The Genetic Heterogeneity of Multimodal Human Brain Age

Junhao Wen[1,2,*], Bingxin Zhao[3], Zhijian Yang[1], Guray Erus[1], Ioanna Skampardoni[1], Elizabeth Mamourian[1], Yuhan Cui[1], Gyujoon Hwang[1], Jingxuan Bao[4], Aleix Boquet-Pujadas[5], Zhen Zhou[1], Yogasudha Veturi[6], Marylyn D. Ritchie[7] , Paul M. Thompson[8], Haochang Shou[1], Li Shen[4], Arthur W. Toga[9], Christos Davatzikos[1]

[1]Artificial Intelligence in Biomedical Imaging Laboratory (AIBIL), Center for AI and Data Science for Integrated Diagnostics (AI[2]D), Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.
[2]Laboratory of AI and Biomedical Science (LABS), Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Marina del Rey, California, USA.
[3]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA, USA
[4]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, USA
[5]Biomedical Imaging Group, EPFL, Lausanne, Switzerland
[6]Department of Biobehavioral Health and Statistics, Penn State University, University Park, PA, USA
[7]Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[8]Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Marina del Rey, California
[9]Laboratory of Neuro Imaging, Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Marina del Rey, California, USA

[*]Corresponding author:
Junhao Wen, junhao.wen89@gmail.com
3700 Hamilton Walk, 7th Floor, Philadelphia, PA 19104

Word counts: 4514 words

Abstract

The complex biological mechanisms underlying human brain aging remain incompletely understood. To investigate this, we utilized multimodal magnetic resonance imaging and artificial intelligence (AI) to examine the genetic heterogeneity of the brain age gap (BAG) derived from gray matter volume (GM-BAG), white matter tract (WM-BAG), and functional connectivity (FC-BAG). Sixteen significant genomic loci were identified, with GM-BAG loci showing abundant associations with neurodegenerative and neuropsychiatric traits, WM-BAG for cancer and Alzheimer's disease (AD), and FC-BAG for only insomnia. The gene-drug-disease network further corroborated these associations by highlighting genes linked to GM-BAG for the treatment of neurodegenerative and neuropsychiatric disorders, and WM-BAG genes for cancer therapy. GM-BAG showed the highest enrichment of heritability in conserved regions, while in WM-BAG, the 5' untranslated regions exhibited the highest heritability enrichment; oligodendrocytes and astrocytes showed significant heritability enrichment in WM and FC-BAG, respectively. Notably, Mendelian randomization identified risk causal effects of triglyceride-to-lipid ratio in VLDL and type 2 diabetes on GM-BAG, and AD on WM-BAG. These findings suggest that interventions targeting these factors and diseases may ameliorate human brain health. Overall, our results provide valuable insights into the genetic heterogeneity of human brain aging, with potential implications for lifestyle and therapeutic interventions.

Main

The pursuit of precision medicine has involved more than a century-long effort to comprehend the human brain – the most enigmatic and intricate human organ[1]. The advent of artificial intelligence (AI) has made significant progress in its ability to decipher various aspects of human brain health[2,3], such as normal brain aging[4], neurodegenerative disorders like Alzheimer's disease (AD)[5], and brain cancer[6]. Utilizing magnetic resonance imaging (MRI), the AI-derived human brain age[7–9] has emerged as a valuable biomarker for the evaluation of brain health. More precisely, the difference between an individual's predicted brain age and their chronological age – brain age gap (BAG) – provides a means of quantifying an individual's brain health by measuring deviation from the typical aging trajectory. This deviation can be caused by adverse factors, such as AD[10], leading to an accelerated brain age (i.e., a positive BAG), or protective factors, such as positive physical activity or education, resulting in a decelerated brain age (i.e., a negative BAG).

