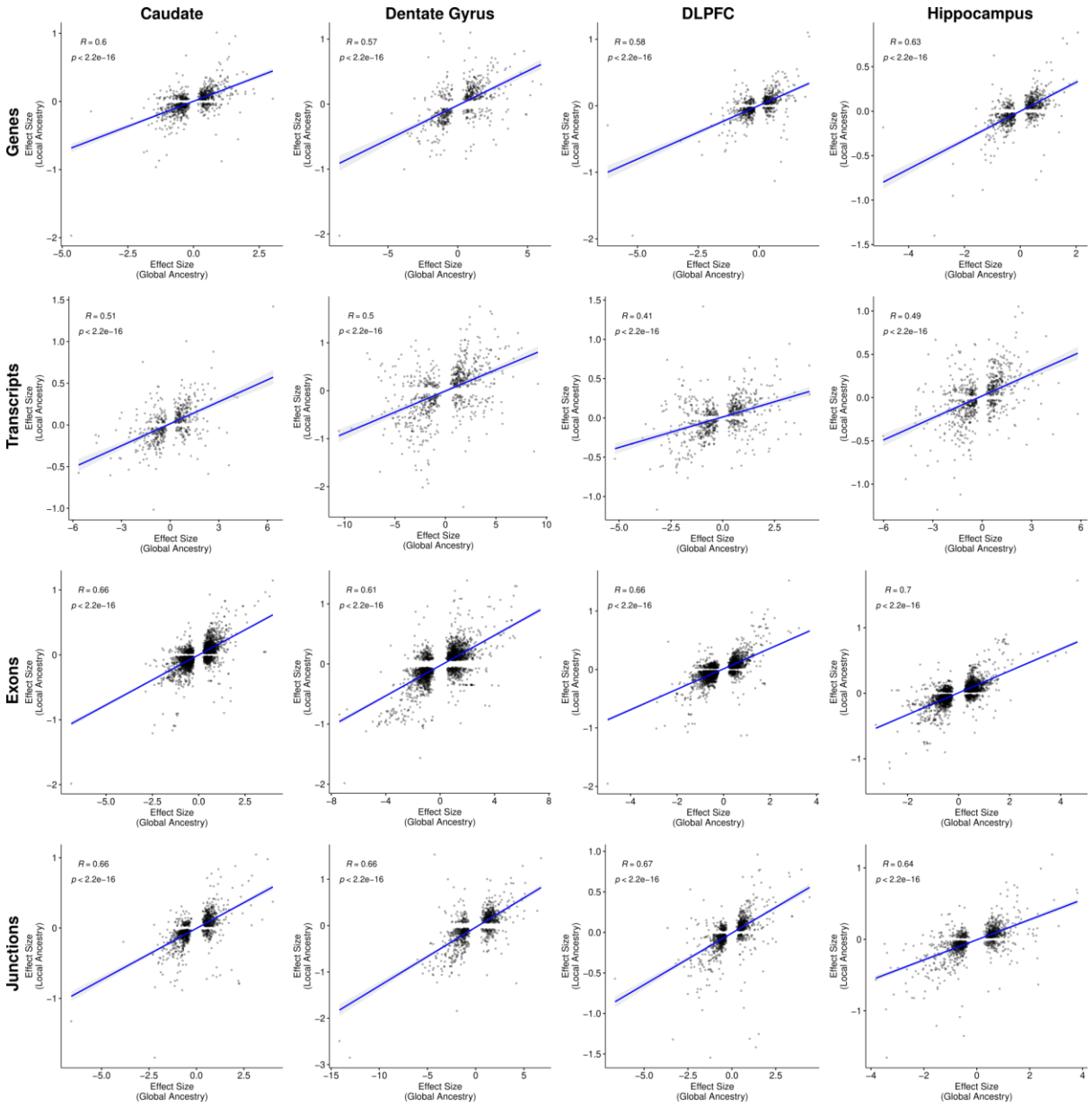

Fig. S3: Significant correlation between global and local ancestry shared features with local ancestry showing smaller effect sizes. Correlation (two-sided, Spearman) of local ancestry-associated DE features (i.e., gene, transcript, exon, and junction) effect sizes (y-axis) versus global ancestry-associated DE features effect sizes (x-axis) across brain regions. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

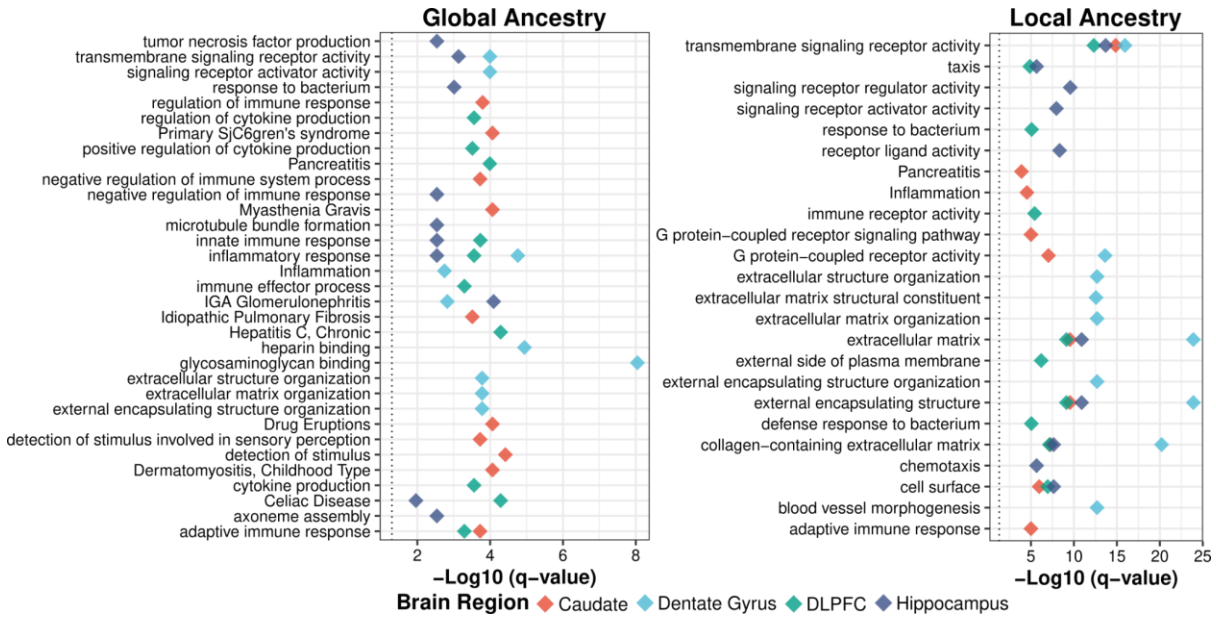

Fig. S4: Significant enrichment of immune-related pathways for ancestry-associated DEGs. GO enrichment (hypergeometric, q-value < 0.05) of all global (left) and local (right) ancestry-associated DEGs across brain regions, highlighting terms associated with immune response.

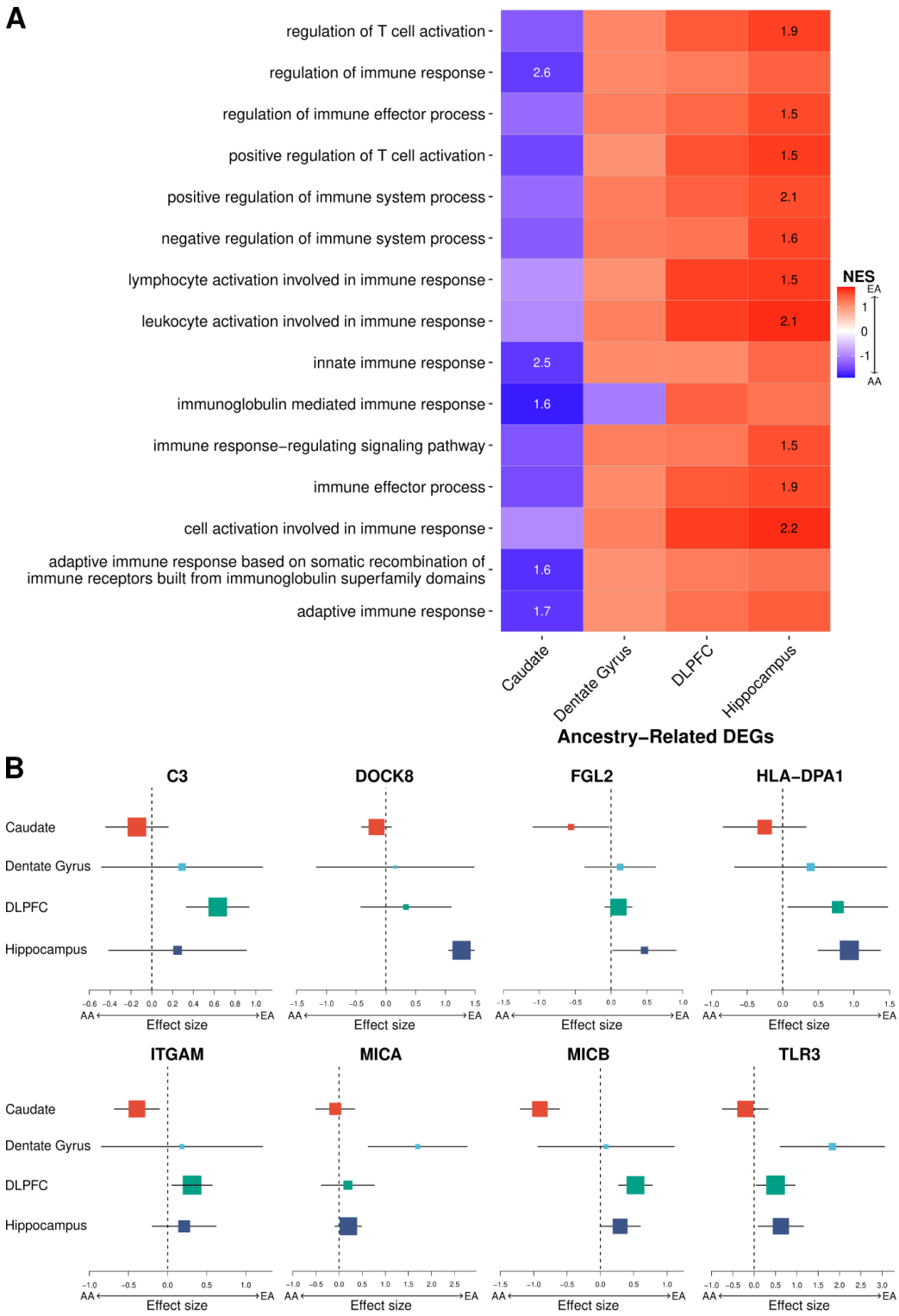

Fig. S5: Immune-related pathways show consistent direction of effect expression across brain regions. A. Heatmap showing direction of effect (increase AA proportion [blue] or increased EA proportion [red]) associated with immune-related GO terms across brain regions. Significant enrichments (GSEA, q-values < 0.05; -$\log_{10}$ transformed) annotated within tiles. **B.** Metaplot showing examples of immune-related genes associated with significantly enriched pathways (GSEA enrichment analysis) across brain regions.

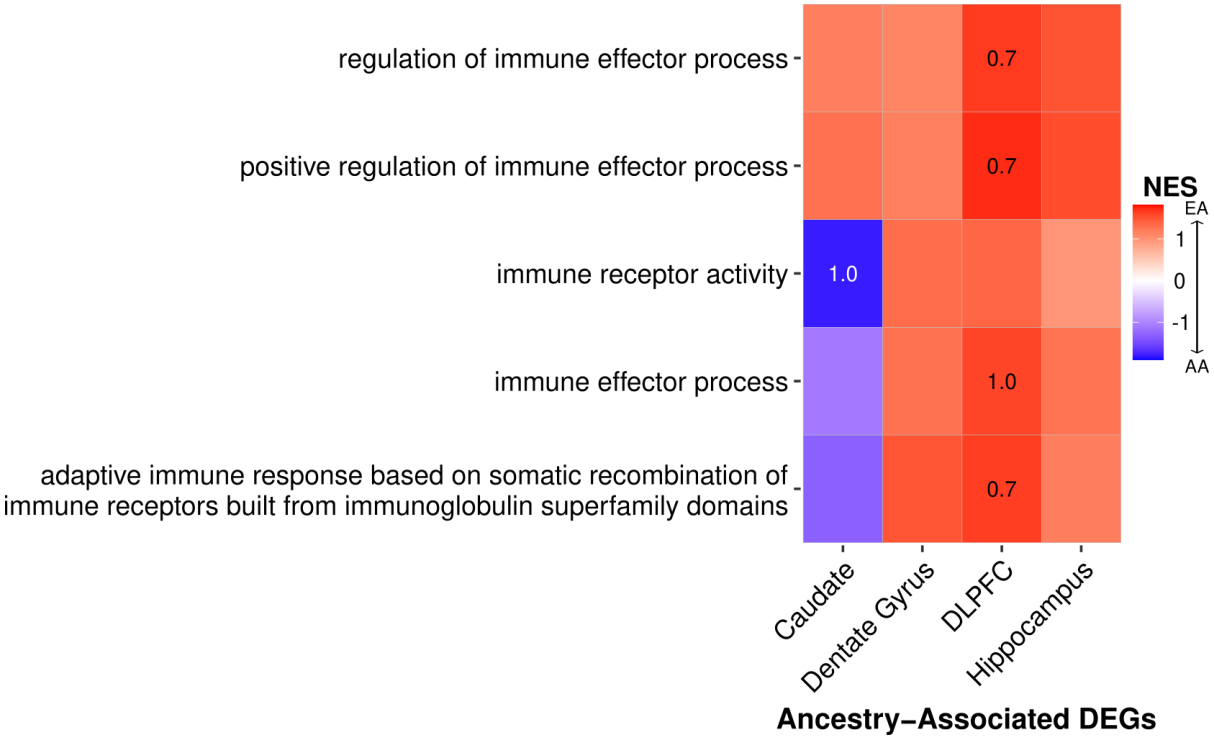

Fig. S6: Local ancestry expression show similar pattern of direction of effect in immune-related pathways across brain regions as global ancestry. Heatmap showing direction of effect (increase AA proportion [blue] or increased EA proportion [red]) associated with immune-related GO terms across brain regions. Enrichment trends (GSEA, q-values < 0.25; $-\log_{10}$ transformed) annotated within tiles.

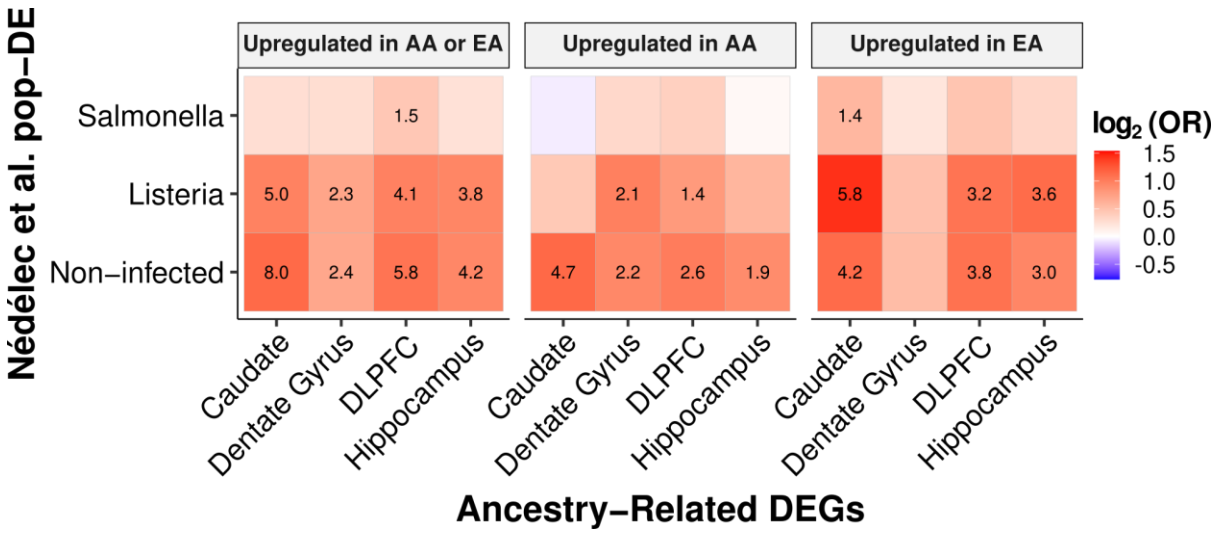

Fig. S7: Significant enrichment of ancestry-associated DEGs with population differences in immune responses. Heatmaps showing significant enrichment (red; two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) of ancestry-associated DEGs (adjusted p-value < 0.05) and population differences in primary macrophages (*18*) separated by infection status and direction of effect. Significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values $-\log_{10}$ transformed) annotated within tiles.

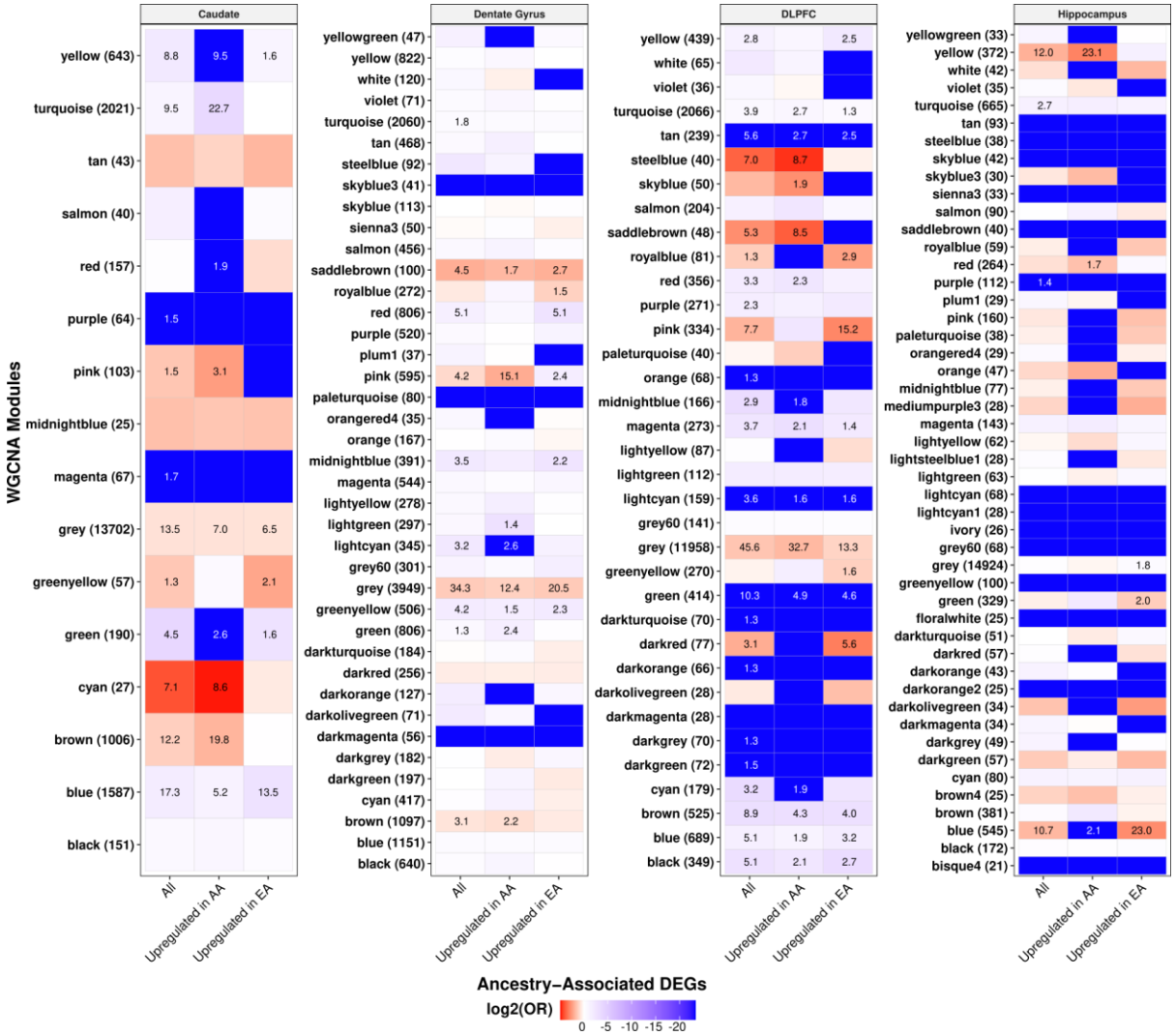

Fig. S8: Extensive enrichment of gene co-expression modules with ancestry-associated DEGs in admixed Black American individuals across brain regions. Heatmap of enrichment analysis (two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) showing significant enrichment (red) and depletion (blue) across WGCNA modules for ancestry-associated DEGs (adjusted p-value < 0.05) separated by direction of effect. Significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values -$\log_{10}$ transformed) annotated within tiles.