Brain imaging genomics[11], an emerging scientific field that combines computational statistics and AI, uses imaging-derived phenotypes (IDP) from MRI and genetic variants to shed light on the mechanistic and pathological insights of the human brain. Recent large-scale genome-wide association studies (GWAS)[12–19] have identified a diverse set of genomic loci linked to gray matter (GM)-IDP from T1-weighted MRI, white matter (WM)-IDP from diffusion MRI for, and functional connectivity (FC)-IDP from functional MRI. While previous GWAS[20] have associated BAG with common genetic variants (e.g., SNP), they primarily focused on GM-BAG[10,21]. Identifying diverse genetic factors underlying multimodal human brain age and pinpointing their causal consequences are vital for developing gene-inspired therapeutic interventions, because numerous risk/protective lifestyle and environmental factors, and

neurobiological processes may exert independent, synergistic, antagonistic, sequential, or differential influences on human brain health. In this context, we postulate that GM, WM, and FC-BAG can serve as robust and complementary endophenotypes[22] – close to underlying etiology – for digitizing multiple facets of human brain health.

The present study sought to uncover the genetic heterogeneity of multimodal BAG and explore the causal relationships between protective/risk factors and decelerated/accelerated brain age. To accomplish this, we analyzed multimodal MRI scans from 42,089 participants from the UK Biobank study[23] and obtained 119 GM-IDP, 48 fractional anisotropy (FA) WM-IDP, and 210 FC-IDP (See **Method 1**). We first objectively compared the age prediction performance of four machine learning models using these IDPs. We then conducted GWAS to identify genomic loci associated with GM, WM, and FC-BAG, constructed a gene-drug-disease network, and estimated genetic correlation with several clinical traits, and heritability enrichment in various functional categories or specific cell/tissue types. Finally, we performed Mendelian Randomization (MR)[24] to infer causal effects between several clinical traits and multimodal BAG (i.e., GM, WM, and FC-BAG).

Results

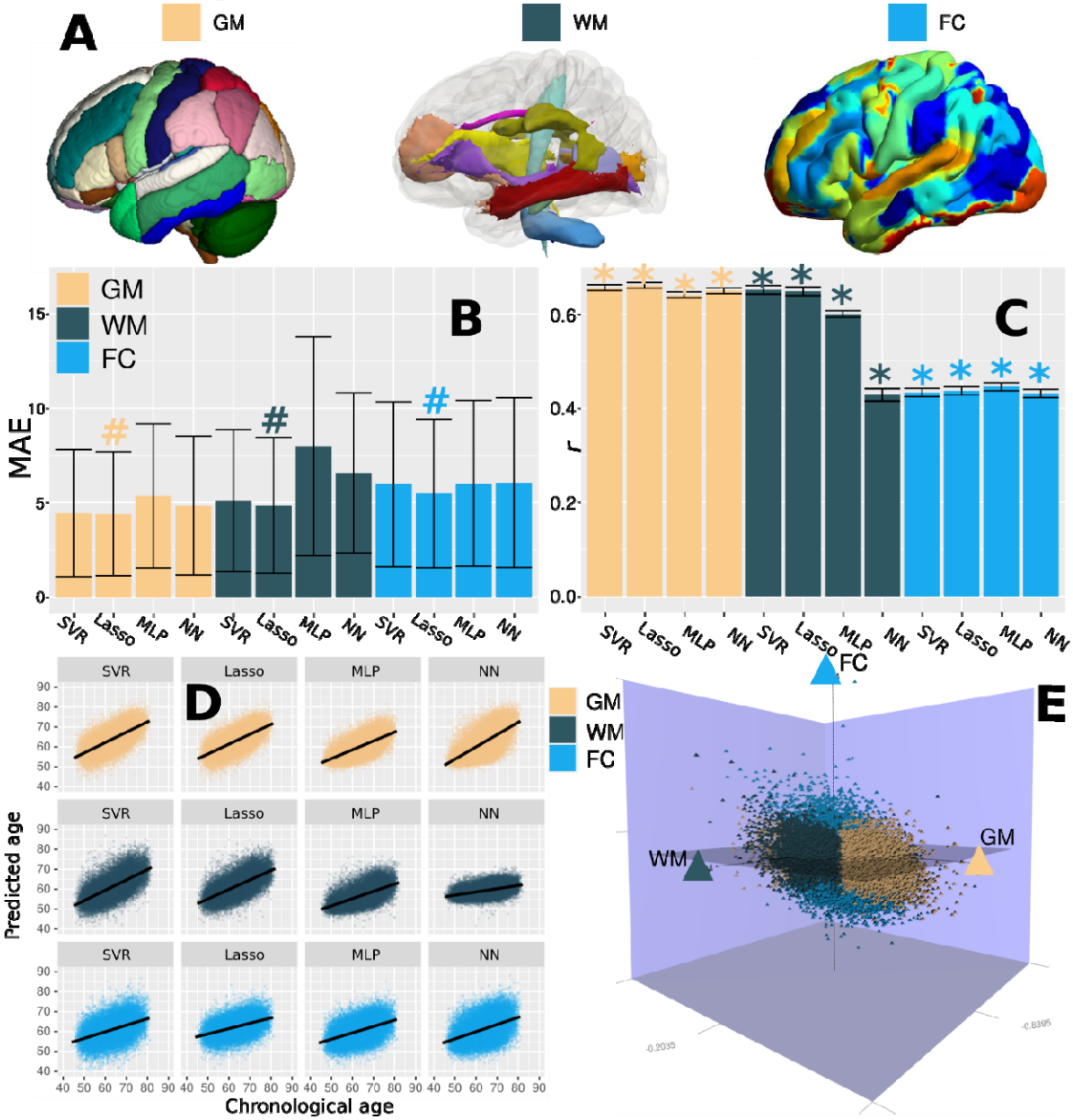
Brain age prediction using multimodal imaging-derived phenotypes