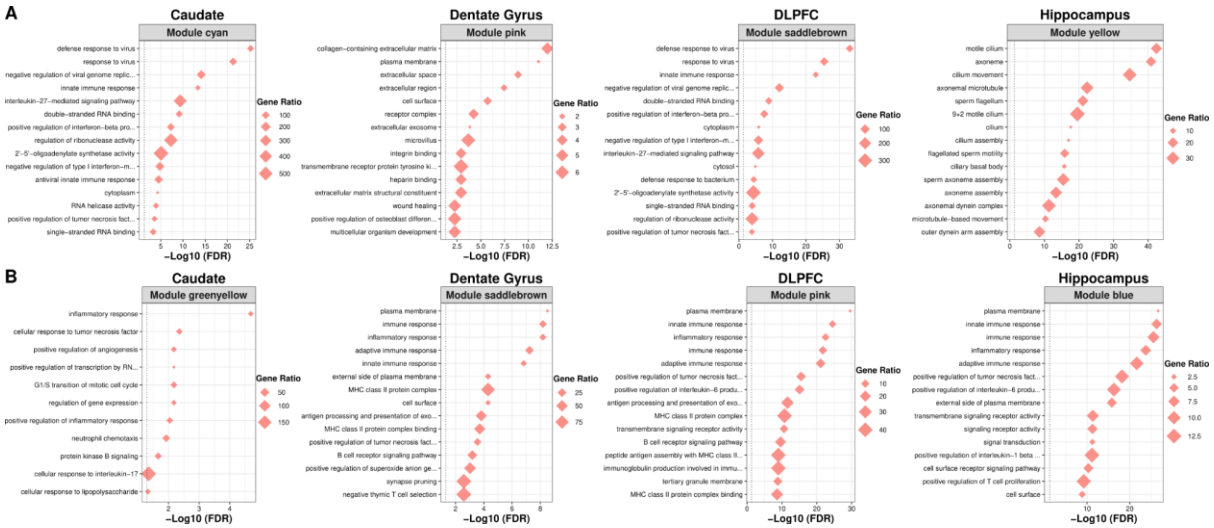

Fig. S9: Functional enrichment of gene co-expression network modules and ancestry-associated DEGs in admixed Black American across brain regions. Top 15 enriched GO terms for most significantly enriched WGCNA module for ancestry-associated DEGs that show **A.** upregulation with increasing AA proportion or **B.** upregulation with increasing EA proportion.

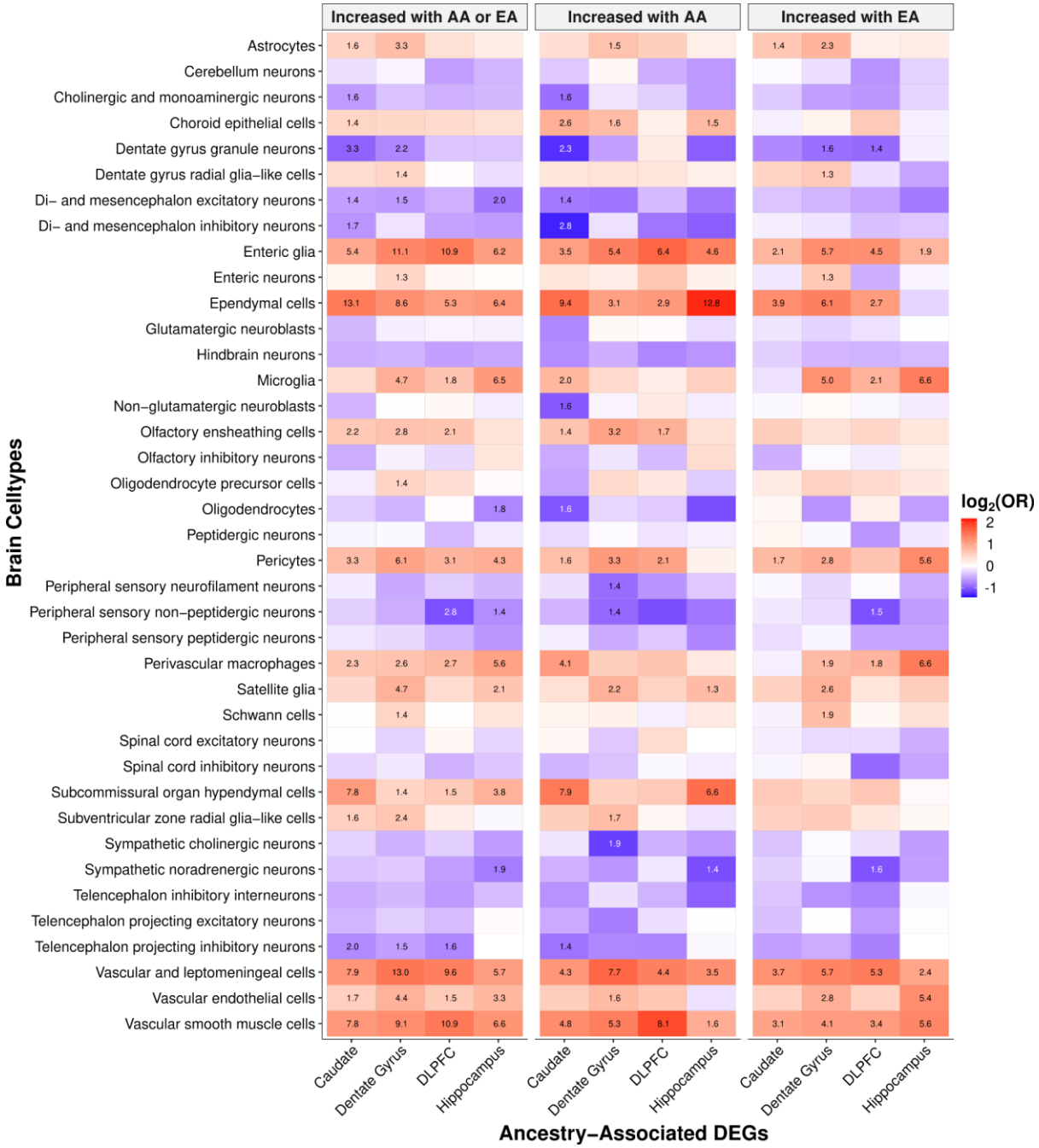

Fig. S10: Global ancestry-associated DEGs show significant enrichment of immune-related cell types (i.e., microglia and macrophages). Heatmap showing enrichment analysis (two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) of significantly enriched (red) or depleted (blue) across brain cell types (*24*) for ancestry-associated DEGs (adjusted p-value < 0.05) separated by direction of effect. Significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values -log10 transformed) annotated within tiles.

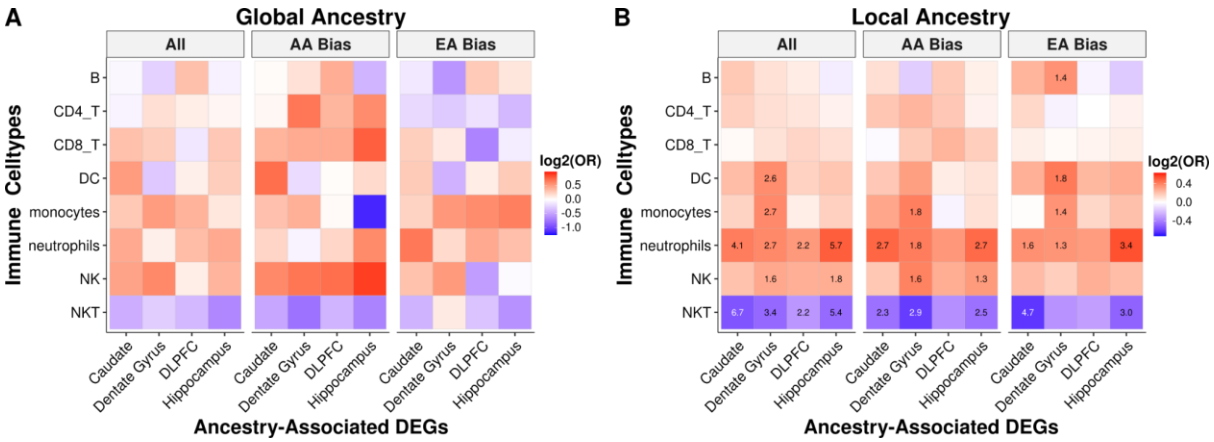

Fig. S11: Enrichment of non-brain immune cell types for local but not global ancestry DEGs. Heatmap showing enrichment analysis (two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) of significantly enriched (red) or depleted (blue) across peripheral blood mononuclear cells (PBMCs) cell types (*25*) for ancestry-associated DEGs (adjusted p-value < 0.05) separated by direction of effect. Significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values -log10 transformed) annotated within tiles.

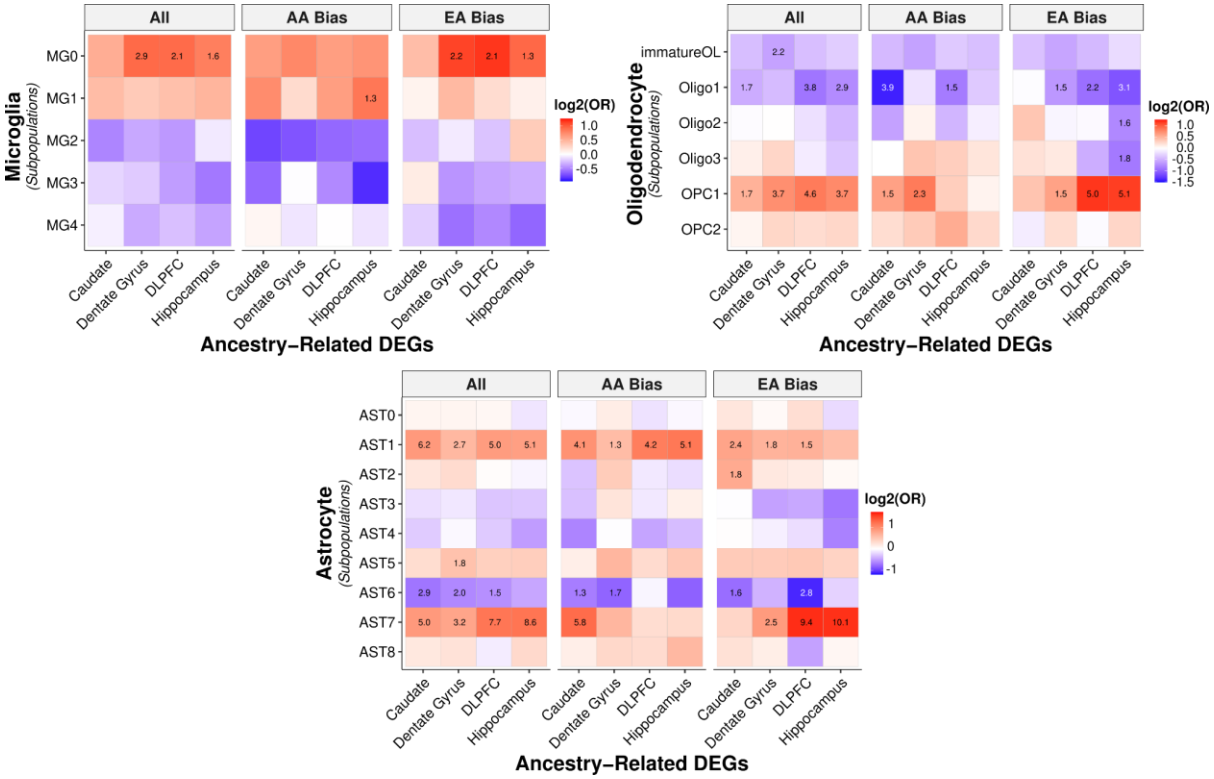

Fig. S12: Distinct, but not specific enrichment of global ancestry-associated DEGs for glial cells. Heatmap showing enrichment analysis (two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) of significantly enriched (red) or depleted (blue) across brain immune cell subtypes (*26*) for ancestry-associated DEGs (adjusted p-value < 0.05) separated by direction of effect. Significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values -log10 transformed) annotated within tiles.

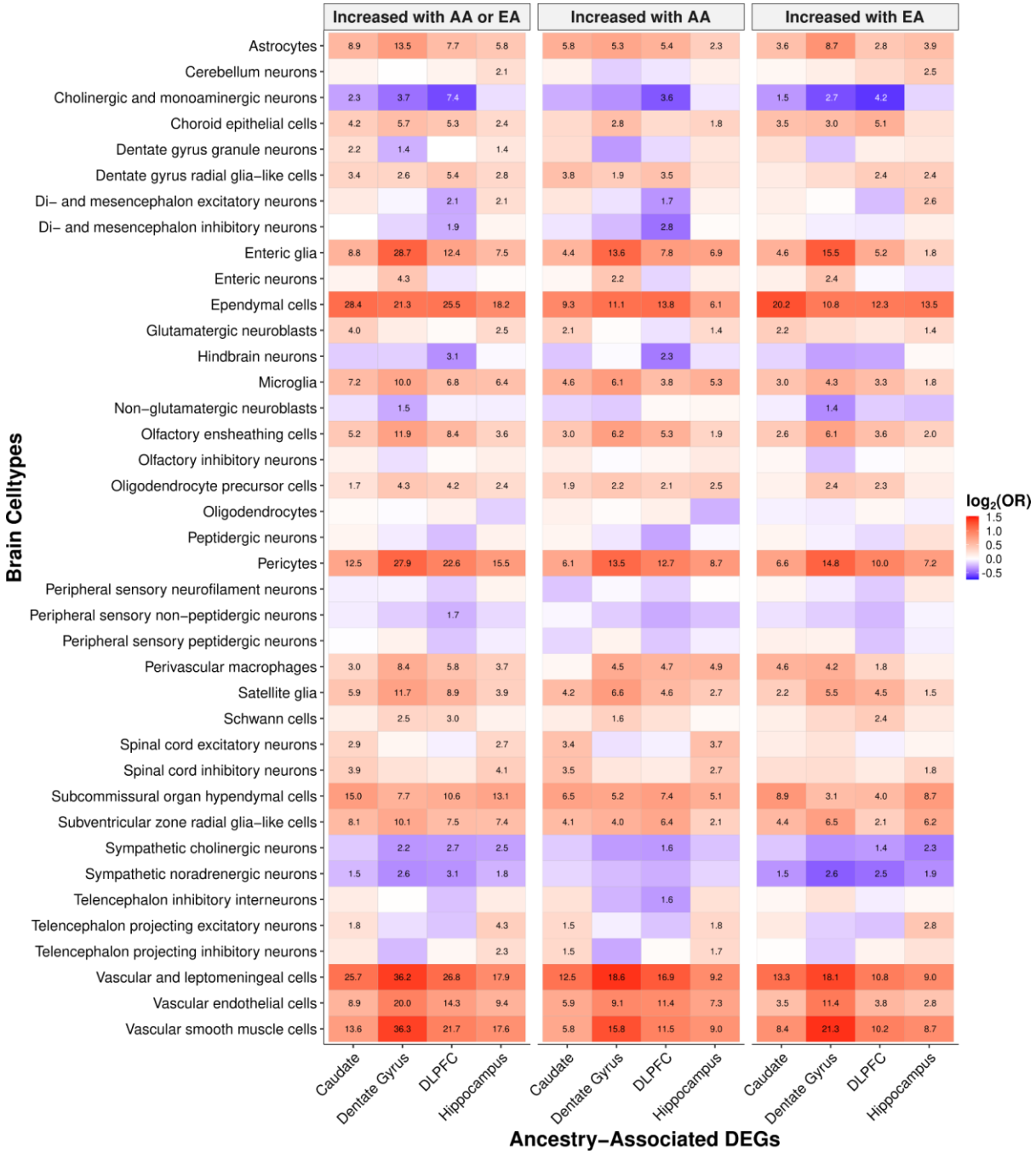

Fig. S13: Local ancestry-associated DEGs show significant enrichment of immune-related cell types (i.e., microglia and macrophages). Heatmap showing enrichment analysis (two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) of significantly enriched (red) or depleted (blue) across brain cell types (*24*) for ancestry-associated DEGs (adjusted p-value < 0.05) separated by direction of effect. Significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values -log10 transformed) annotated within tiles.

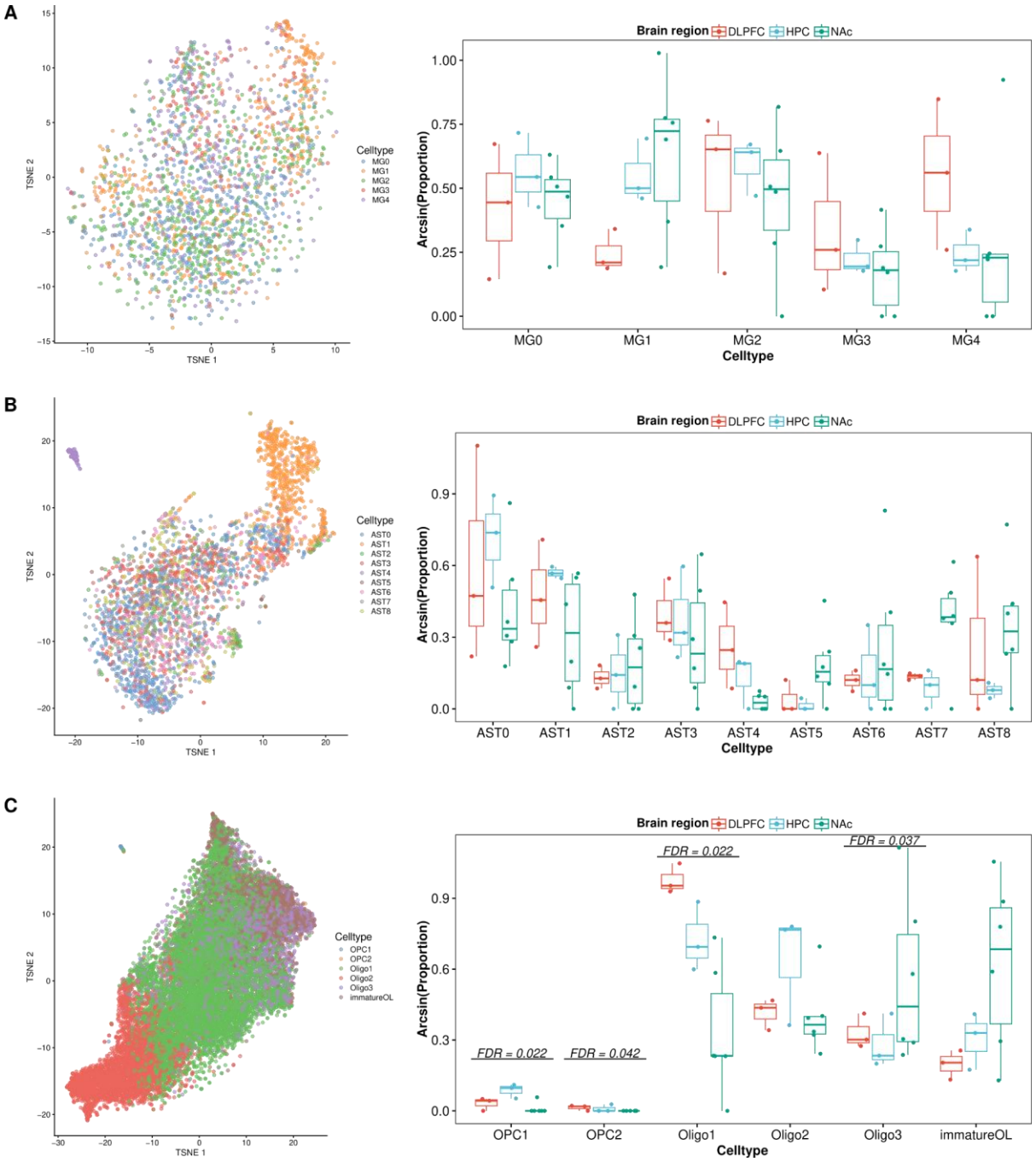

Fig. S14: The majority of glial cell composition is not significantly different across brain regions. t-SNE and cell proportion box plots (two-way, ANOVA) of single-cell data from the DLPFC (n=3), hippocampus (HPC; n=3), and nucleus accumbens (NAc; n=8) (*27*) after annotating for **A.** microglia subpopulations, **B.** astrocyte subpopulation, and **C.** oligodendrocyte lineage (*26*).Box plots show the median and first and third quartiles, and whiskers extend to 1.5× the interquartile range. Y-axis is arcsine transformed counts.