To objectively evaluate the performance of four machine learning methods in brain age prediction based on GM, WM, and FC-IDP (**Fig. 1A**), we employed a nested cross-validation (CV) procedure[25,26] and used an independent test dataset – not touched until we finalized the models during training and validation (**Method 1**). The four machine learning models included support vector regression (SVR), lasso regression, multilayer perceptron (MLP), and a five-layer neural network (i.e., three linear layers and one Relu layer; thereafter NN)[27] (**Method 2**). Other studies have undertaken a more thorough evaluation of a broad range of machine learning models for predicting brain age[28].

Several findings were observed based on the results from the independent test dataset. First, GM-IDP (4.39<MAE<5.35; 0.64<r<0.66), WM (4.92<MAE<7.95; 0.42<r<0.65), and FC-IDP (5.48<MAE<6.05; 0.43 <r<0.46) achieved gradually higher mean absolute errors (MAE) and smaller Pearson's correlation (r) (**Fig. 1B**, **C**, and **D**). Second, lasso regression always obtained the lowest MAE for GM, WM, and FC-IDP, and linear models obtained a lower MAE than non-linear networks (**Fig. 1B**). Third, all models generalized well from the training/validation/test dataset to the independent test dataset. However, simultaneously incorporating WM-IDP from FA, mean diffusivity (MD), neurite density index (NDI), and orientation dispersion index (ODI) resulted in a severe overfitting phenomenon. (**Supplementary eTable 1** and **Supplementary eFigure 1**). The observed overfitting may be attributed to the substantially large number of parameters (N=38,364) in the network, as well as the strong correlations among the diffusion metrics (i.e., FA, MD, ODI, and NDI). Fourth, the experiments stratified by sex did not exhibit substantial differences, except for a stronger

overfitting tendency observed in females compared to males using WM-IDP incorporating the four diffusion metrics (**Supplementary eTable 1**). Finally, in this independent dataset, considerable heterogeneity existed in the expression of GM, WM, and FC-BAG (**Fig. 1E**). The detailed results of the CV procedure, including the training, validation, and test performance, as well as sex-stratified experiments, are presented in **Supplementary eTable 1**. Taken together, these results propelled us to investigate the underlying genetic determinants responsible for this phenotypic heterogeneity (**Fig. 1E**). In all subsequent genetic analyses, we reported the results using BAG derived from the three lasso models with the lowest MAE in each modality (**Fig. 1A**), with "age bias" corrected as in De Lange et al.[29].