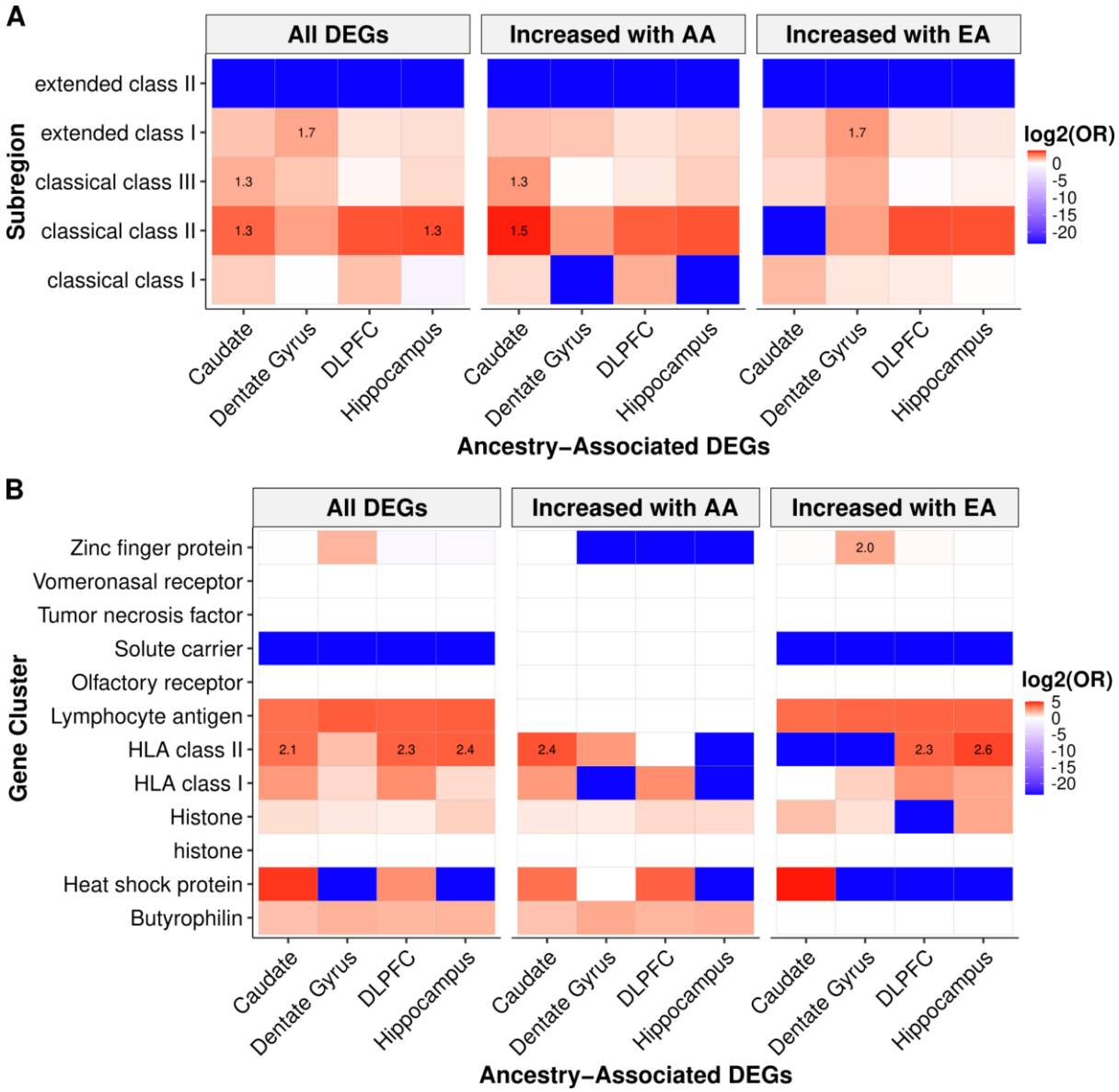

Fig. S15: Global ancestry-associated DEGs are significantly enriched for HLA class II genes. Heatmap showing enrichment analysis (two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) of significantly enriched (red) or depleted (blue) across extended MHC region for ancestry-associated DEGs (adjusted p-value < 0.05) separated by direction of effect. **A.** Subregions of the extended MHC region. **B.** Gene clusters of the extended MHC region. Significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values -log10 transformed) annotated within tiles.

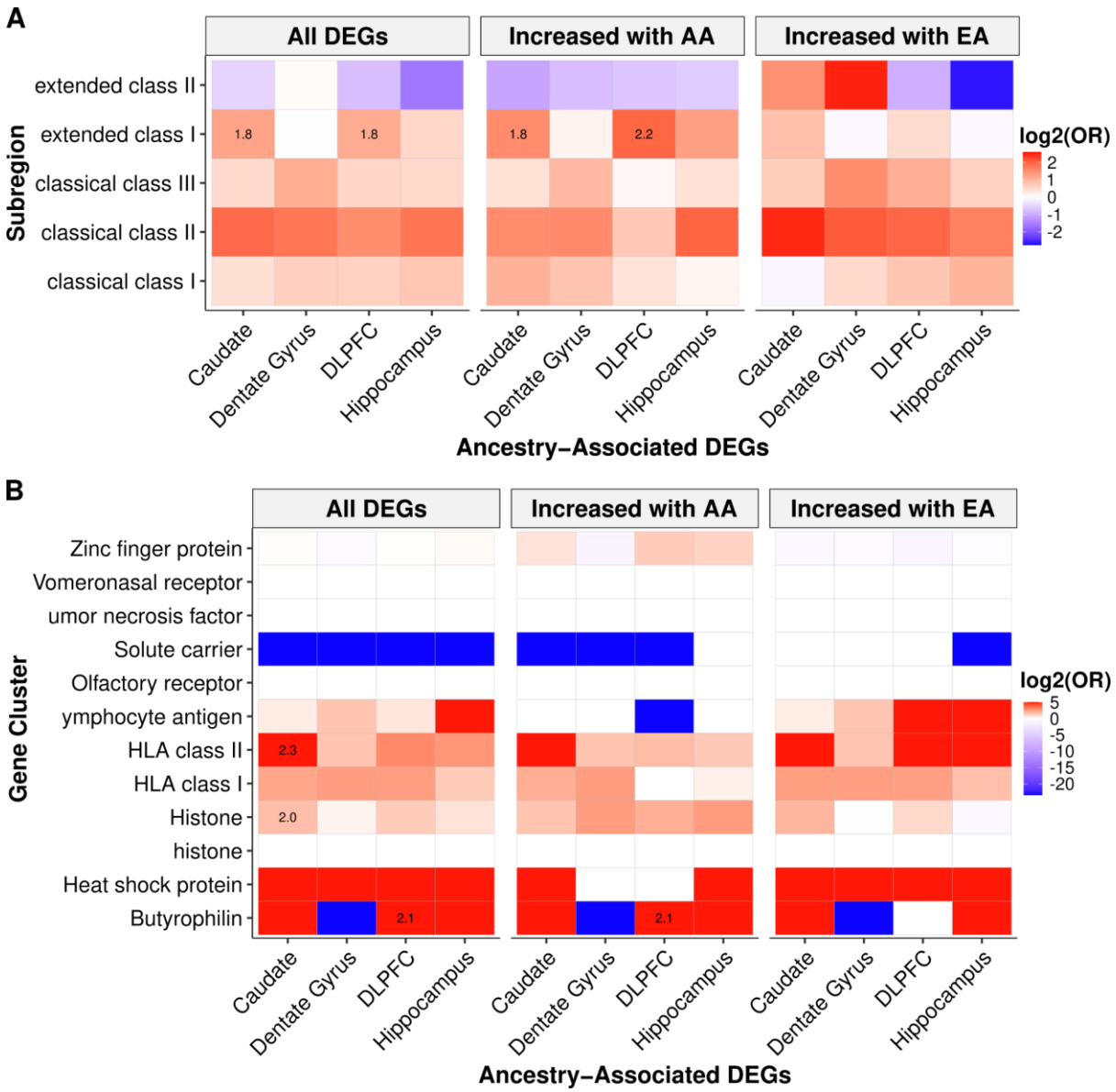

Fig. S16: Local ancestry-associated DEGs are significantly enriched for HLA class II genes. Heatmap showing enrichment analysis (two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) of significantly enriched (red) or depleted (blue) across extended MHC region for ancestry-associated DEGs (adjusted p-value < 0.05) separated by direction of effect. **A.** Subregions of the extended MHC region. **B.** Gene clusters of the extended MHC region. Significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values -log10 transformed) annotated within tiles.

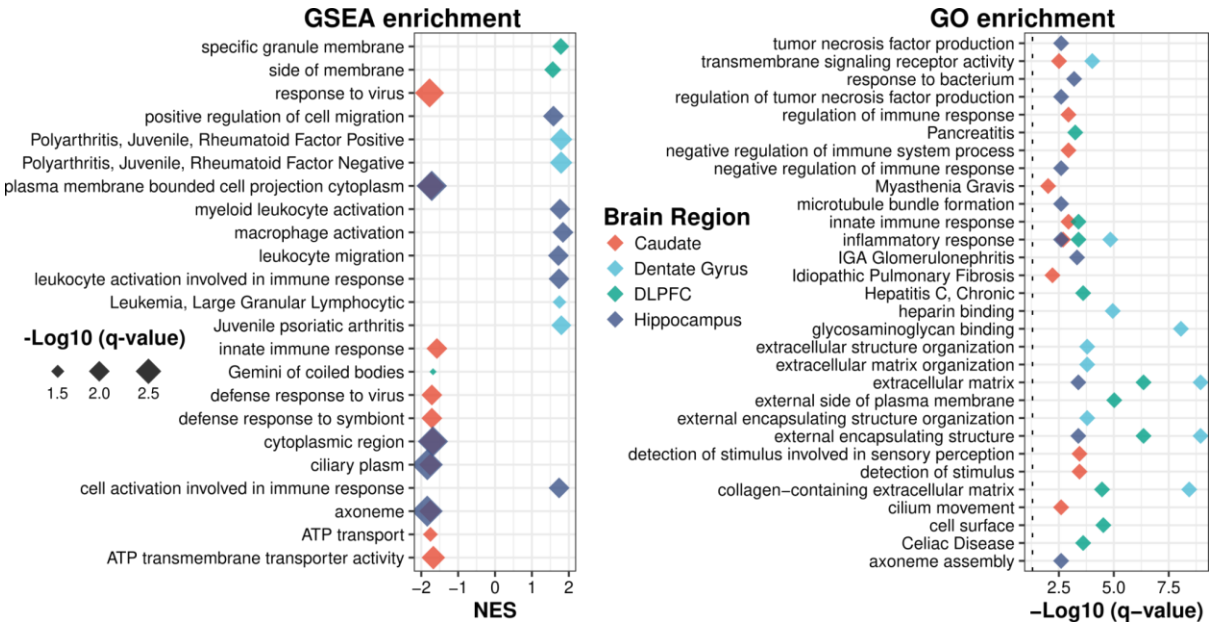

Fig. S17: HLA genes do not drive enrichment for immune-related pathways of global ancestry-associated DEGs. GSEA and GO enrichment of DEG without HLA genes across brain regions. GSEA analysis highlighting terms associated with increased AA (African ancestry) or EA (European ancestry) proportions. GO enrichment including all DEGs.

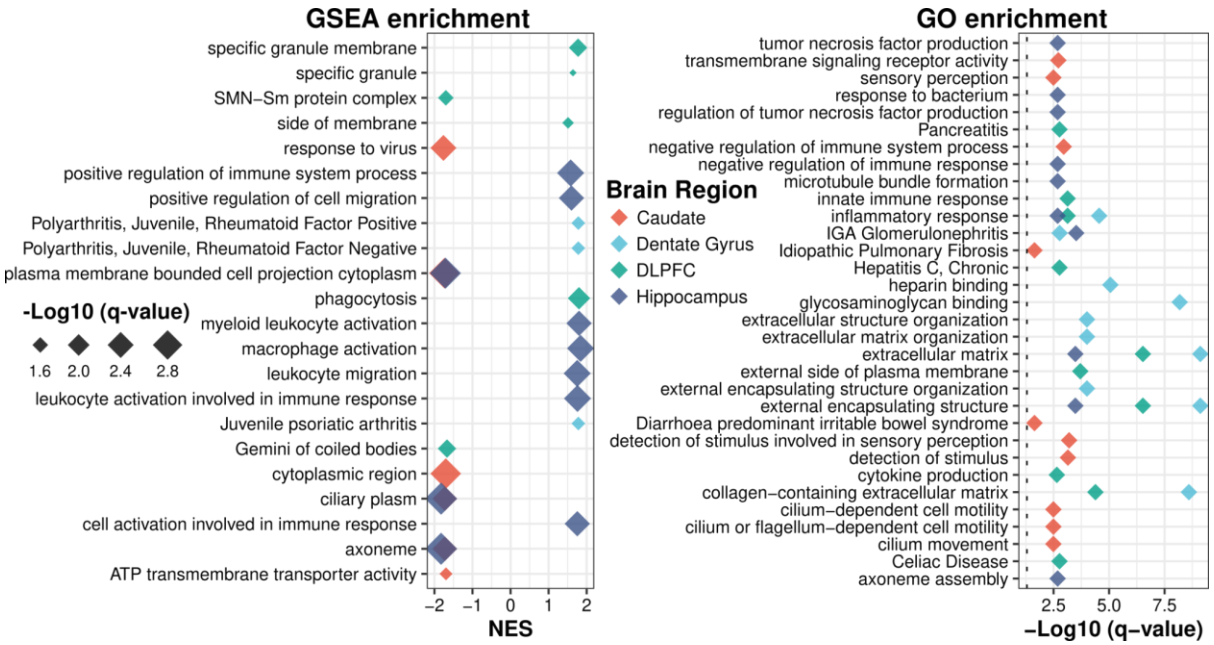

Fig. S18: MHC region does not drive enrichment for immune-related pathways of global ancestry-associated DEGs. GSEA and GO enrichment of DEG without MHC region across brain regions. GSEA analysis highlighting terms associated with increased AA (African ancestry) or EA (European ancestry) proportions. GO enrichment including all DEGs.

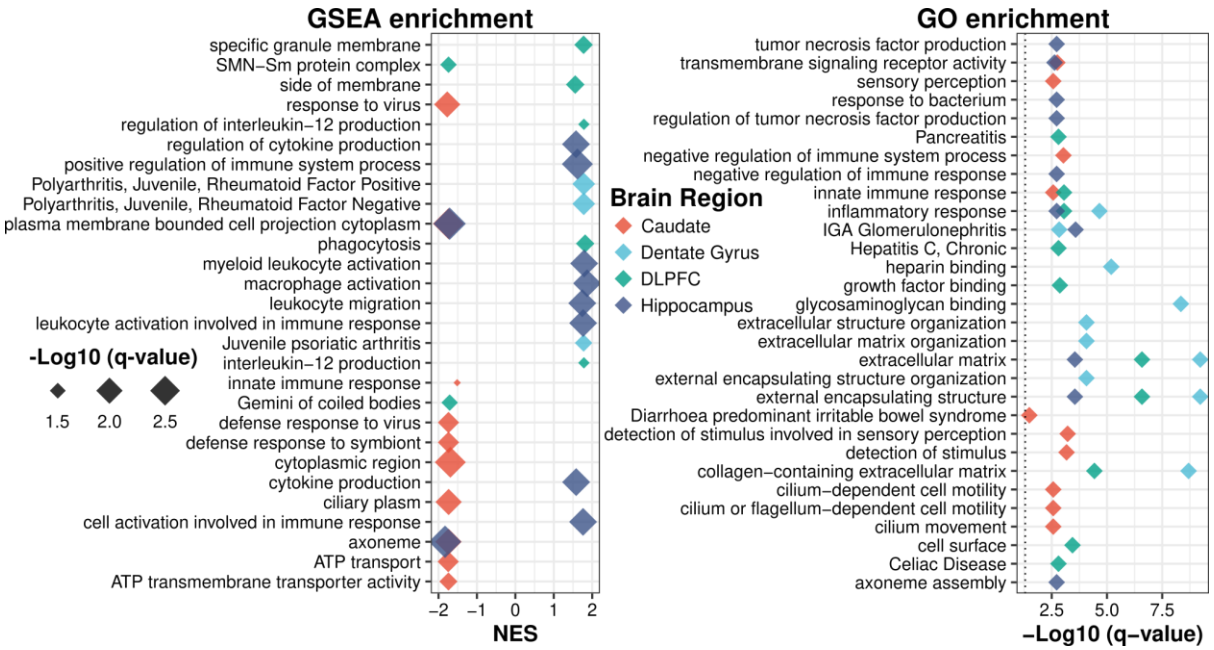

Fig. S19: Extended MHC region does not drive enrichment for immune-related pathways of global ancestry-associated DEGs. GSEA and GO enrichment of DEG without extended MHC region across brain regions. GSEA analysis highlighting terms associated with increased AA (African ancestry) or EA (European ancestry) proportions. GO enrichment including all DEGs.

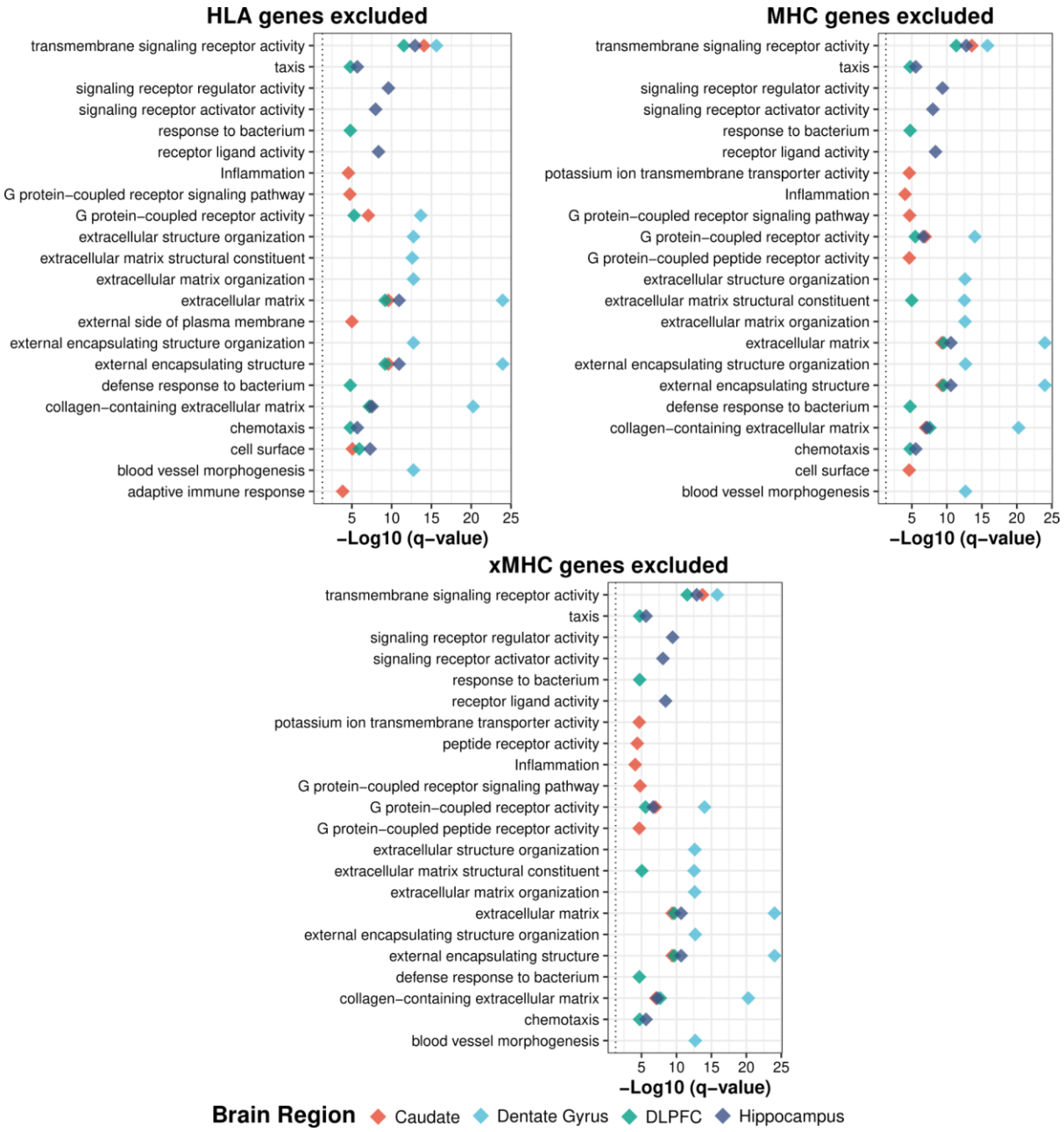

Fig. S20: Similar to global ancestry, extended MHC region does not drive enrichment for immune-related pathways of local ancestry-associated DEGs. GO enrichment of DEG without extended MHC region across brain regions. GO enrichment including all DEGs.