Figure 1: **Brain age prediction using multimodal MRI and four machine learning models**



A) Multimodal brain MRI data were used to derive imaging-derived phenotypes (IDP) for T1-weighted MRI (119 GM-IDP), diffusion MRI (48 WM-IDP), and resting-state functional MRI (210 FC-IDP). IDPs for each modality were indicated using different colors based on predefined brain atlases or ICA for FC-IDP. **B**) Linear models achieved lower mean absolute errors (MAE) than non-linear models using support vector regression (SVR), lasso regression, multilayer perceptron (MLP), and a five-layer neural network (NN). The MAE for the independent test dataset is presented, and the model with the lowest MAE for each modality is indicated by the # symbol. Error bars represent standard deviation (SD). **C**) Pearson's correlation (*r*) between the predicted brain age and chronological age was computed, and statistical significance (P-value<0.05) after adjustment for multiple comparisons using the FDR method was denoted by the * symbol. Error bars represent the 95% confidence interval (CI). **D**) A scatter plot for the predicted brain age and the chronological age. **E**) Participants from the independent test dataset

showed considerable heterogeneity in the manifestation of GM, WM, and FC-BAG. The colored triangles represent the absolute maximum magnitude of the multimodal BAG for each participant. The presented brain age was corrected for "age bias" from the three best-MAE models (#); genetic analyses were reported using the corrected BAG from the three best-MAE models.

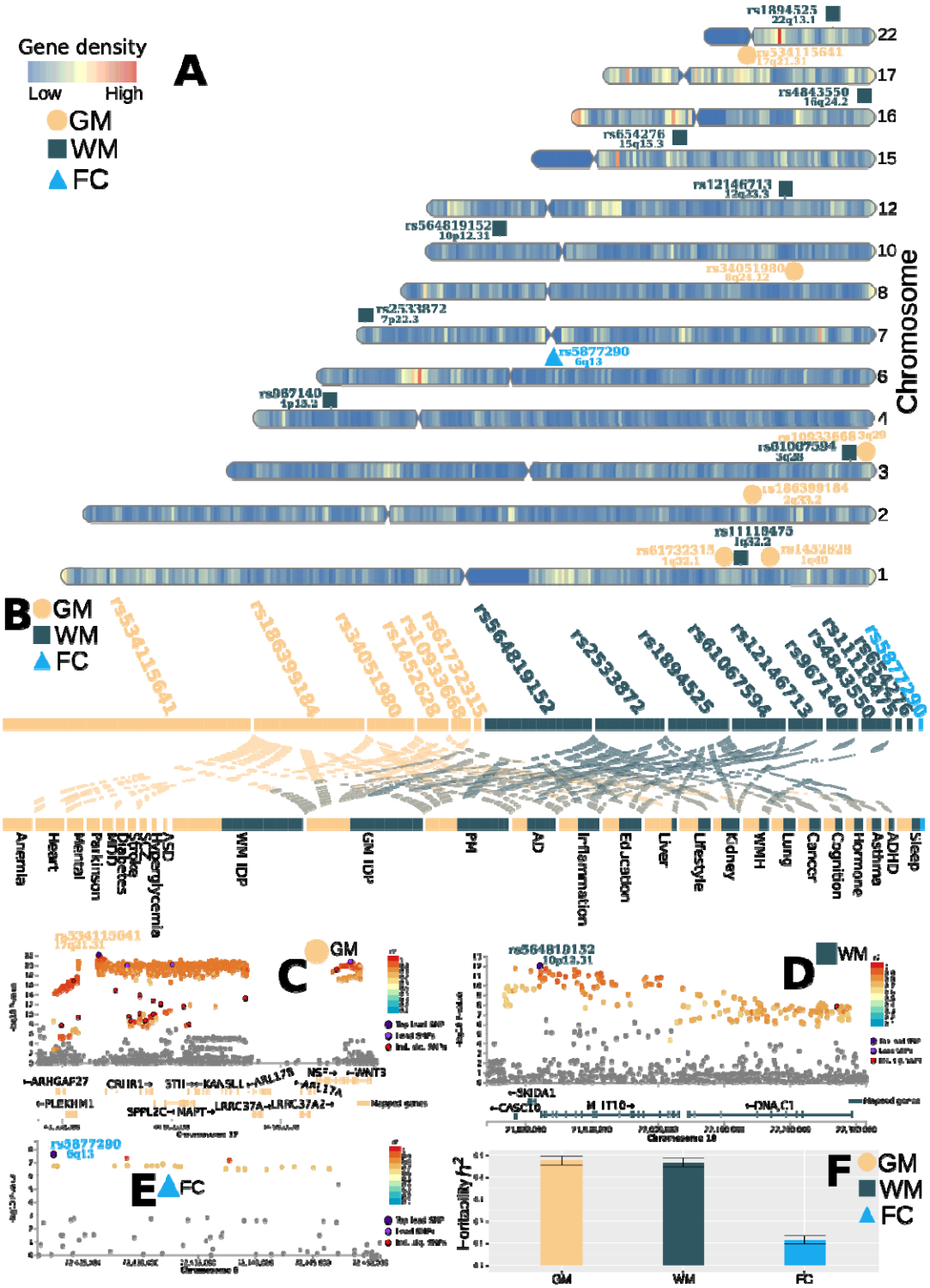
Genomic loci associated with multimodal brain age gaps

GWAS (**Method 4A**) revealed 6, 9, and 1 genomic locus linked to GM, WM, and FC-BAG, respectively (**Fig. 2A**, **Supplementary eTable 2**). All multimodal BAGs were highly heritable (**Fig. 2F**). We performed a phenome-wide association query for these independent variants within each locus (**Method 4C**). Notably, the independent SNPs within each locus were frequently linked to other traits previously reported in the literature. Specifically, we found that the GM-BAG loci were uniquely associated with a range of neuropsychiatric disorders like Major Depressive Disorder (MDD), as well as heart disease, such as cardiovascular disease. Additionally, we observed associations between these loci and other diseases (such as anemia), as well as biomarkers from various human organs (e.g., liver) (**Fig. 2B**). As an illustration, within a specific genomic locus (top lead SNP: rs534115641, chr17:43667537:C>T, P-value=8.44×10^{-23}), the independent significant SNPs were predominantly linked with traits related to AD (rs7207400[30], rs199499[30]), as well as traits related to mental health (rs9303521 and rs7225384 for depressed affect[31]). One of the mapped genes, namely *MAPT*, has a well-established association with increased susceptibility to AD[32]. The loci for WM-BAG were also found to be associated with physical measures (e.g., BMI), cancer (e.g., brain cancer), and AD-related biomarkers (e.g., tau protein levels, **Fig. 2B**). Finally, the novel genomic locus associated with FC-BAG was only associated with insomnia (**Supplementary eFile 1**).

We then performed positional and functional annotations to map SNPs to genes associated with GM and WM-BAG loci (**Method 4B**). A genomic locus (top lead SNP:

rs534115641, **Fig. 2C**) was identified for GM-BAG that contains multiple protein-encoding genes annotated by position, eQTL, and chromatin interaction. The *NSF* gene, which encodes N-ethylmaleimide-sensitive fusion proteins, plays a key role in the transfer of membrane vesicles between cellular compartments. This gene has been linked to several conditions, including Parkinson's disease (PD)[33], epithelial ovarian cancer[34], cognitive traits[35], and fibromuscular dysplasia[36]. The *CRHR1* gene encodes a G protein-coupled receptor which specifically binds to neuropeptides of the corticotropin-releasing hormone family. These neuropeptides are recognized as key regulators of the hypothalamic-pituitary-adrenal pathway. A prior GWAS[37] demonstrated the association of this gene with environmental stress, providing evidence for the intimate involvement of the hypothalamic-pituitary-adrenal axis, the central nervous system, and the endocrine system in stress response[38]. We also identified a highly polygenic genomic locus (top lead SNP: rs564819152, **Fig. 2D**) for WM-BAG, mapped to the *SKIDA1*, *CASC10*, *MLLT10*, and *DNAJC1* genes – all implicated in various types of cancer. In contrast, the FC-BAG locus was novel and did not map to any genes. All mapped genes for GM and WM-BAG are presented in **Supplementary eTable 2**. Taken together, the broad range of associations from our results indicate potential genetic overlaps and causalities between the multimodal BAG and other clinical traits.