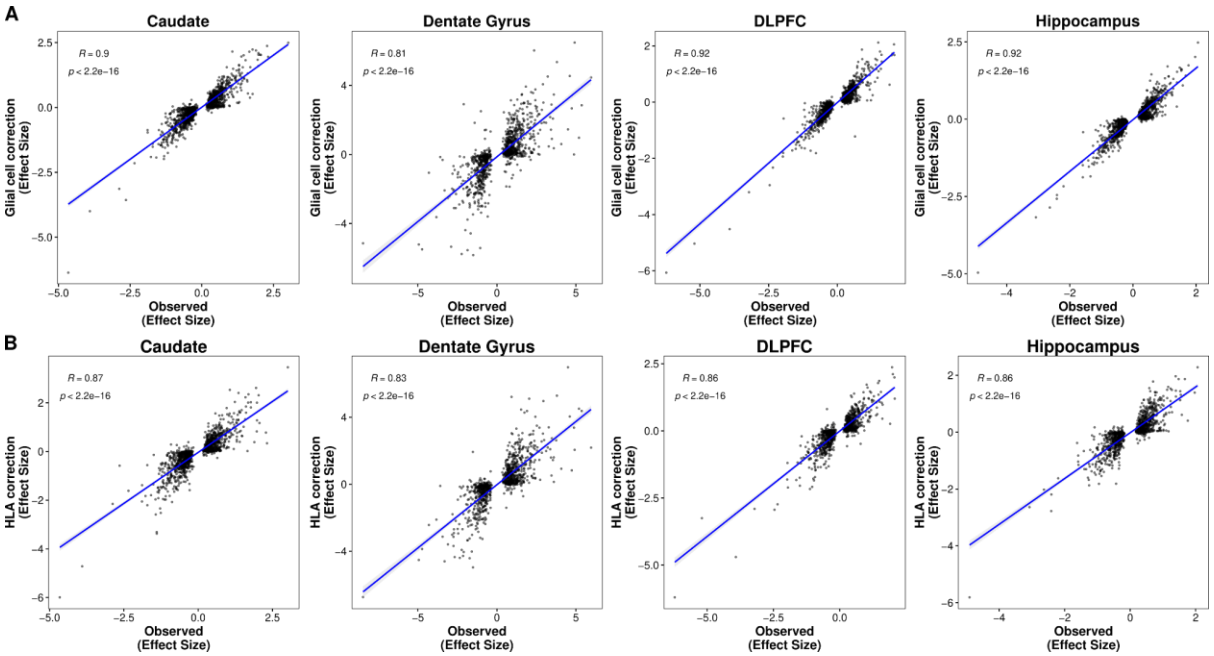

Fig. S21: Immune variation contributes only minimally to transcriptional changes of ancestry-associated DEGs. Scatter plot of global ancestry-associated DEGs comparing effect sizes from general model (x-axis) and model with covariates (y-axis) associated with **A.** HLA variation or **B.** glial cell proportion. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

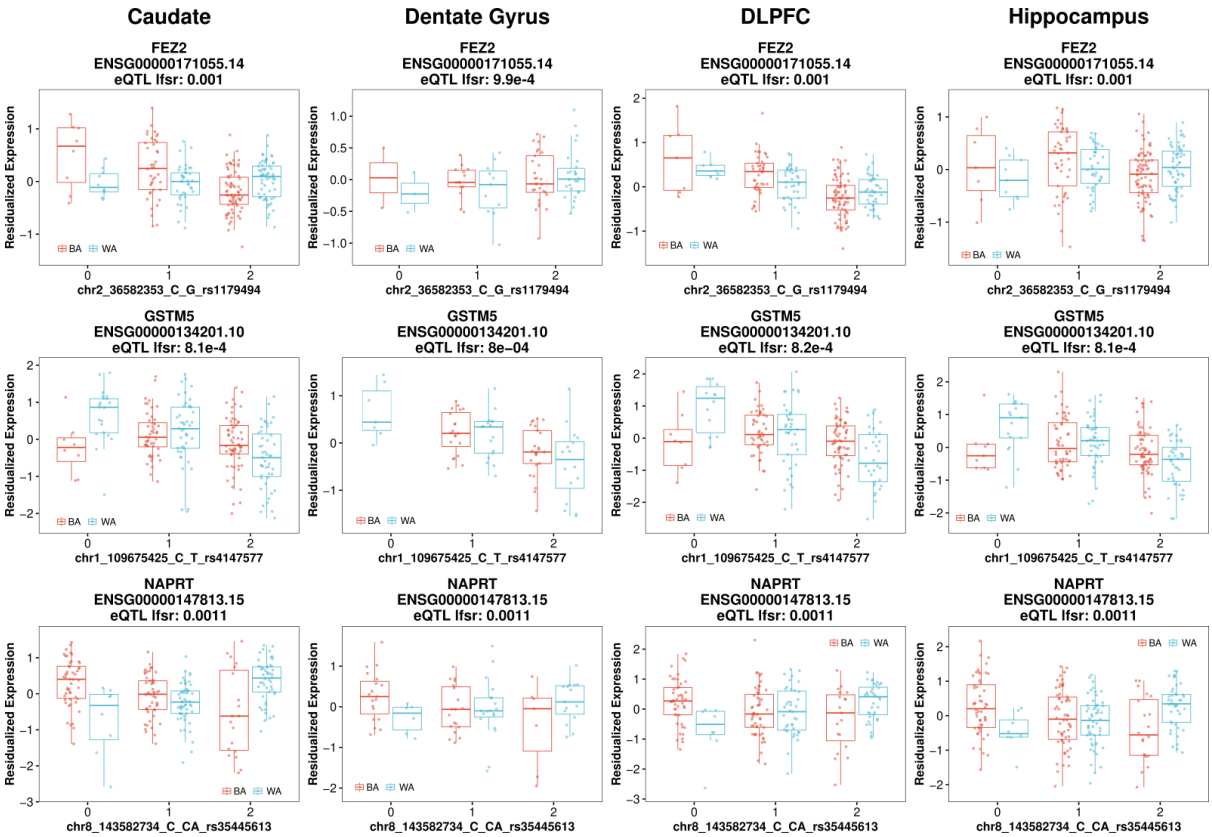

Fig. S22: Ancestry-dependent eQTL examples showing shared direction of effect across brain regions. Box plot of the most significant variant per eGene showing ancestry-dependent eQTL using combined (Black [red] and White [blue] Americans). Box plots show the median and first and third quartiles, and whiskers extend to 1.5× the interquartile range.

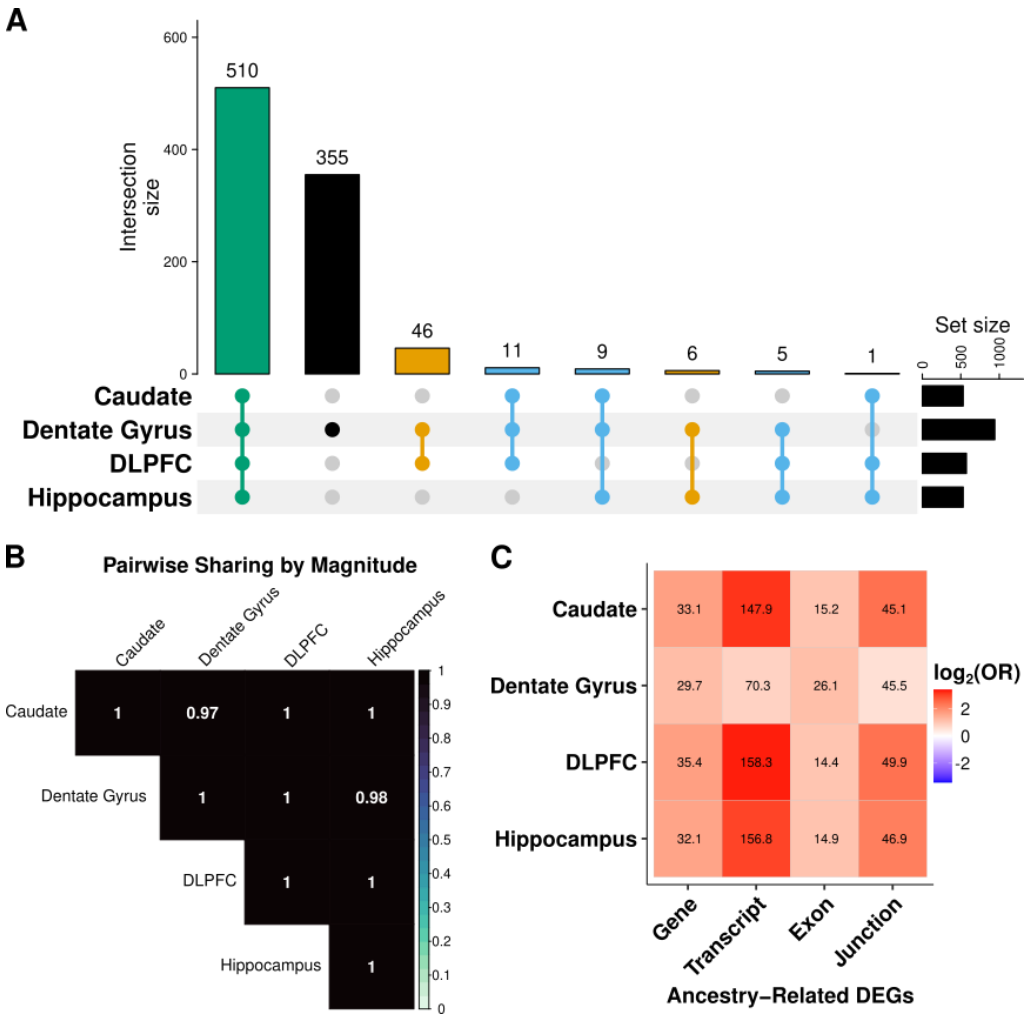

Fig. S23: **Ancestry-dependent eQTL shared across brain regions and enriched for main effect eQTL. A.** UpSet plot showing sharing of eGenes across brain regions. **B.** Heatmap showing significant sign matched. **C.** Enrichment heatmap of ancestry-dependent eGenes with ancestry-associated DEGs; significant enrichments (two-sided, Fisher's exact test with FDR corrected p-values $-\log_{10}$ transformed) annotated within tiles.

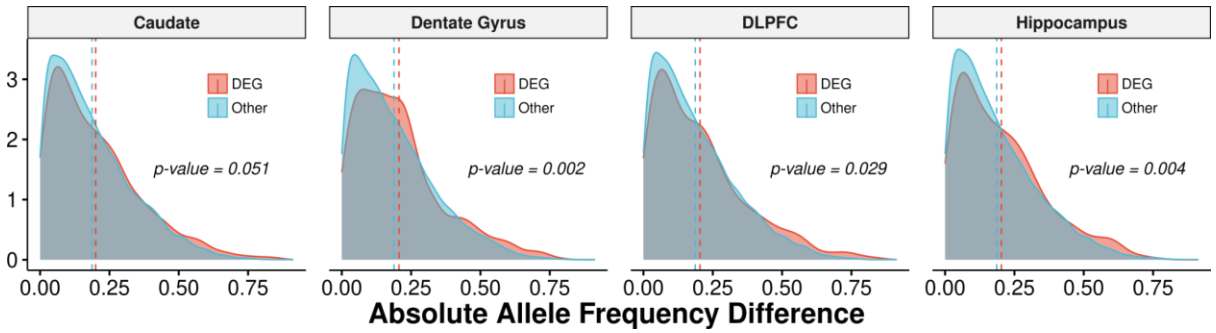

Fig. S24: Significant increase of absolute allele frequency difference for ancestry-associated DEGs compared with non-DEGs. Density plot showing significant increase in absolute allele frequency differences (AFD; one-sided, Mann-Whitney U, p-value < 0.05) for global ancestry-associated DEGs (red) compared with non-DEGs (blue) across brain regions. A dashed line marks the mean absolute AFD. Absolute AFD calculated from the most significant SNP per gene.

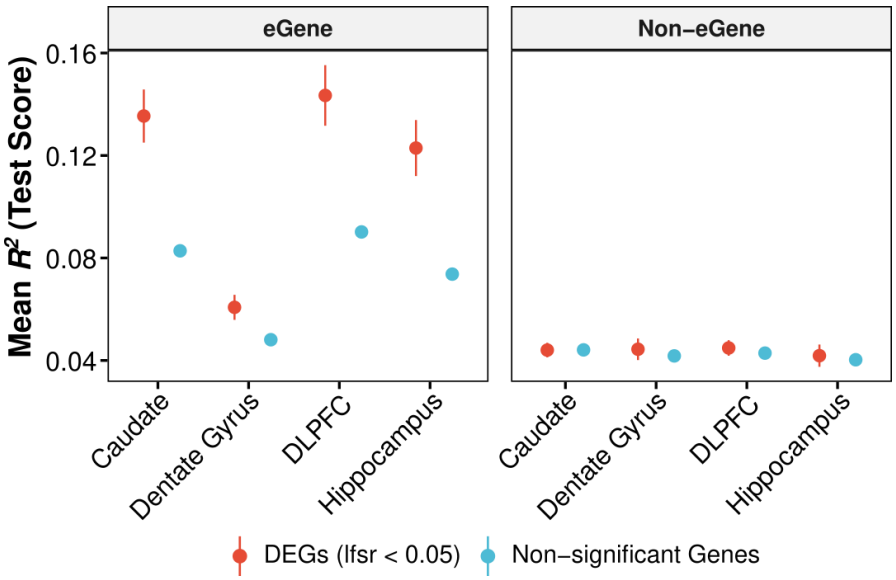

Fig. S25: Summary of test R2 for elastic net model separated by genes with an eQTL and those without. Error bars correspond to 95% confidence intervals. eGene are unique genes associated with an eQTL. DEGs are global ancestry-associated differential expressed genes.

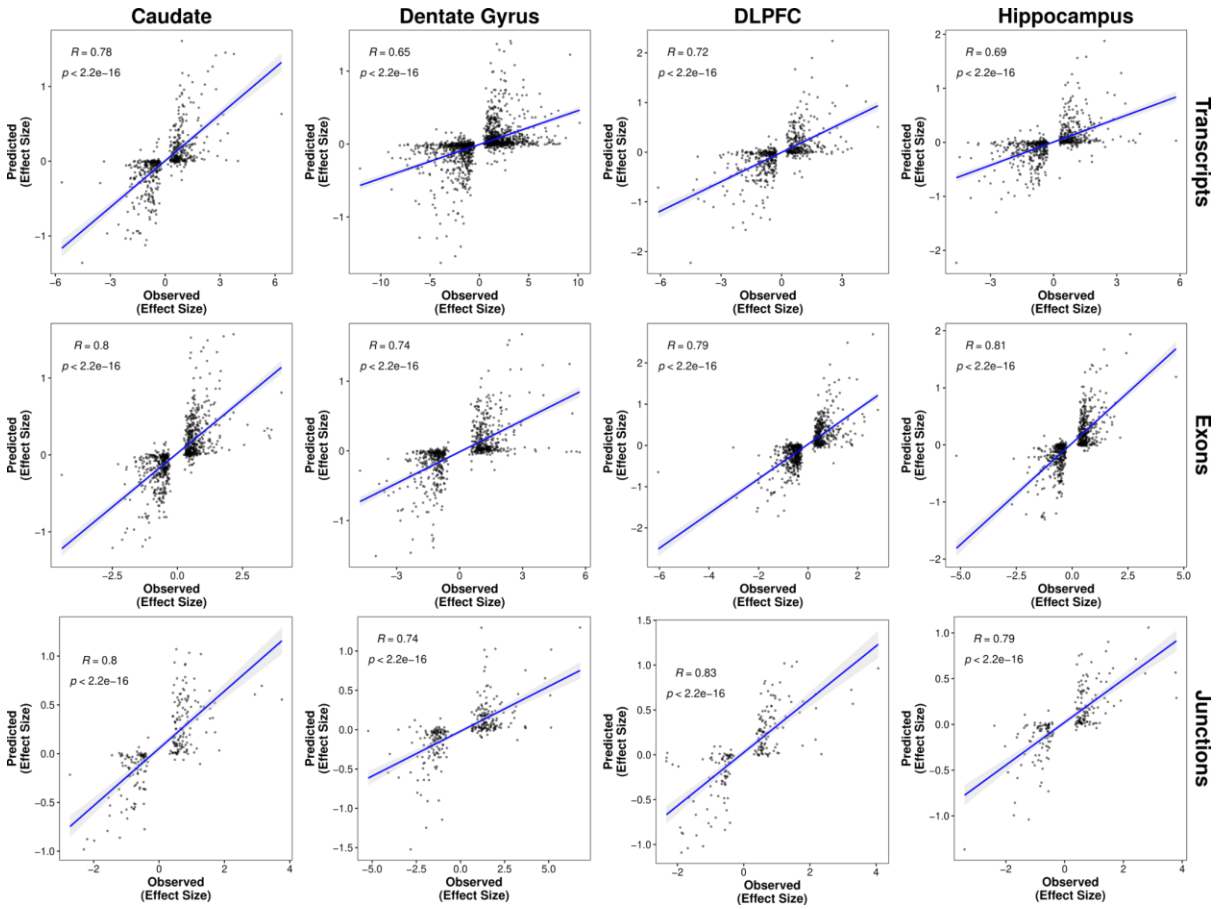

Fig. S26: Elastic net model captures more genetic contribution of genetic ancestry-associated expression changes in the brain on the isoform level. Correlation (two-sided, Spearman) of elastic net predicted (y-axis) versus observed (x-axis) global ancestry-associated differences in expression among ancestry-associated DE features (i.e., transcript, exon, and junction) with an eQTL across brain regions. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

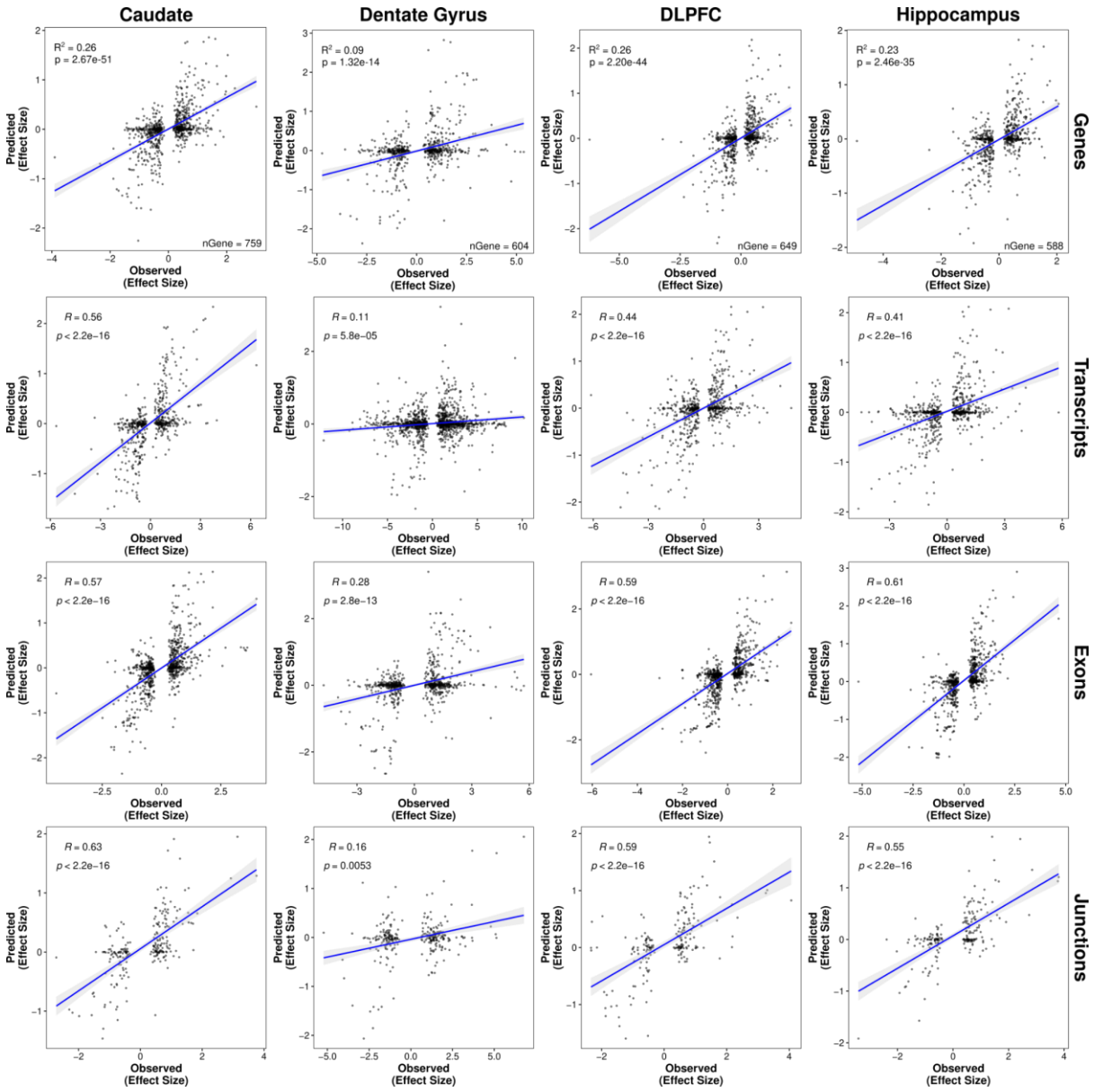

Fig. S27: The most significant eQTL explains roughly 20% of genetic ancestry expression differences in the brain. Correlation (two-sided, Spearman) of cis-predicted (y-axis) versus observed (x-axis) global ancestry-associated differences in expression among ancestry-associated DE features (i.e., gene, transcript, exon, and junction) with an eQTL across brain regions. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

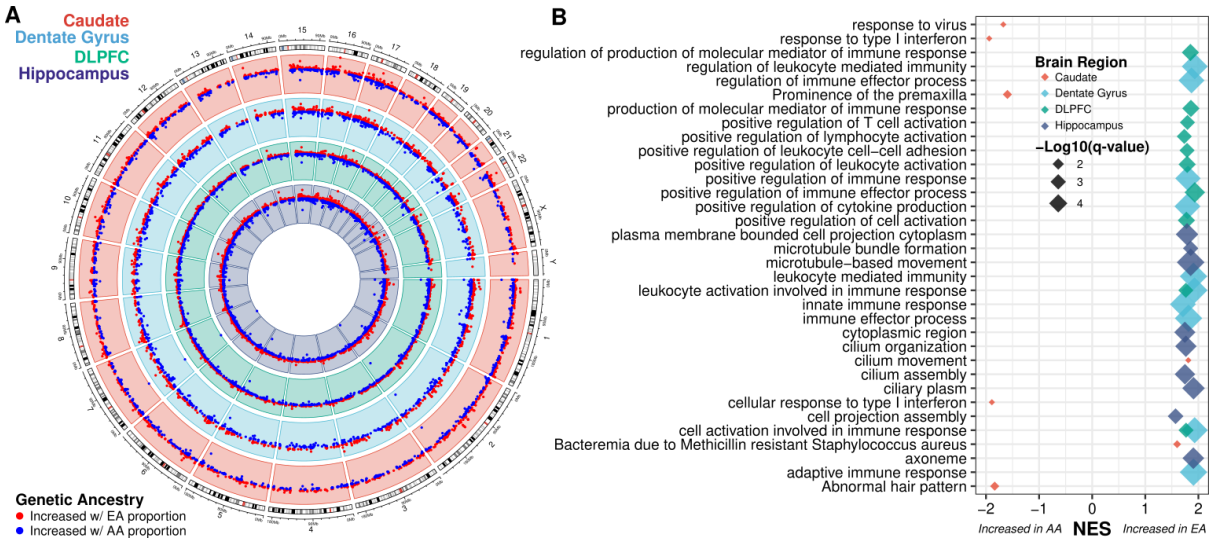

Fig. S28: Extensive ancestry-associated expression changes across the brain highlighting impact of environment. A. Circos plot showing ancestry DEGs from binary internal replication analysis across the caudate (red), dentate gyrus (blue), DLPFC (green), and hippocampus (purple). **B.** Gene set enrichment analysis of differential expression analysis across brain regions, highlighting terms associated with increased AA (African ancestry) or EA (European ancestry) proportions.