Figure 2: **Genome-wide associations of multimodal brain age gaps**



A) Genome-wide associations identified sixteen genomic loci associated with GM (6), WM (9), and FC-BAG (1) using a genome-wide P-value threshold [$-\log_{10}$(P-value) > 7.30]. Each locus is represented by the top lead SNP and cytogenetic region. **B**) Phenome-wide association query from GWAS Catalog[39]. Independent significant SNPs inside each locus were largely associated

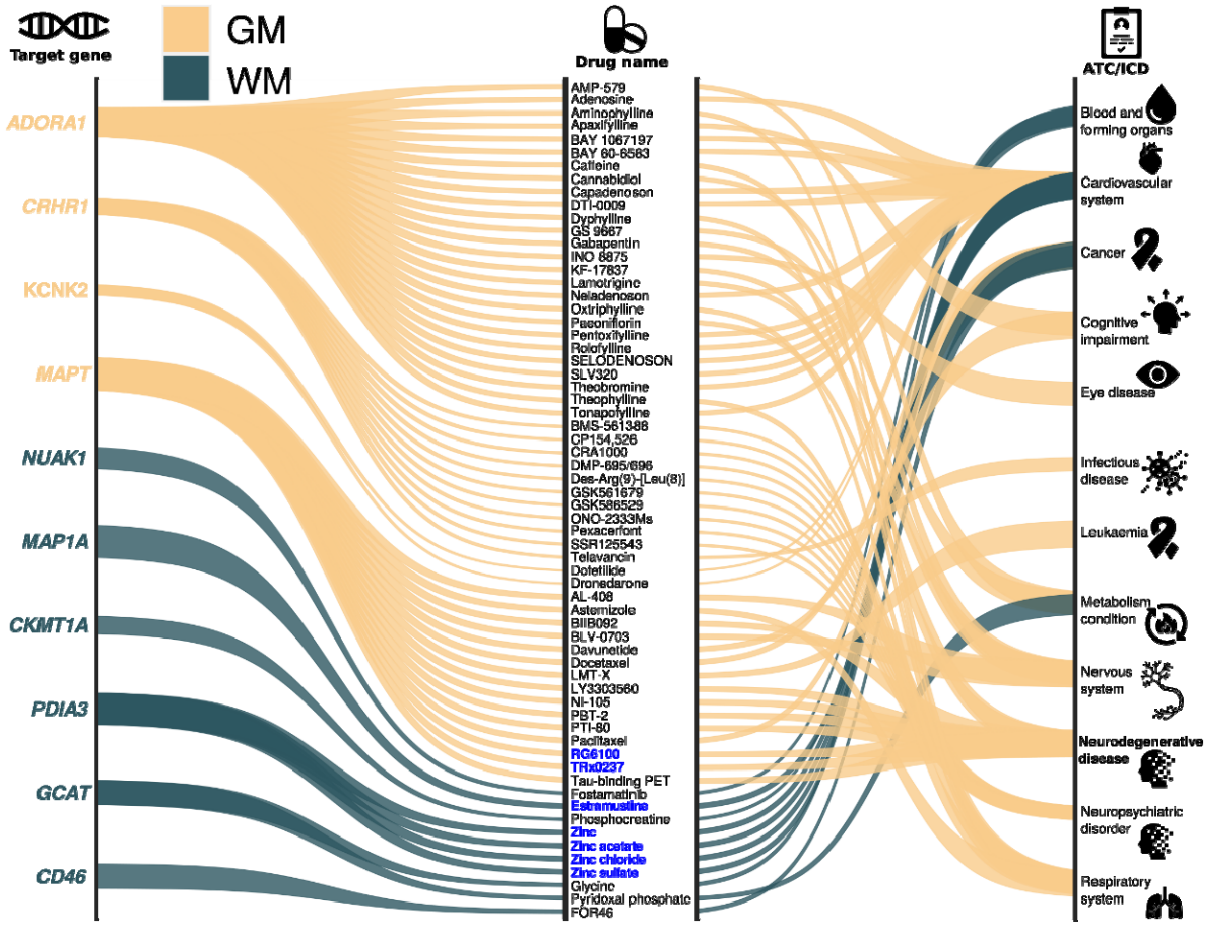
with many traits, which were further classified into many trait categories, including biomarkers from multiple body organs (e.g., heart and liver), neurological disorders (e.g., Alzheimer's disease and Parkinson's disease), and lifestyle risk factors (e.g., alcohol consumption). **C**) Regional plot for a genomic locus associated with GM-BAG. Color-coded SNPs are decided based on their highest r^2 to one of the nearby independent significant SNPs. Gray-colored SNPs are below the r^2 threshold. The top lead SNP, lead SNPs, and independent significant SNPs are denoted as dark-purple, purple, and red, respectively. Mapped, orange-colored genes of the genomic locus were annotated by positional mapping, eQTL mapping, and chromatin interaction mapping (**Method 4B**). **D**) Regional plot for a genomic locus associated with WM-BAG. **E**) The novel genomic locus associated with FC-BAG did not map to any genes. We used the Genome Reference Consortium Human Build 37 (GRCh37) in all genetic analyses. **F**) The SNP-based heritability estimates[40] (h^2) are shown for GM, WM, and FC-BAG. Abbreviation: AD: Alzheimer's Disease; ASD: Autism Spectrum Disorder; PD: Parkinson's Disease; ADHD: Attention-Deficit/Hyperactivity Disorder.