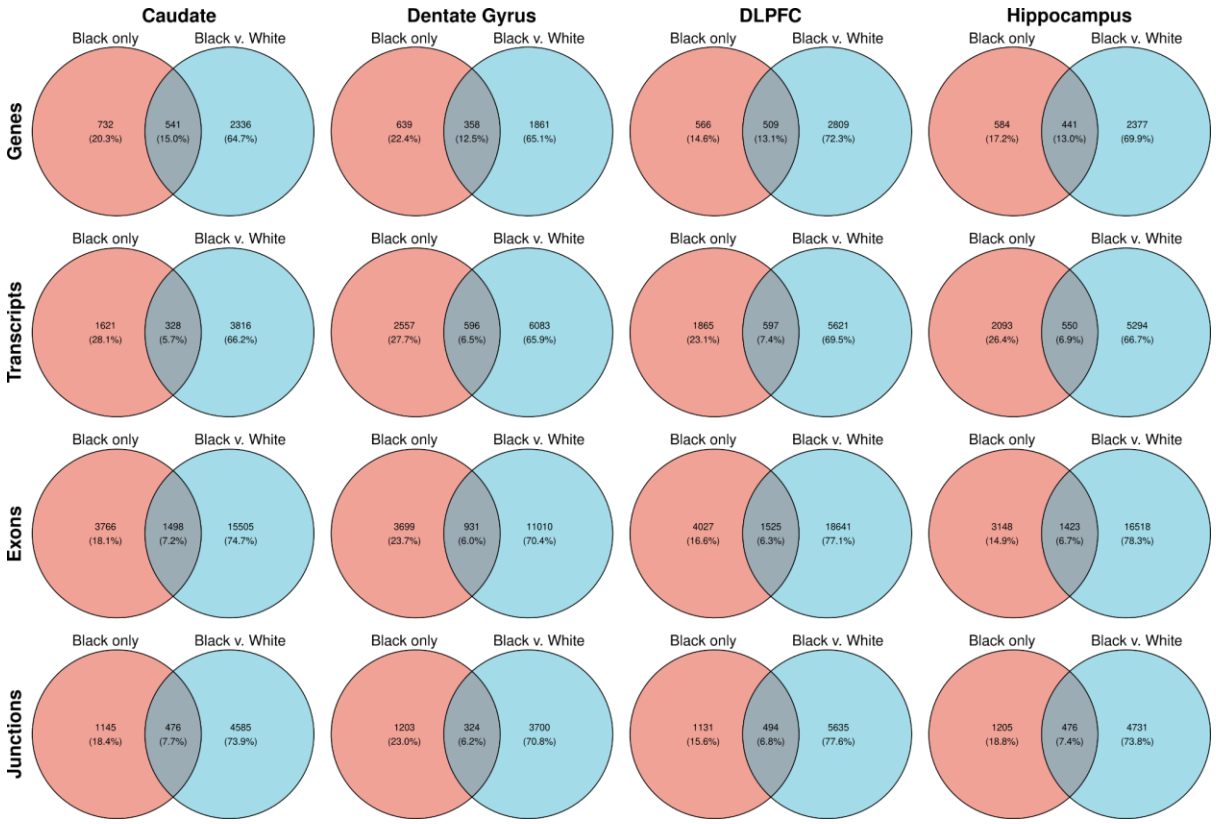

Fig. S29: Black versus White American binary analysis potentially confounded by environmental factors. Venn diagram showing the overlap of global ancestry-associated DE features (i.e., gene, transcript, exon, and junction) between within Black and other (Black v. White Americans) DE analysis across brain regions.

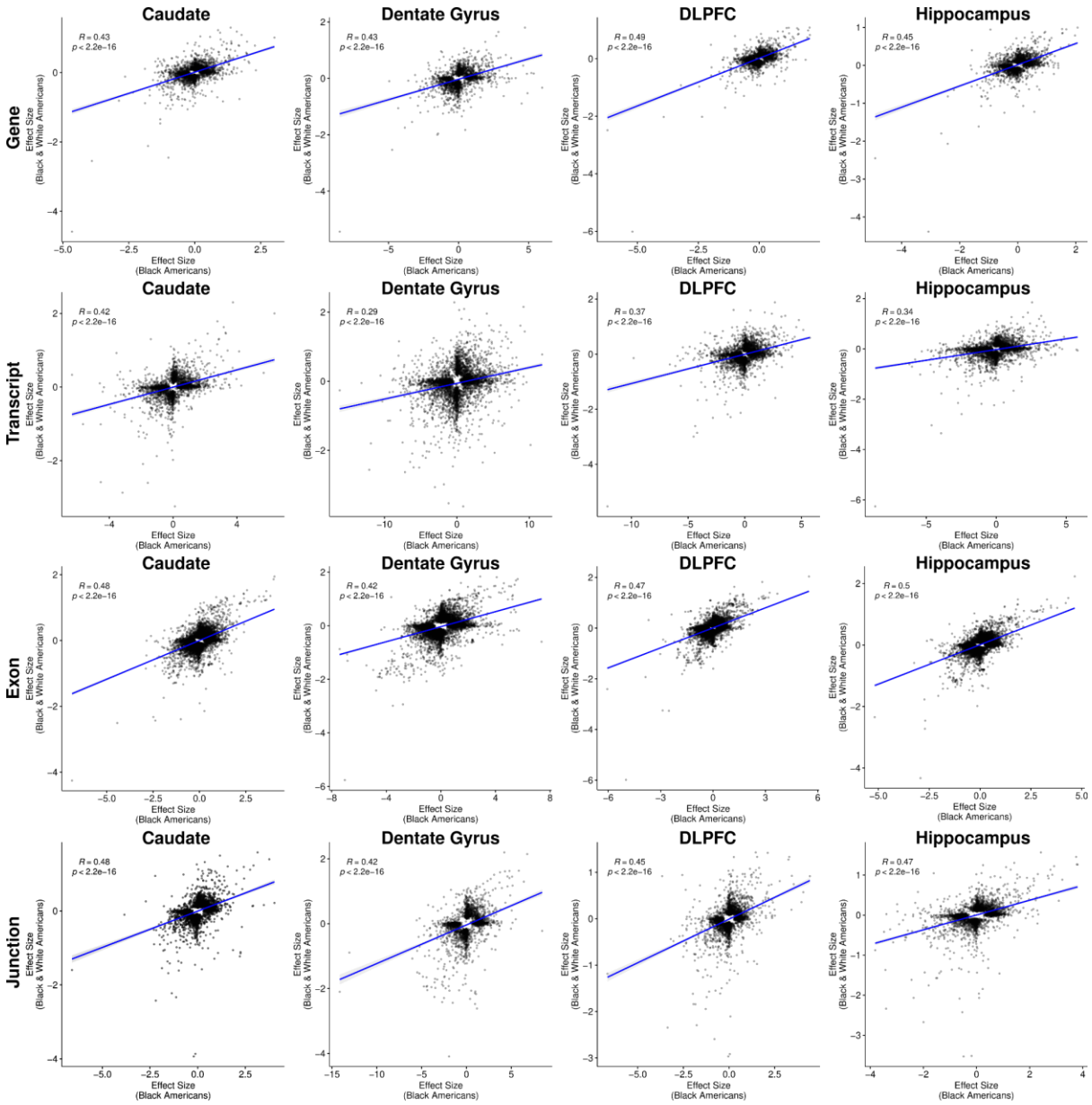

Fig. S30: Significant correlation of effect sizes between genetic ancestry association in admixed Black American individuals and self-reported race between Black and White Americans. Scatter plot showing significant correlation (two-sided, Spearman correlation) between Black American-only analysis (x-axis) and combined analysis (Black and White Americans; y-axis) for each brain region and feature (i.e., gene, transcript, exon, and junction). A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

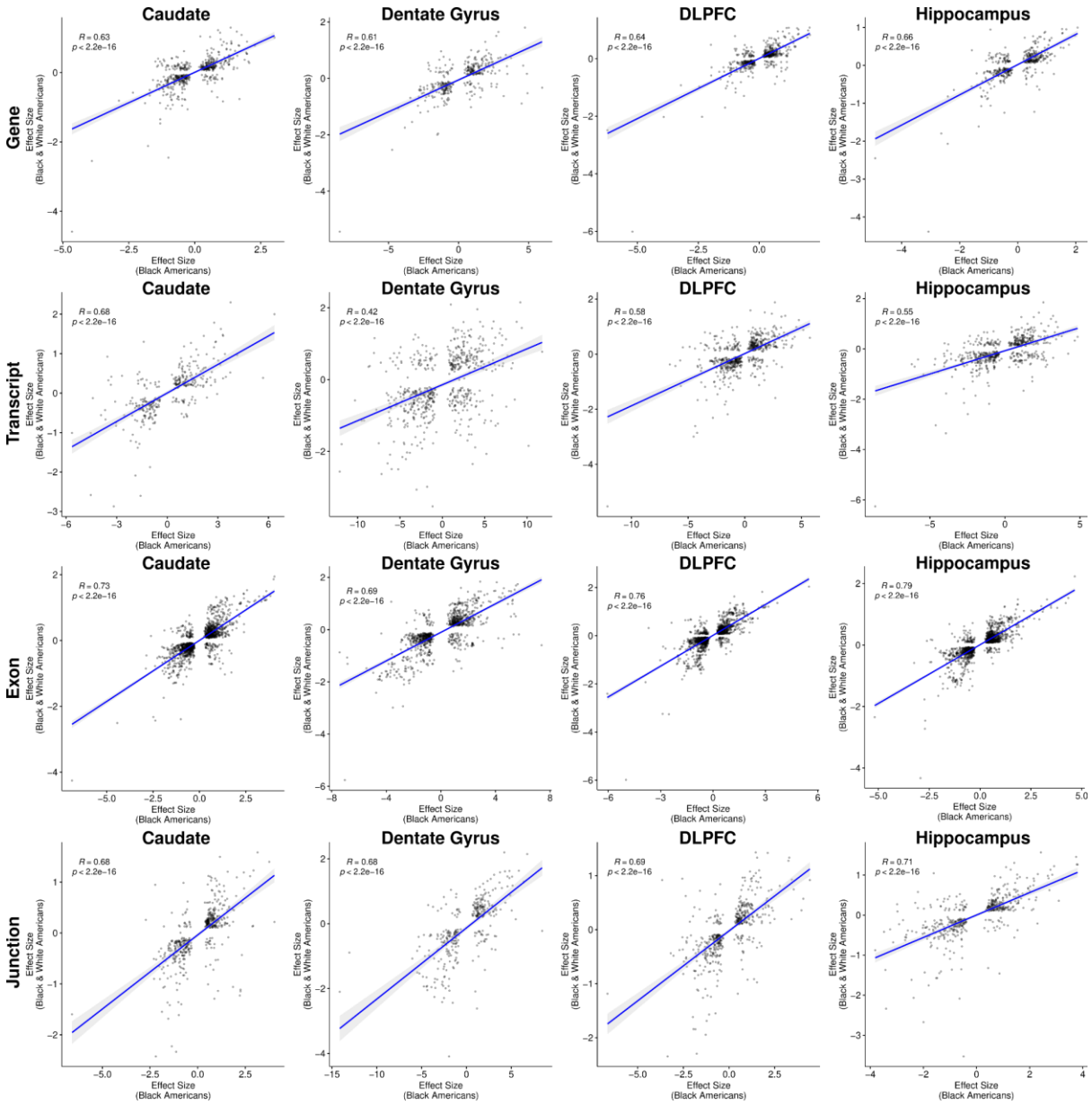

Fig. S31: Increased correlation of effect sizes for shared features between genetic ancestry association in admixed Black American individuals and self-reported race between Black and White Americans. Scatter plot showing significant correlation (two-sided, Spearman correlation) between Black American-only analysis (x-axis) and combined analysis (Black and White Americans; y-axis) for each brain region and feature (i.e., gene, transcript, exon, and junction). A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

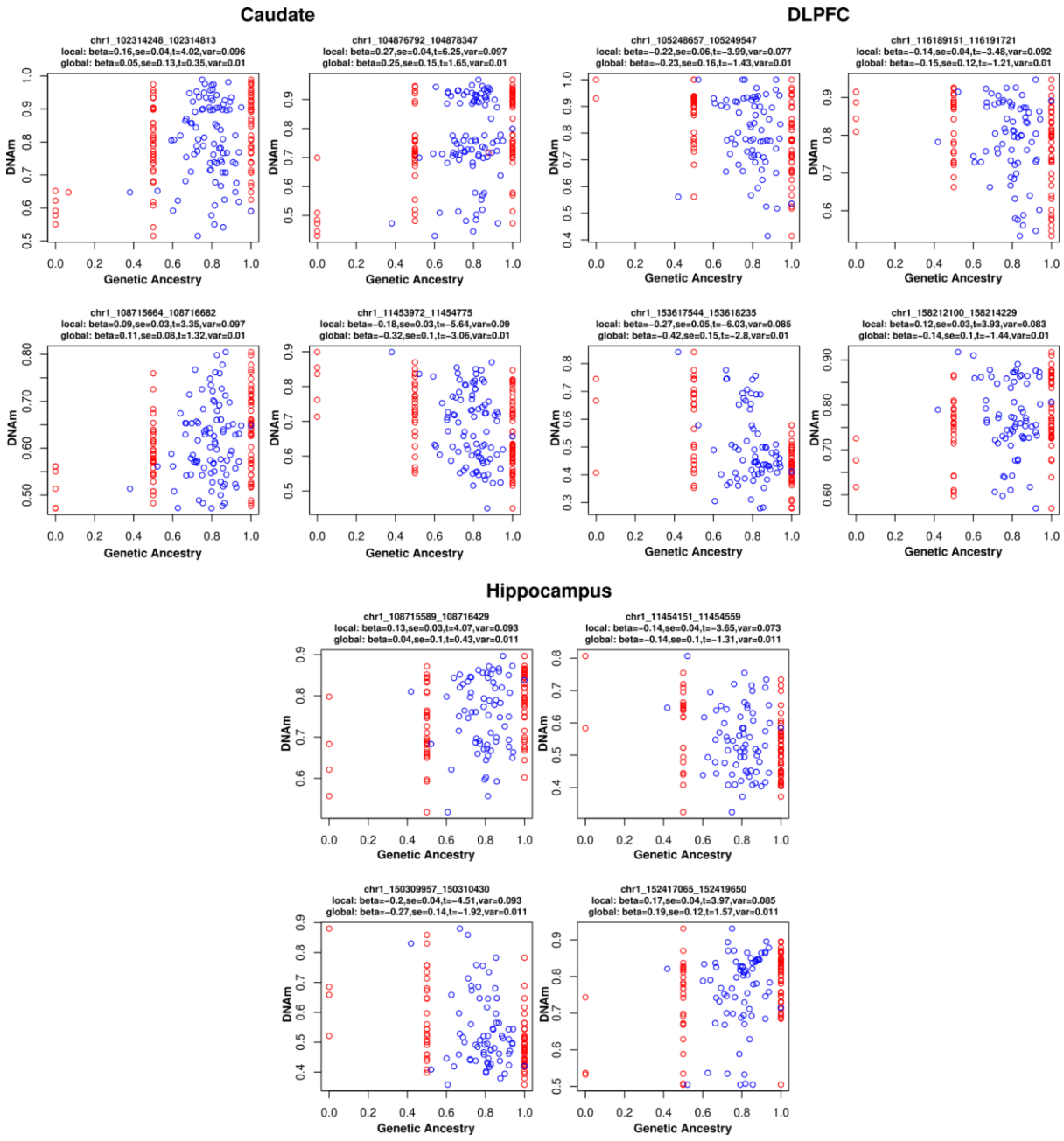

Fig. S32: Local ancestry is more variable than global ancestry. Example scatter plots showing local (red) and global (blue) ancestry associated with DNAm across brain regions. DMR test results (effect size [beta], standard error [se], and variance [var]) annotated on top of each example VMR.

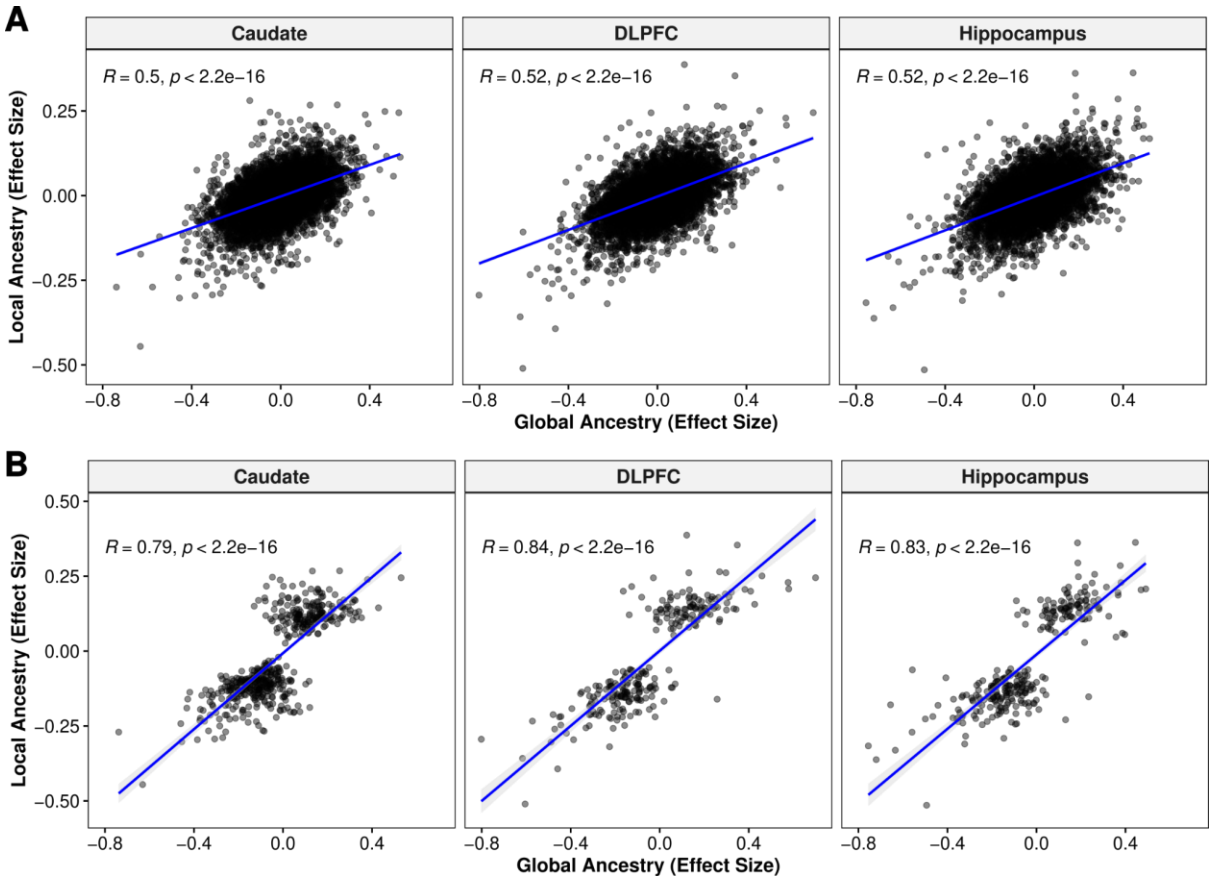

Fig. S33: Significant correlation of DNAm levels between local and global ancestry-associated DMRs across brain regions. Scatter plot comparing global (x-axis) and local (y-axis) for **A.** all ancestry-associated DMRs and **B.** significant ancestry-associated DMRs (FDR < 0.05). A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

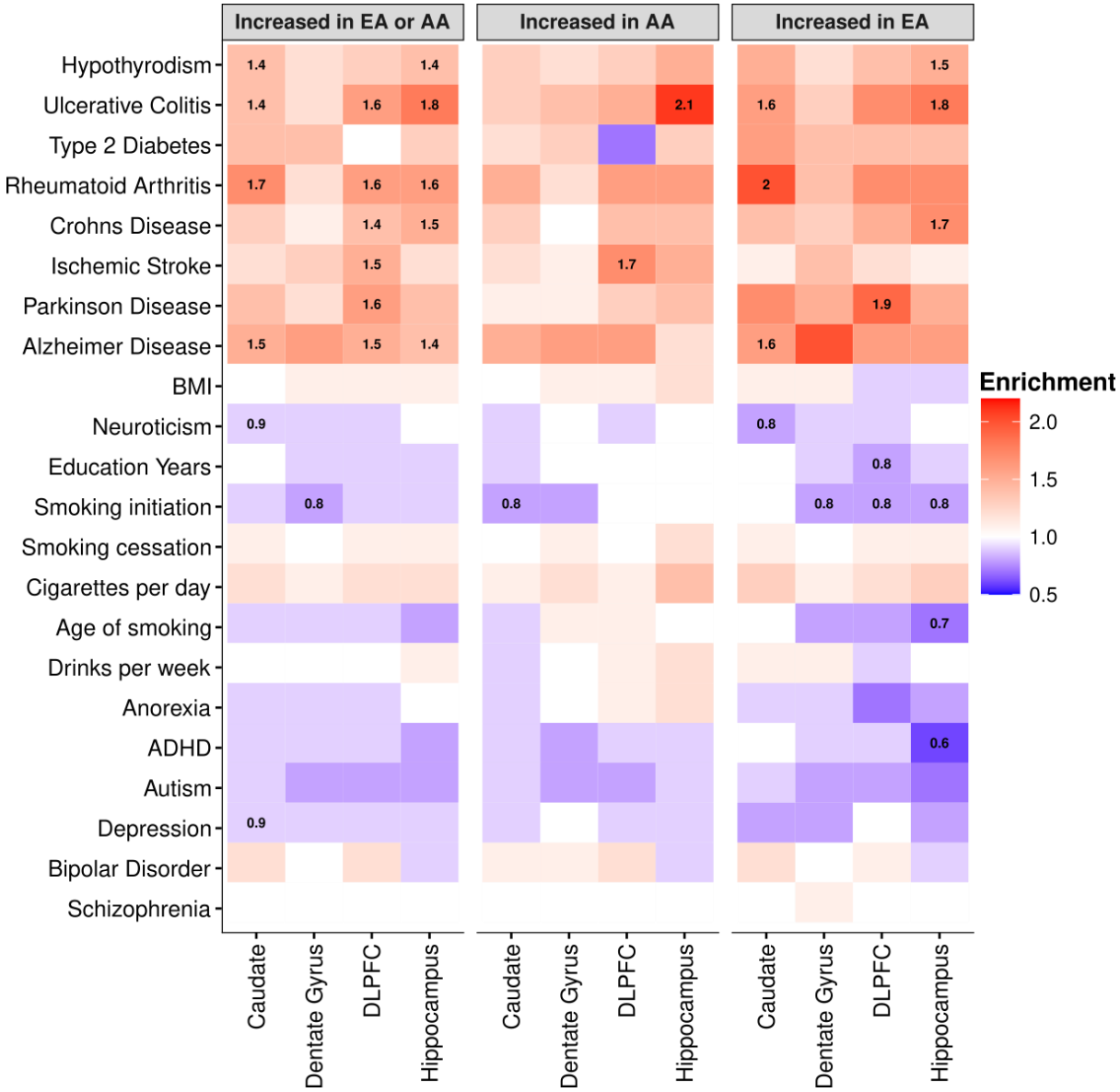

Fig. S34: **Global ancestry-associated DEGs show general enrichment for heritability of neurological and immune-related traits.** Heatmap for ancestry-associated DEGs that show no enrichment (red) nor depletion (blue) for heritability of brain- and immune-related traits from S-LDSC analysis. Numbers within tiles are levels of enrichment (> 1) or depletion (< 1) that are significant after multiple testing correction (FDR < 0.05). The left panel shows results for all DEG in each brain region. The middle and right panels show results for DEG increased with AA or EA proportions for each brain region, respectively.

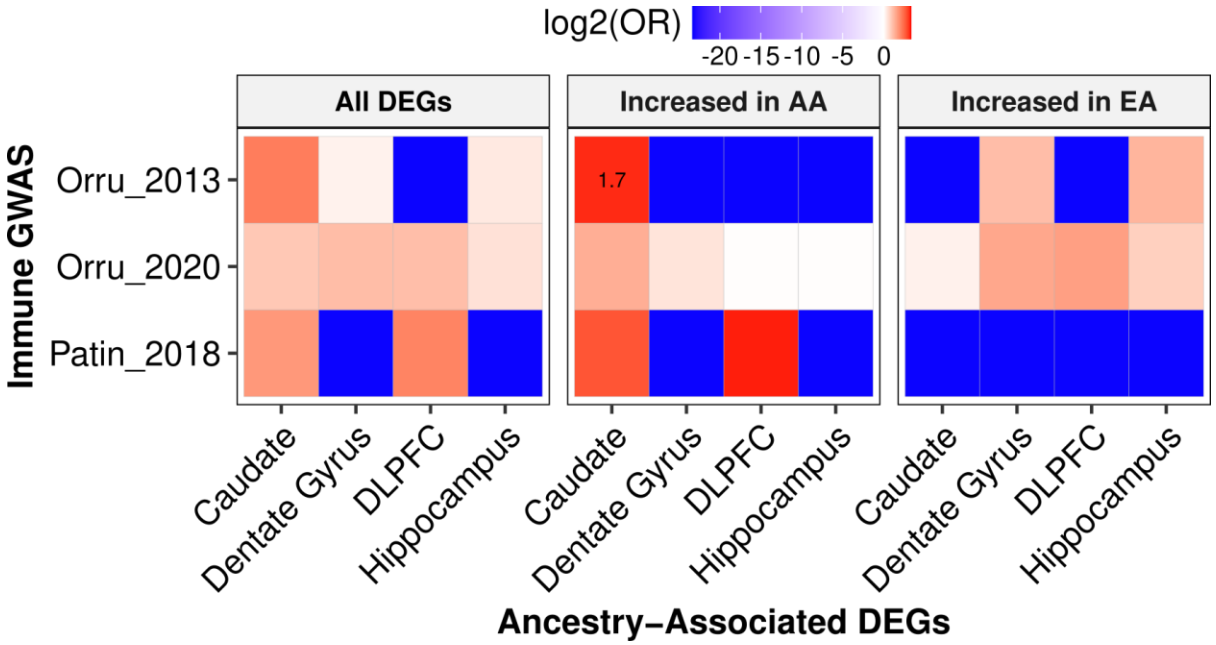

Fig. S35: Limited enrichment of non-brain immune function GWAS prioritized genes with global ancestry-associated DEGs. Heatmap showing enrichment analysis (two-sided, Fisher's exact test with p-values corrected for multiple testing with Benjamini-Hochberg) of significantly enriched (red) or depleted (blue) immune function GWAS prioritized genes for ancestry-associated DEGs (lfsr < 0.05) separated by direction of effect. Significant enrichments (-log10 transformed) annotated within tiles.

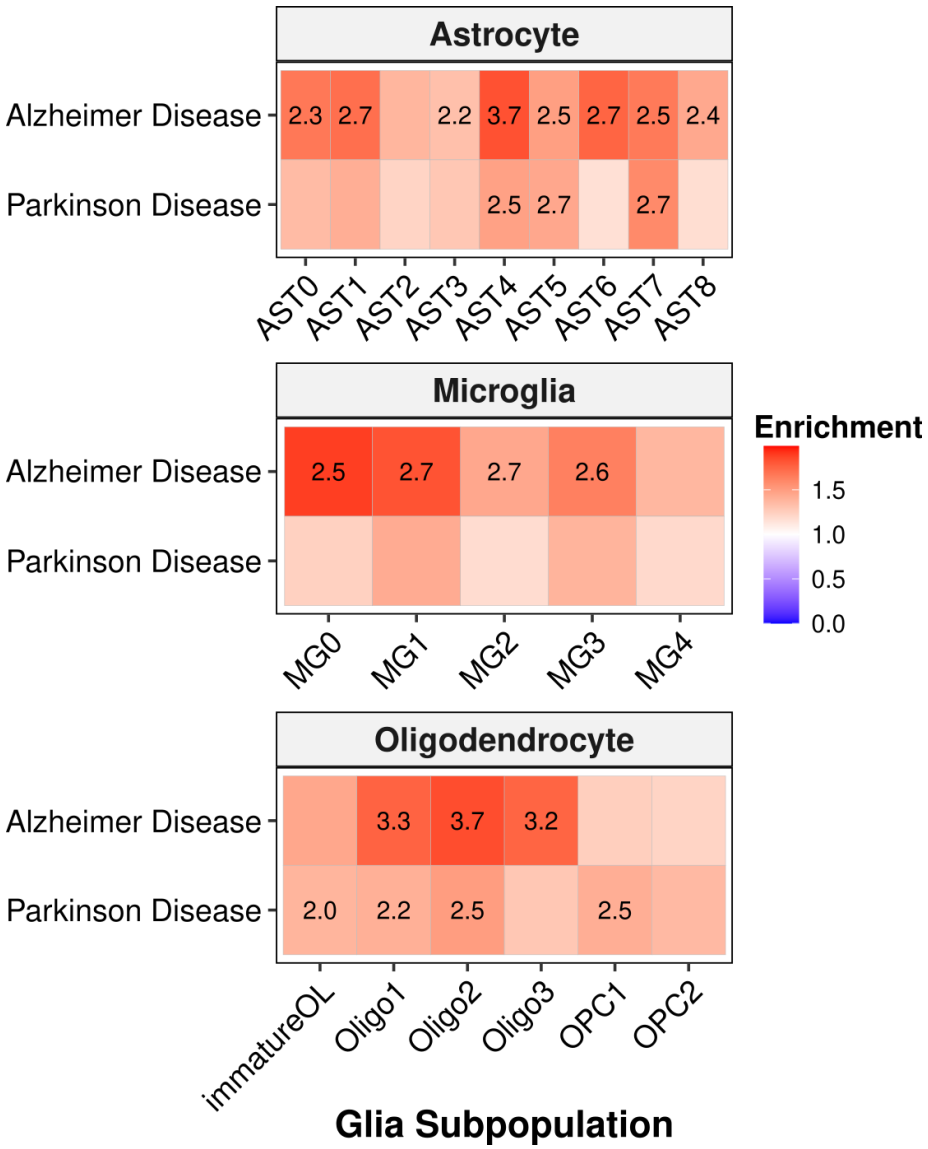

Fig. S36: Ancestry-associated glial cell subpopulations show general enrichment for heritability of Alzheimer's and Parkinson's diseases. Heatmap for glial cell subpopulations that show enrichment (red) for heritability of Alzheimer's and Parkinson's diseases from S-LDSC analysis. Numbers within tiles are levels of enrichment (> 1) or depletion (< 1) that are significant after multiple testing correction (FDR < 0.01).

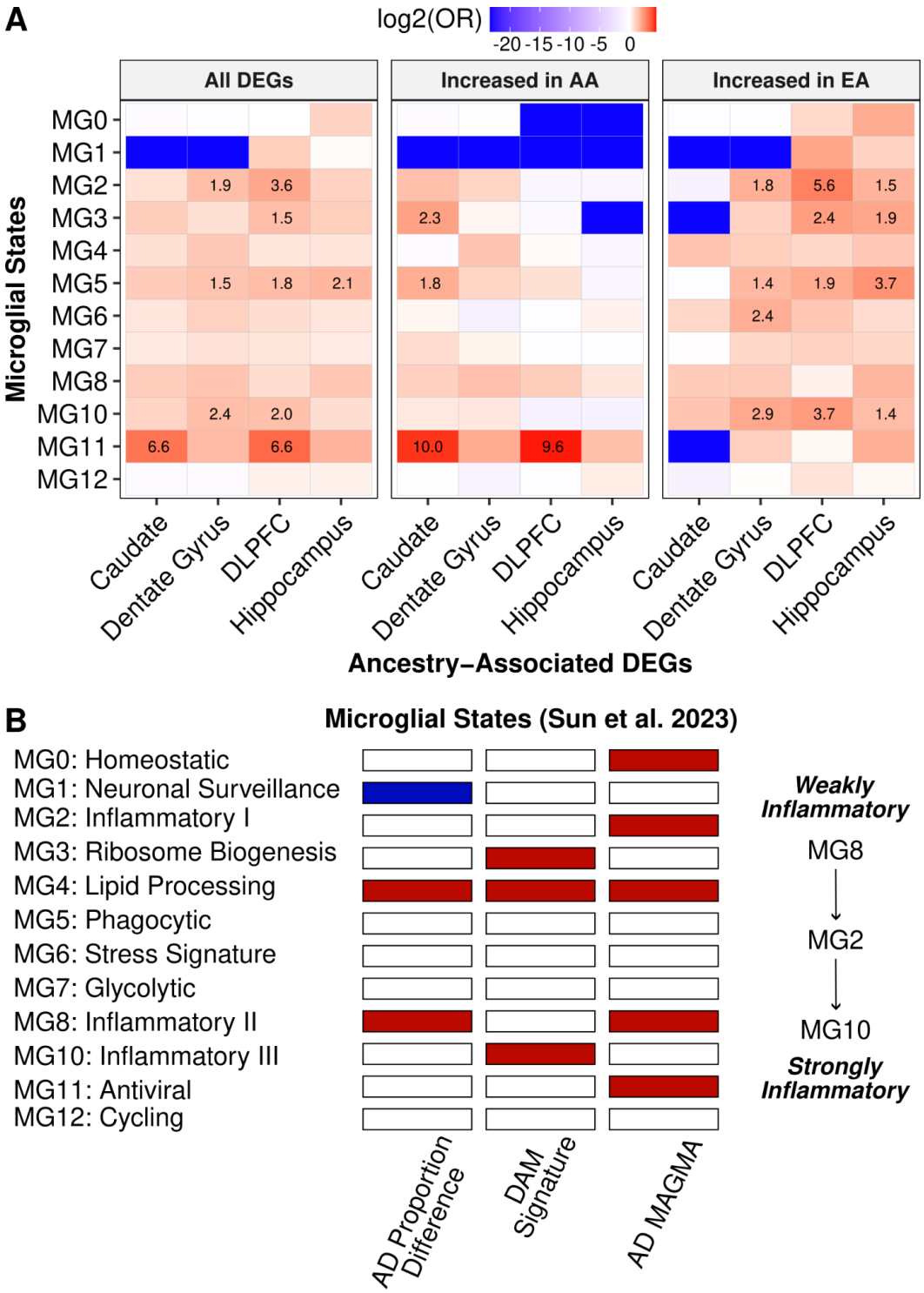

Fig. S37: Significant enrichment of ancestry-associated DEGs for activated microglia and disease-associated microglia. A. Heatmap showing enrichment analysis (two-sided, Fisher's exact test) of significantly enriched (red) or depleted (blue) microglia states (*40*) for ancestry-associated DEGs (lfsr < 0.05) separated by direction of effect. Significant enrichments ($-\log_{10}$ transformed) annotated within tiles. **B.** Annotation of microglia states (*40*) for Alzheimer's disease (AD): significant cell proportion differences, disease-associated microglia [DAM] signature in mouse models, and genetic association via MAGMA enrichment. Depletion is annotated in blue and enrichment in red.

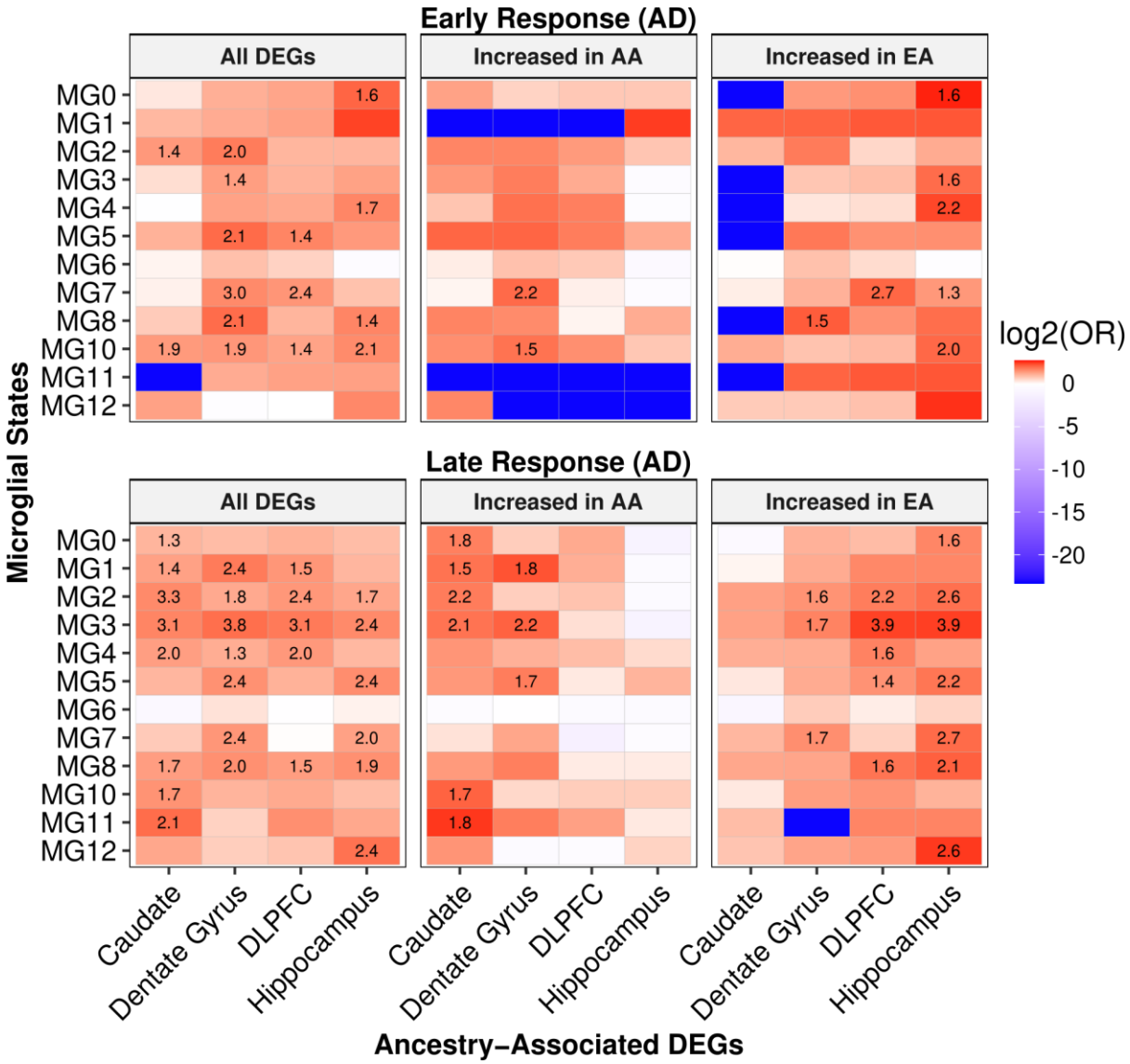

Fig. S38: Ancestry-associated DEGs are primarily enriched for microglial states associated with late-response Alzheimer's disease-related DEGs. Heatmap showing enrichment analysis (two-sided, Fisher's exact test) of ancestry DEGs (lfsr < 0.05) with cell type-specific Alzheimer's disease (AD) DEGs (*40*) separated response stage for ancestry-associated DEGs (lfsr < 0.05) separated by direction of effect. Early response is Alzheimer's DEGs detected between neurotypical control and early Alzhiemer's individual. Late response is Alzheimer's DEGs detected between early and late Alzheimer's individuals. Alzheimer's stage defined in (*40*). Significantly enriched (red) or depleted (blue) tiles are annotated with $-\log_{10}$(FDR).

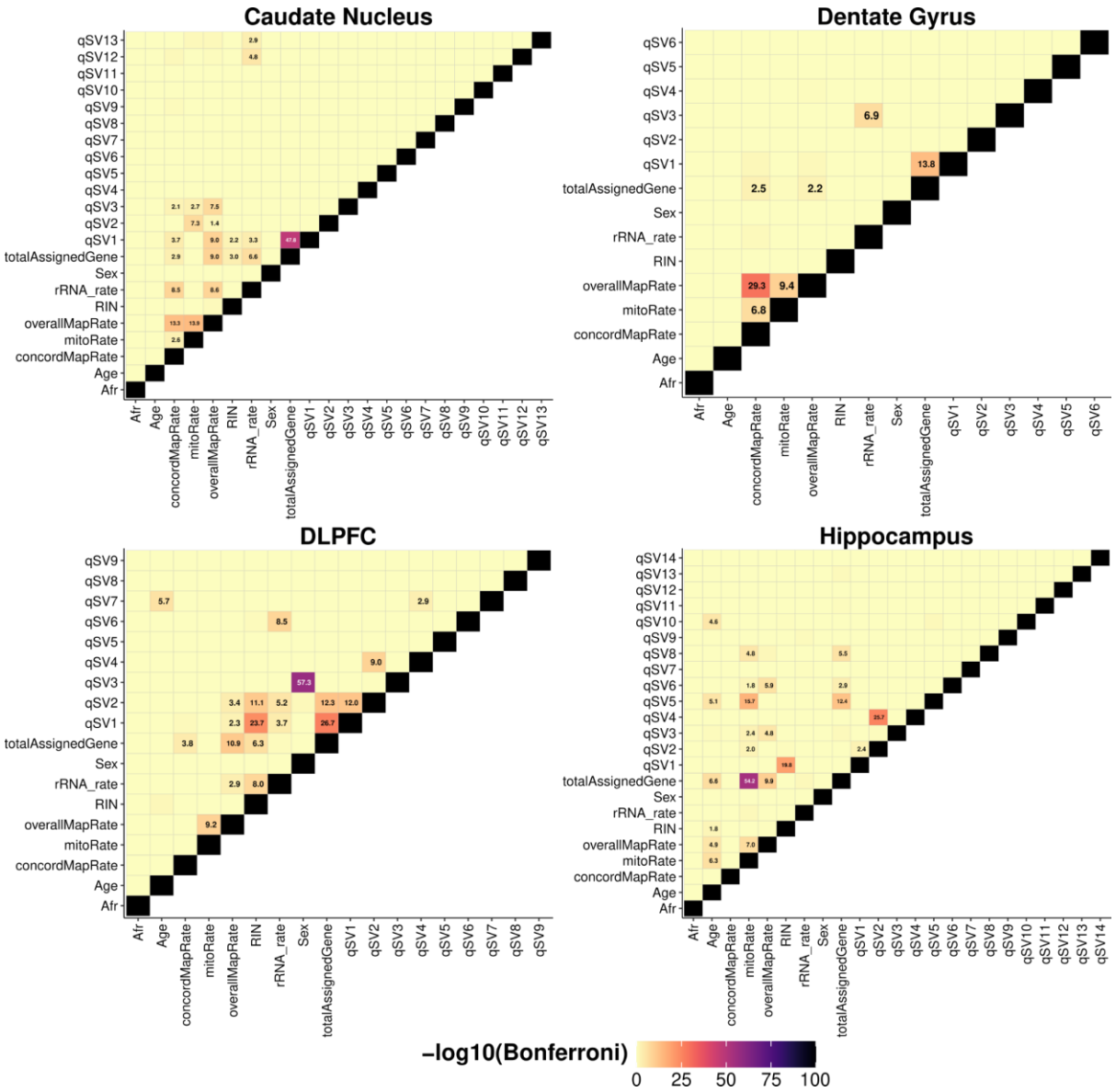

Fig. S39: Limited correlation between covariates across brain regions. Heatmap showing correlation between covariates across brain regions (linear regression, Bonferroni corrected p-values). Significant correlations (-$\log_{10}$ transformed) are denoted in each tile.

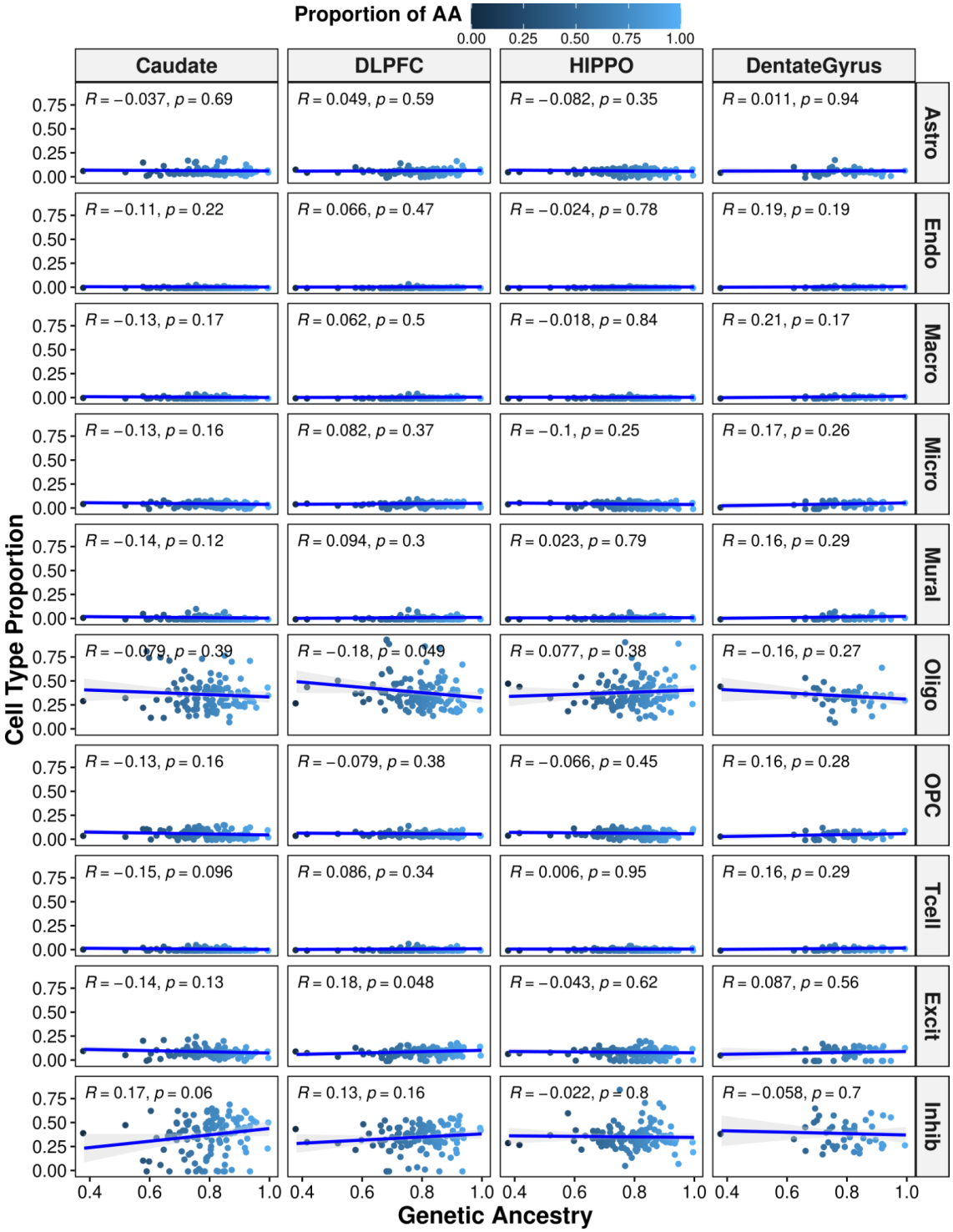

Fig. S40: Cell-type proportions show no correlation with genetic ancestry in Black American donors across brain regions. Scatter plot showing no correlation (two-sided, Spearman) between genetic ancestry and cell-type proportion across the brain. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray. Astro: astrocytes. Endo: endothelial. Micro: microglia. Macro: macrophage. Mural: mural cells. Oligo: oligodendrocytes. OPC: oligodendrocyte progenitor cells. Tcell: T cells. Excit: excitatory neurons. Inhib: inhibitory neurons.

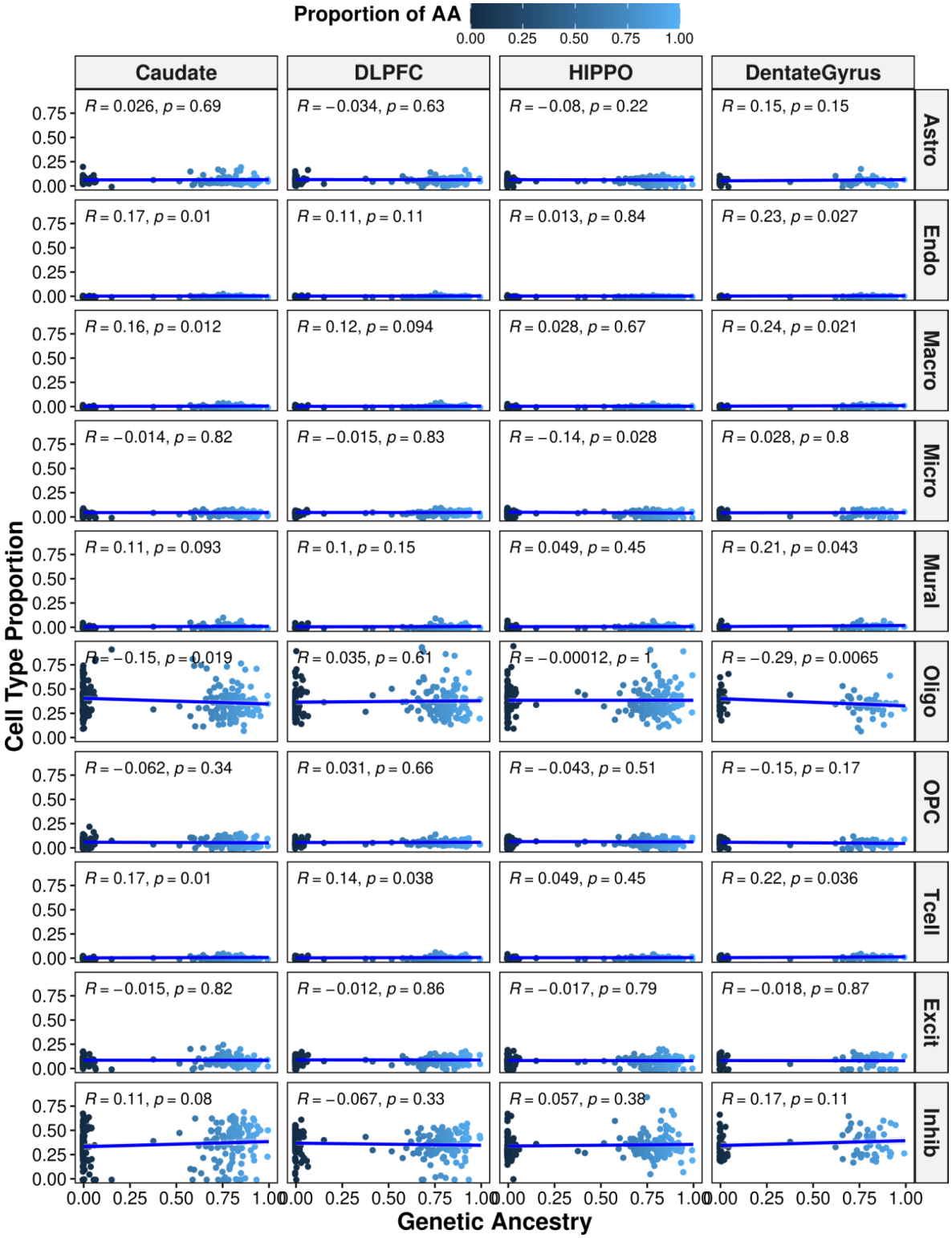

Fig. S41: Significant correlation of several cell-type proportions with genetic ancestry in Black American and White American donors across brain regions. Scatter plot showing correlation (two-sided, Spearman) between genetic ancestry and cell-type proportion across the brain. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray. Astro: astrocytes. Endo: endothelial. Micro: microglia. Macro: macrophage. Mural: mural cells. Oligo: oligodendrocytes. OPC: oligodendrocyte progenitor cells. Tcell: T cells. Excit: excitatory neurons. Inhib: inhibitory neurons.

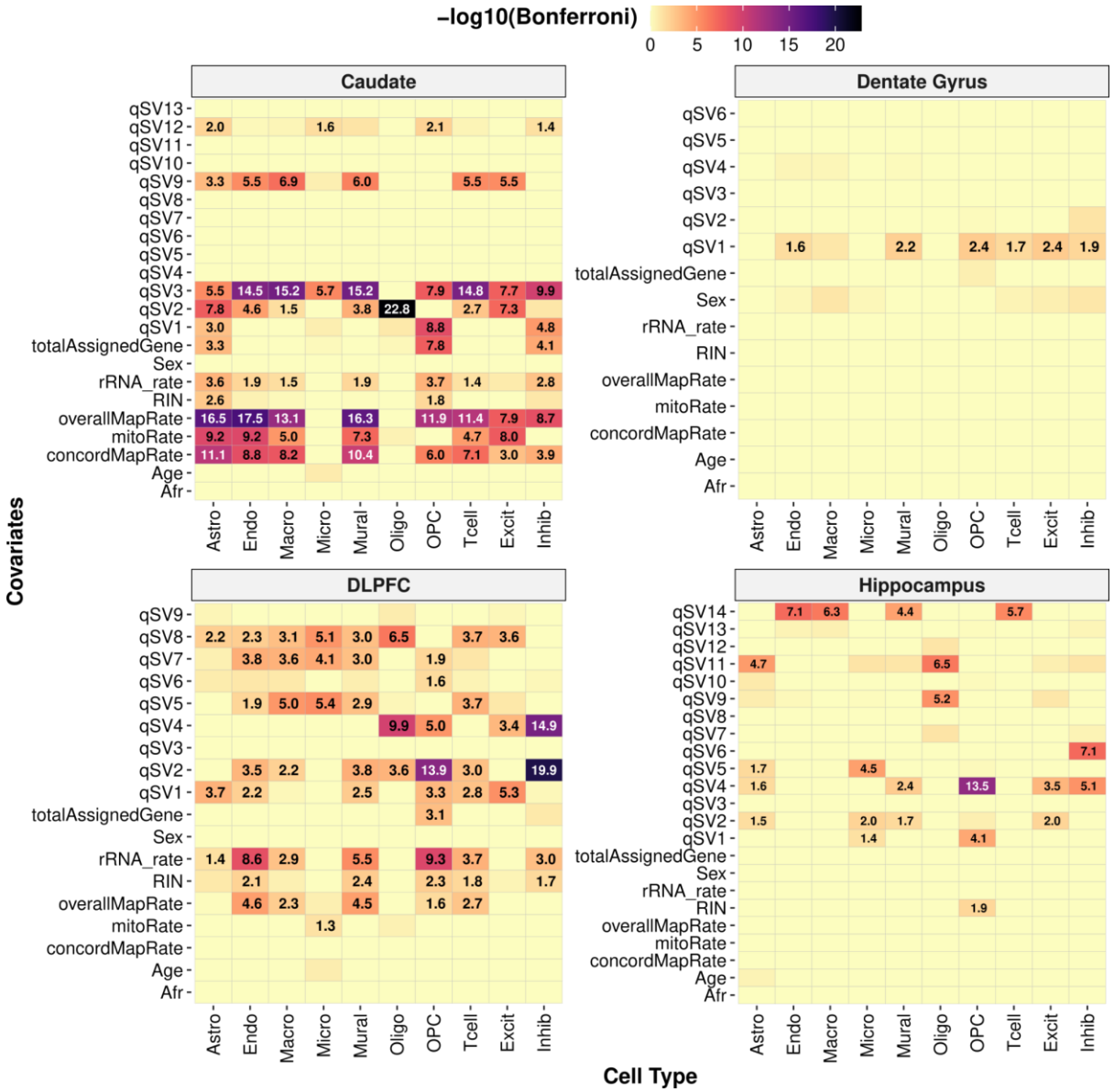

Fig. S42: Significant correlation between cell-type proportion and model covariations. Heatmap of correlation between covariates and cell-type proportion across the brain (linear regression, Bonferroni corrected p-values). Significant correlations (-$\log_{10}$ transformed) are denoted in each tile. Astro: astrocytes. Endo: endothelial. Micro: microglia. Macro: macrophage. Mural: mural cells. Oligo: oligodendrocytes. OPC: oligodendrocyte progenitor cells. Tcell: T cells. Excit: excitatory neurons. Inhib: inhibitory neurons.

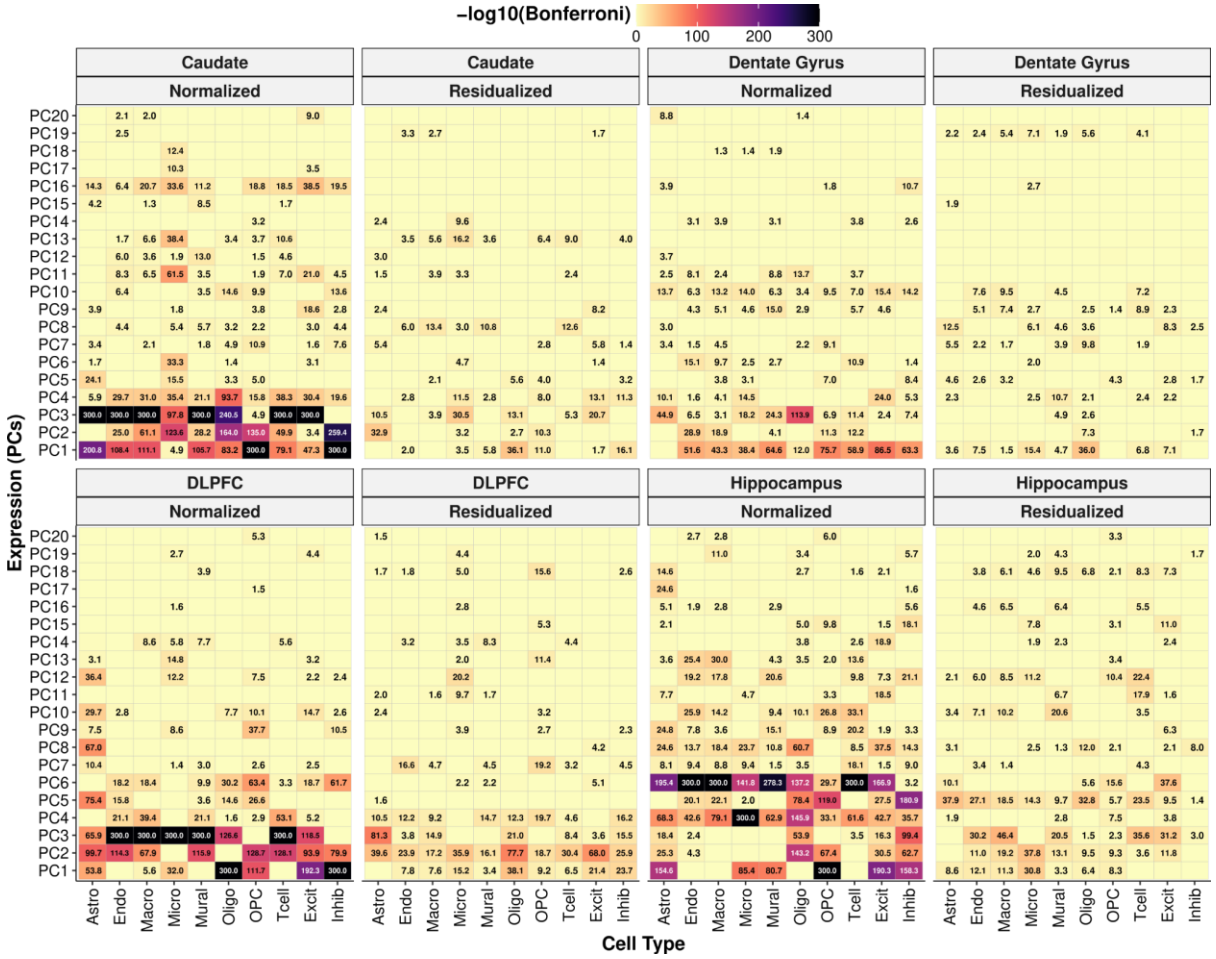

Fig. S43: Cell-type proportion partially corrected for by covariates. Heatmap of correlation between PCA of gene expression and cell-type proportions before (normalized) and after (residualized) adjusting for covariates including qSVs across brain regions (linear regression, Bonferroni corrected p-values). Significant correlations ($-\log_{10}$ transformed) are denoted in each tile. Astro: astrocytes. Endo: endothelial. Micro: microglia. Macro: macrophage. Mural: mural cells. Oligo: oligodendrocytes. OPC: oligodendrocyte progenitor cells. Tcell: T cells. Excit: excitatory neurons. Inhib: inhibitory neurons.

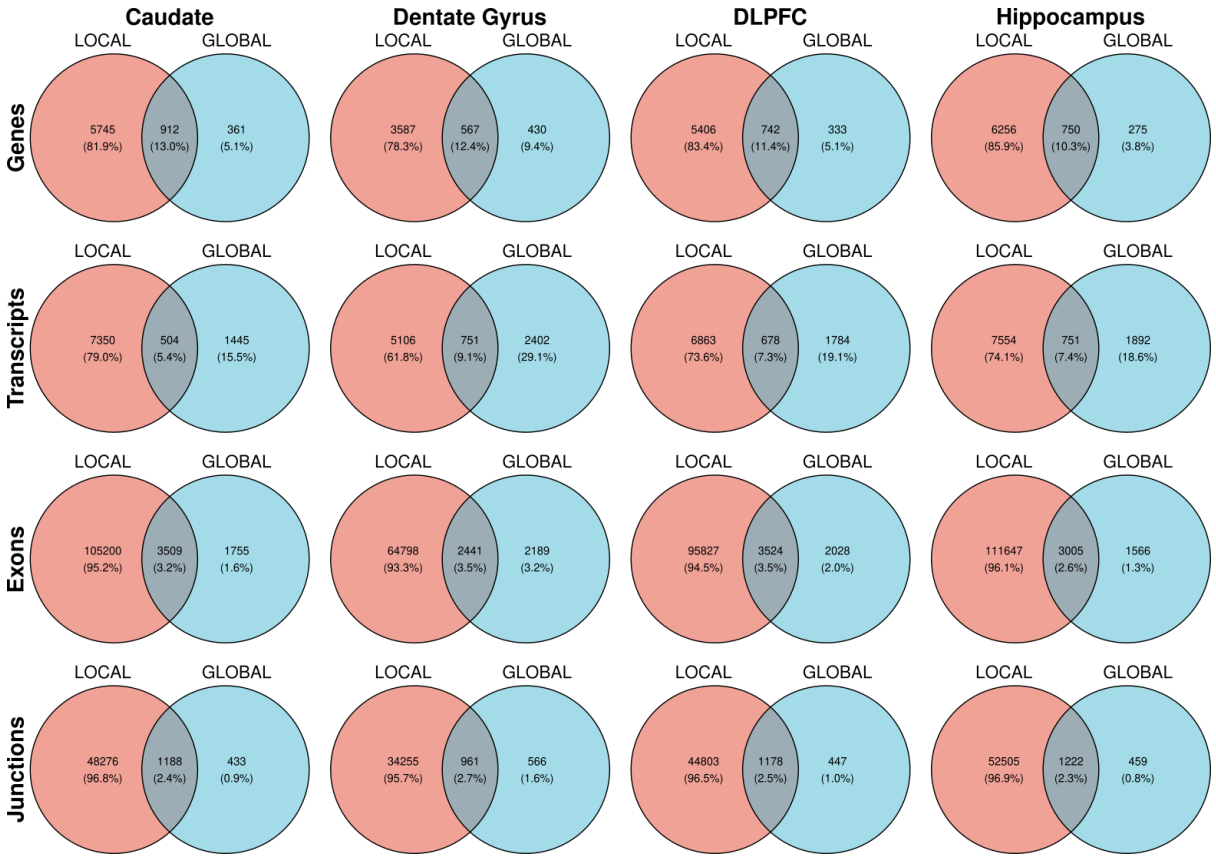

Fig. S44: Significant overlap of global ancestry-associated DE features with local ancestry analysis. Venn diagram showing the overlap between local ancestry-associated DE features (i.e., gene, transcript, exon, and junction) and global ancestry-associated DE features across brain regions. Significant overlap tested with two-sided, Fisher's exact test (FDR < 0.05).

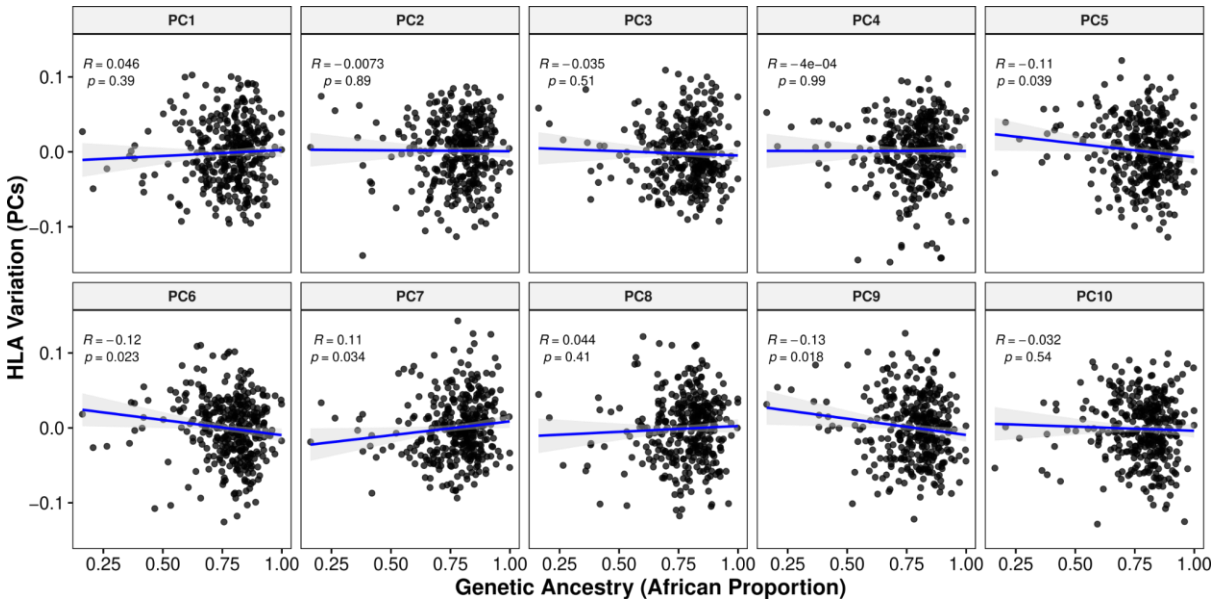

Fig. S45: Little to no correlation between HLA variation PCs and global genetic ancestry. Scatter plot showing correlation (two-sided, Spearman) between genetic ancestry and HLA variation PCs. A fitted trend line is presented in blue as the mean values +/- standard deviation. The standard deviation is shaded in light gray.

Supplementary Tables

Table S1: Black American sample characteristics for adult (age > 17) neurotypical control postmortem caudate, dentate gyrus, DLPFC, and hippocampus (*10–12*). Abbreviations: RNA integrity number (RIN).

	Brain Region			
Characteristic	**Caudate** N = 122[1]	**Dentate Gyrus** N = 47[1]	**DLPFC** N = 123[1]	**Hippocampus** N = 133[1]
Sex				
Female	50 (41%)	16 (34%)	48 (39%)	53 (40%)
Male	72 (59%)	31 (66%)	75 (61%)	80 (60%)
Age	46 (15)	46 (16)	44 (15)	43 (15)
RIN	7.83 (0.80)	5.45 (1.22)	7.70 (0.89)	7.72 (0.98)

[1] n (%); Mean (SD)

Table S2. Summary of global ancestry-associated differential expression results (lfsr < 0.05) by feature (gene, transcript, exon, and exon-exon junction) for ancestry differences within admixed AA (n=151) in the caudate (n=122), dentate gyrus (n=47), DLPFC (n=123), and hippocampus (n=133). The number of unique genes associated with transcript, exon, or junction is in parentheses.

Brain Region	Gene	Transcript (Geneid)	Exon (Geneid)	Junction (Geneid)
Caudate	1,273	1,949 (1,728)	5,264 (2,991)	1,621 (1,116)
Dentate Gyrus	997	3,153 (2,701)	4,630 (2,737)	1,527 (1,105)
DLPFC	1,075	2,462 (2,126)	5,552 (3,140)	1,625 (1,138)
Hippocampus	1,025	2,643 (2,263)	4,571 (2,717)	1,681 (1,177)

Table S3. Summary of local ancestry-associated differential expression results (lfsr < 0.05) by feature (gene, transcript, exon, and exon-exon junction) for ancestry differences within admixed AA (n=149) in the caudate (n=120), dentate gyrus (n=45), DLPFC (n=121), and hippocampus (n=131). The number of unique genes associated with transcript, exon, or junction is in parentheses.

Brain Region	Gene	Transcript (Geneid)	Exon (Geneid)	Junction (Geneid)
Caudate	6,657	7,854 (5,746)	108,709 (12,287)	49,464 (8,385)
Dentate Gyrus	4,154	5,857 (4,613)	67,239 (10,974)	35,216 (7,918)
DLPFC	6,148	7,541 (5,561)	99,351 (12,051)	45,981 (8,282)
Hippocampus	7,006	8,305 (5,996)	114,652 (12,411)	53,727 (8,510)

Table S4: Summary of main effect cis-eQTL (lfsr < 0.05) in Black American admixed individuals (n=148) by feature (gene, transcript, exon, and exon-exon junction) across the caudate (n=120), dentate gyrus (n=45), DLPFC (n=121), and hippocampus (n=131). eFeature: unique feature. eGene: unique gene ID.

		Caudate	**Dentate Gyrus**	**DLPFC**	**Hippocampus**
Gene	*eQTL*	698,047	605,755	728,533	688,628
	eFeature	10,867	11,664	11,173	10,408
	eGene	10,867	11,664	11,173	10,408
Transcript	*eQTL*	934,240	1,062,967	968,814	940,405
	eFeature	17,759	32,342	18,422	17,581
	eGene	10,369	14,674	10,710	10,320
Exon	*eQTL*	1,503,349	1,532,590	1,551,027	1,461,400
	eFeature	29,203	37,894	30,091	27,560
	eGene	10,423	12,675	10,612	9,927
Junction	*eQTL*	502,183	601,181	496,775	480,246
	eFeature	11,135	25,694	10,831	10,165
	eGene	3,084	4,768	3,022	2,874

Table S5: Summary of ancestry-dependent cis-eQTL (lfsr < 0.05) in Black American admixed individuals (n=148) by feature (gene, transcript, exon, and exon-exon junction) across the caudate (n=120), dentate gyrus (n=45), DLPFC (n=121), and hippocampus (n=131). eFeature: unique feature. eGene: unique gene ID.

		Caudate	**Dentate Gyrus**	**DLPFC**	**Hippocampus**
Gene	*eQTL*	3,281	5,484	3,441	3,371
	eFeature	531	942	573	531
	eGene	531	942	573	531
Transcript	*eQTL*	3,849	28,102	3,853	4,315
	eFeature	617	4,443	619	716
	eGene	580	3,529	582	671
Exon	*eQTL*	2,353	4,346	2,189	2,196
	eFeature	510	773	483	483
	eGene	330	481	314	312
Junction	*eQTL*	2,452	61,494	2,609	3,111
	eFeature	436	8,427	508	613
	eGene	275	4,748	332	419

Table S6. Black and White Americans sample breakdown for adult (age > 17) neurotypical control postmortem caudate, dentate gyrus, DLPFC, and hippocampus (*10–12*). Abbreviations: Female (F), Male (M), Black American (BA), White American (WA), and RNA integrity number (RIN).

	Caudate		Dentate Gyrus		DLPFC		Hippocampus	
Characteristic	**BA** N = 122[1]	**WA** N = 118[1]	**BA** N = 47[1]	**WA** N = 43[1]	**BA** N = 123[1]	**WA** N = 89[1]	**BA** N = 133[1]	**WA** N = 110[1]
Sex								
Female	50 (41%)	21 (18%)	16 (34%)	10 (23%)	48 (39%)	18 (20%)	53 (40%)	21 (19%)
Male	72 (59%)	97 (82%)	31 (66%)	33 (77%)	75 (61%)	71 (80%)	80 (60%)	89 (81%)
Age	46 (15)	51 (17)	46 (16)	50 (13)	44 (15)	47 (14)	43 (15)	46 (15)
RIN	7.83 (0.80)	7.87 (0.79)	5.45 (1.22)	5.14 (1.22)	7.70 (0.89)	7.70 (0.88)	7.72 (0.98)	7.76 (0.96)

[1] n (%); Mean (SD)

Table S7. Summary of differential expression results (lfsr < 0.05) by feature (gene, transcript, exon, and exon-exon junction) for ancestry differences in the caudate (n=240), dentate gyrus (n = 90), DLPFC (n=212), and hippocampus (n=243). The number of unique genes associated with transcript, exon, or junction is in parentheses.

Brain Region	Gene	Transcript (Geneid)	Exon (Geneid)	Junction (Geneid)
caudate	4,238	7,752 (5,451)	30,858 (9,303)	9,426 (4,181)
Dentate Gyrus	3,395	9,758 (6,543)	23,407 (8,145)	7,770 (3,848)
DLPFC	4,226	9,396 (6,288)	31,834 (9,524)	9,467 (4,231)
Hippocampus	4,025	9,456 (6,370)	29,340 (9,095)	9,034 (4,071)

Supplementary Data

1. **Data S1. BrainSeq_ancestry_4features_4regions_allFeatures.txt.gz**: Compressed text file of differential expression analysis after mash modeling for global genetic ancestry (continuous) across the caudate, dentate gyrus, DLPFC, and hippocampus for four features (gene, transcript, exon, and junction).
2. **Data S2. BrainSeq_ancestry_local_4features_4regions_allFeatures.txt.gz**: Compressed text file of differential expression analysis after mash modeling for local genetic ancestry (continuous by feature) across the caudate, dentate gyrus, DLPFC, and hippocampus for four features (gene, transcript, exon, and junction).
3. **Data S3. DE_functional_enrichment_ancestry_AAonly.xlsx**: Excel file of GO-term enrichment and gene set enrichment analysis (GSEA) for genetic ancestry (continuous) differentially expressed genes across the caudate, dentate gyrus, DLPFC, and hippocampus.
4. **Data S4**. **WGCNA_functional_enrichment_analysis_ancestry_AAonly.xlsx**: Excel file of GO-term enrichment for genetic ancestry-associated WGCNA modules across brain regions.
5. **Data S5. WGCNA_DEG_enrichment_modules_GO_analysis.tar.gz**: Compressed directory of ancestry-associated DEGs enriched for WGCNA module functional enrichment results (i.e., GO-term enrichment) for the caudate, dentate gyrus, DLPFC, and hippocampus.
6. **Data S6. BrainSeq_main_eQTL_4features_4regions_significant.txt.gz**: Compressed text file of main effect eQTL results (lfsr < 0.05), variant-feature pairs across the caudate, dentate gyrus, DLPFC, and hippocampus for four features (gene, transcript, exon, and junction).
7. **Data S7. BrainSeq_ancestry_dependent_eQTL_4features_4regions_significant.txt.gz**: Compressed text file of genetic ancestry-dependent eQTL results (lfsr < 0.05), variant-feature pairs across the caudate, dentate gyrus, DLPFC, and hippocampus for four features (gene, transcript, exon, and junction).
8. **Data S8. BrainSeq_ancestry_binary_4features_4regions_allFeatures.txt.gz**: Compressed text file of binary differential expression analysis (10 permutations) for genetic ancestry (binary) across the caudate, dentate gyrus, DLPFC, and hippocampus for four features (gene, transcript, exon, and junction).
9. **Data S9**. **DE_binary_validation_functional_enrichment_ancestry_AA_EA.xlsx**: Excel file of GO-term enrichment and GSEA for internal validation for genetic ancestry (binary) differentially expressed genes across the caudate, dentate gyrus, DLPFC, and hippocampus for four features (gene, transcript, exon, and junction).
10. **Data S10. BrainSeq_DEancestry_LDSC_AAonly.xlsx**: Excel file of stratified LD score regression of admixed Black American differential expression analysis separated by direction of effect (all DEGs, upregulated in AA, or upregulated in EA) for genes (SNP proportion > 0.01) across the caudate, dentate gyrus, DLPFC, and hippocampus.
11. **Data S11**. **DMR_functional_enrichment_localAncestry_AAonly.xlsx**: Excel file of GO-term enrichment for genetic ancestry differential methylation regions across the caudate, DLPFC, and hippocampus.
12. **Data S12**. **BrainSeq_DMR_global_local_comparison.tar.gz**: Compressed directory of PDF of scatter plots comparing DNAm association with local and global ancestry for the caudate, DLPFC, and hippocampus. Plots are annotated with genetic ancestry DMR test results.
13. **Data S13. BrainSeq_est_prop_Bisque.Rdata**: R variable containing estimated cell type proportions using Bisque for the caudate, dentate gyrus, DLPFC, and hippocampus.
14. **Data S14. gwas_summary_statistics_ldsc.xlsx**: Excel file of GWAS summary statistics for heritability enrichment analysis.

Supplementary References

56. W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler, The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

57. S. Das, L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

58. C. Fuchsberger, G. R. Abecasis, D. A. Hinds, minimac2: faster genotype imputation. *Bioinformatics*. **31**, 782–784 (2015).

59. P.-R. Loh, P. Danecek, P. F. Palamara, C. Fuchsberger, Y. A Reshef, H. K Finucane, S. Schoenherr, L. Forer, S. McCarthy, G. R. Abecasis, R. Durbin, A. L Price, Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

60. S. Purcell, C. Chang, *PLINK* (2021; http://www.cog-genomics.org/plink/2.0/).

61. Y. Luo, M. Kanai, W. Choi, X. Li, K. Yamamoto, K. Ogawa, M. Gutierrez-Arcelus, P. K. Gregersen, P. E. Stuart, J. T. Elder, J. Fellay, M. Carrington, D. W. Haas, X. Guo, N. D. Palmer, Y.-D. I. Chen, Jerome. I. Rotter, Kent. D. Taylor, Stephen. S. Rich, A. Correa, S. Raychaudhuri, A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. *medRxiv* (2020), doi:10.1101/2020.07.16.20155606.

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

63. K. A. Perzel Mandell, N. J. Eagles, A. Deep-Soboslay, R. Tao, S. Han, R. Wilton, A. S. Szalay, T. M. Hyde, J. E. Kleinman, A. E. Jaffe, D. R. Weinberger, Molecular phenotypes associated with antipsychotic drugs in the human caudate nucleus. *Mol. Psychiatry*. **27**, 2061–2067 (2022).

64. F. Krueger, F. James, P. Ewels, E. Afyounian, B. Schuster-Boeckler, TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. *Zenodo* (2021), doi:10.5281/zenodo.5127899.

65. R. Wilton, X. Li, A. P. Feinberg, A. S. Szalay, Arioc: GPU-accelerated alignment of short bisulfite-treated reads. *Bioinformatics*. **34**, 2673–2675 (2018).

66. G. G. Faust, I. M. Hall, SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*. **30**, 2503–2505 (2014).

67. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

68. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. **27**, 1571–1572 (2011).

69. K. D. Hansen, B. Langmead, R. A. Irizarry, BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* **13**, R83 (2012).

70. B. K. Maples, S. Gravel, E. E. Kenny, C. D. Bustamante, RFMix: a discriminative modeling

approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).

71. S. Fairley, E. Lowy-Gallego, E. Perry, P. Flicek, The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Res.* **48**, D941–D947 (2020).

72. B. Jew, M. Alvarez, E. Rahmani, Z. Miao, A. Ko, K. M. Garske, J. H. Sul, K. H. Pietiläinen, P. Pajukanta, E. Halperin, Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nat. Commun.* **11**, 1971 (2020).

73. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. **26**, 139–140 (2010).

74. D. J. McCarthy, Y. Chen, G. K. Smyth, Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

75. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

76. E. B. Stovner, P. Sætrom, PyRanges: efficient comparison of genomic intervals in Python. *Bioinformatics*. **36**, 918–919 (2020).

77. J. Bryois, N. G. Skene, T. F. Hansen, L. J. A. Kogelman, H. J. Watson, Z. Liu, Eating Disorders Working Group of the Psychiatric Genomics Consortium, International Headache Genetics Consortium, 23andMe Research Team, L. Brueggeman, G. Breen, C. M. Bulik, E. Arenas, J. Hjerling-Leffler, P. F. Sullivan, Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* **52**, 482–493 (2020).

78. M. L. Speir, A. Bhaduri, N. S. Markov, P. Moreno, T. J. Nowakowski, I. Papatheodorou, A. A. Pollen, B. J. Raney, L. Seninge, W. J. Kent, M. Haeussler, UCSC Cell Browser: visualize your single-cell data. *Bioinformatics*. **37**, 4578–4580 (2021).

79. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, R. Satija, Integrated analysis of multimodal single-cell data. *Cell*. **184**, 3573-3587. (2021).

80. R. A. Amezquita, A. T. L. Lun, E. Becht, V. J. Carey, L. N. Carpp, L. Geistlinger, F. Marini, K. Rue-Albrecht, D. Risso, C. Soneson, L. Waldron, H. Pagès, M. L. Smith, W. Huber, M. Morgan, R. Gottardo, S. C. Hicks, Orchestrating single-cell analysis with Bioconductor. *Nat. Methods*. **17**, 137–145 (2020).

81. D. J. McCarthy, K. R. Campbell, A. T. L. Lun, Q. F. Wills, Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*. **33**, 1179–1186 (2017).

82. L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

83. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, N. Yosef, Deep generative modeling for single-cell transcriptomics. *Nat. Methods*. **15**, 1053–1058 (2018).

84. A. Gayoso, R. Lopez, G. Xing, P. Boyeau, V. Valiollah Pour Amiri, J. Hong, K. Wu, M. Jayasuriya, E. Mehlman, M. Langevin, Y. Liu, J. Samaran, G. Misrachi, A. Nazaret, O. Clivio, C. Xu, T. Ashuach, M. Gabitto, M. Lotfollahi, V. Svensson, N. Yosef, A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).

85. C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, N. Yosef, Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).

86. B. Phipson, C. B. Sim, E. R. Porrello, A. W. Hewitt, J. Powell, A. Oshlack, propeller: testing for differences in cell type proportions in single cell data. *Bioinformatics*. **38**, 4720–4726 (2022).

87. P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, SciPy 1.0 Contributors, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*. **17**, 261–272 (2020).

88. S. Seabold, J. Perktold, in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010), *Proceedings of the python in science conference*, pp. 92–96.

89. V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, L. C. Groop, PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).

90. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. **102**, 15545–15550 (2005).

91. G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. **16**, 284–287 (2012).

92. G. Yu, L.-G. Wang, G.-R. Yan, Q.-Y. He, DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*. **31**, 608–609 (2015).

93. D. V. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. Warwick Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, W. Dampier, C. Dessimoz, P. Flick, H. Tang, GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8**, 10872 (2018).

94. A. Taylor-Weiner, F. Aguet, N. J. Haradhvala, S. Gosai, S. Anand, J. Kim, K. Ardlie, E. M. Van Allen, G. Getz, Scaling computational genomics to millions of individuals with GPUs. *Genome Biol.* **20**, 228 (2019).

95. J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, J. D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. **28**, 882–883 (2012).

96. J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*. **100**, 9440–9445 (2003).

97. J. D. Storey, A. J. Bass, A. Dabney, D. Robinson, qvalue: Q-value estimation for false

discovery rate control (2020).

98. J. R. Davis, L. Fresard, D. A. Knowles, M. Pala, C. D. Bustamante, A. Battle, S. B. Montgomery, An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *Am. J. Hum. Genet.* **98**, 216–224 (2016).

99. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library. *arXiv* (2019), doi:10.48550/arxiv.1912.01703.

100. F. Privé, H. Aschard, A. Ziyatdinov, M. G. B. Blum, Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*. **34**, 2781–2787 (2017).

101. Z. Gu, D. Hübschmann, rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics*. **39** (2023), doi:10.1093/bioinformatics/btac745.

102. S. Lee, D. Cook, M. Lawrence, plyranges: a grammar of genomic data transformation. *Genome Biol.* **20**, 4 (2019).

103. R. G. Cavalcante, M. A. Sartor, annotatr: genomic regions in context. *Bioinformatics*. **33**, 2381–2383 (2017).

104. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. **32**, 2847–2849 (2016).

105. Z. Gu, L. Gu, R. Eils, M. Schlesner, B. Brors, circlize Implements and enhances circular visualization in R. *Bioinformatics*. **30**, 2811–2812 (2014).

106. H. Wickham, *ggplot2 - Elegant Graphics for Data Analysis* (Springer International Publishing, Cham, ed. 2nd, 2016).

107. A. Kassambara, ggpubr: “ggplot2” Based Publication Ready Plots (2020).

108. T. Wei, V. Simko, R package “corrplot”: Visualization of a Correlation Matrix (2021).