Gene-drug-disease network of multimodal brain age gaps

We investigated the potential "druggable genes[41]" from the mapped genes by constructing a gene-drug-disease network (**Method 4F**). The network connects genes with drug-like molecules or drugs targeting specific diseases that are currently active at any stage of clinical trials. Our analysis revealed that 4 and 6 mapped genes, associated with GM-BAG and WM-BAG, respectively. The GM-BAG genes were linked to clinical trials for treating heart disease, neurodegenerative, neuropsychiatric, and neurodevelopmental diseases, as well as respiratory diseases. On the other hand, the WM-BAG genes were primarily targeted for various cancer treatments and cardiovascular diseases (**Fig. 3**). To illustrate, for the GM-BAG *MAPT* gene, several drugs or drug-like molecules are currently being evaluated for treating Alzheimer's disease (AD). For example, Semorinemab (RG6100), an anti-tau IgG4 antibody, was being investigated in a phase-2 clinical trial (trial number: NCT03828747), which targets extracellular tau in AD, to reduce microglial activation and inflammatory responses[42]. Another drug is LMTM (TRx0237), which is a second-generation tau protein aggregation inhibitor and is being tested in a phase-3 clinical trial (trial number: NCT03446001) for the treatment of AD and frontotemporal

dementia[43]. In contrast, we observed that the WM-BAG genes primarily bind with drugs for the treatment of cancer and cardiovascular diseases. For instance, the *PDIA3* gene, which is associated with the folding and oxidation of proteins, has been targeted for the development of several zinc-related drugs approved by the FDA for treating cardiovascular diseases. Another example is the *MAP1A* gene, which encodes microtubule-associated protein 1A, and is linked to the development of estramustine, an FDA-approved drug for treating prostate cancer (Fig. 3). The detailed results are presented in **Supplementary eFile 2**.

Figure 3: **Gene-drug-disease network of multimodal brain age gaps**



The gene-drug-disease network derived from the mapped genes revealed a broad spectrum of targeted diseases and cancer, spanning brain cancer, cardiovascular system, Alzheimer's disease, and obstructive airway disease, among others. The thickness of the lines represented by the brain tissue-specific gene set enrichment results using the GTEx v8 dataset. We highlight several drugs under blue-colored, bold text for illustration. Abbreviation: ATC: Anatomical Therapeutic Chemical; ICD: International Classification of Diseases.

Genetic correlation with other clinical traits

To examine genetic covariance between multimodal BAG and other clinical traits, we calculated the genetic correlation (r_g) (**Method 4D**) by utilizing GWAS summary statistics from 16 clinical traits. These traits encompassed common brain diseases and their AI-derived disease subtypes, as well as education and intelligence (**Fig. 4A** and **Supplementary eTable 3**).

Our analysis revealed significant genetic correlations between GM-BAG and disease subtypes of AD (AD1[5]), ASD (ASD1 and ASD3[44]), schizophrenia (SCZ1[45]), and Obsessive-Compulsive Disorder (OCD)[46]; WM-BAG and AD1, ASD1, SCZ1, and SCZ2; and FC-BAG and education[47] and SCZ1. The detailed results for r_g estimates are presented in **Supplementary eTable 4**. These results highlight the disease subtypes, rather than the "1-for-all" unitary disease diagnosis, as robust endophenotypes[22] to re-test relevant hypotheses in future studies for patient stratification and personalized medicine.

Partitioned heritability enrichment in annotated functional categories

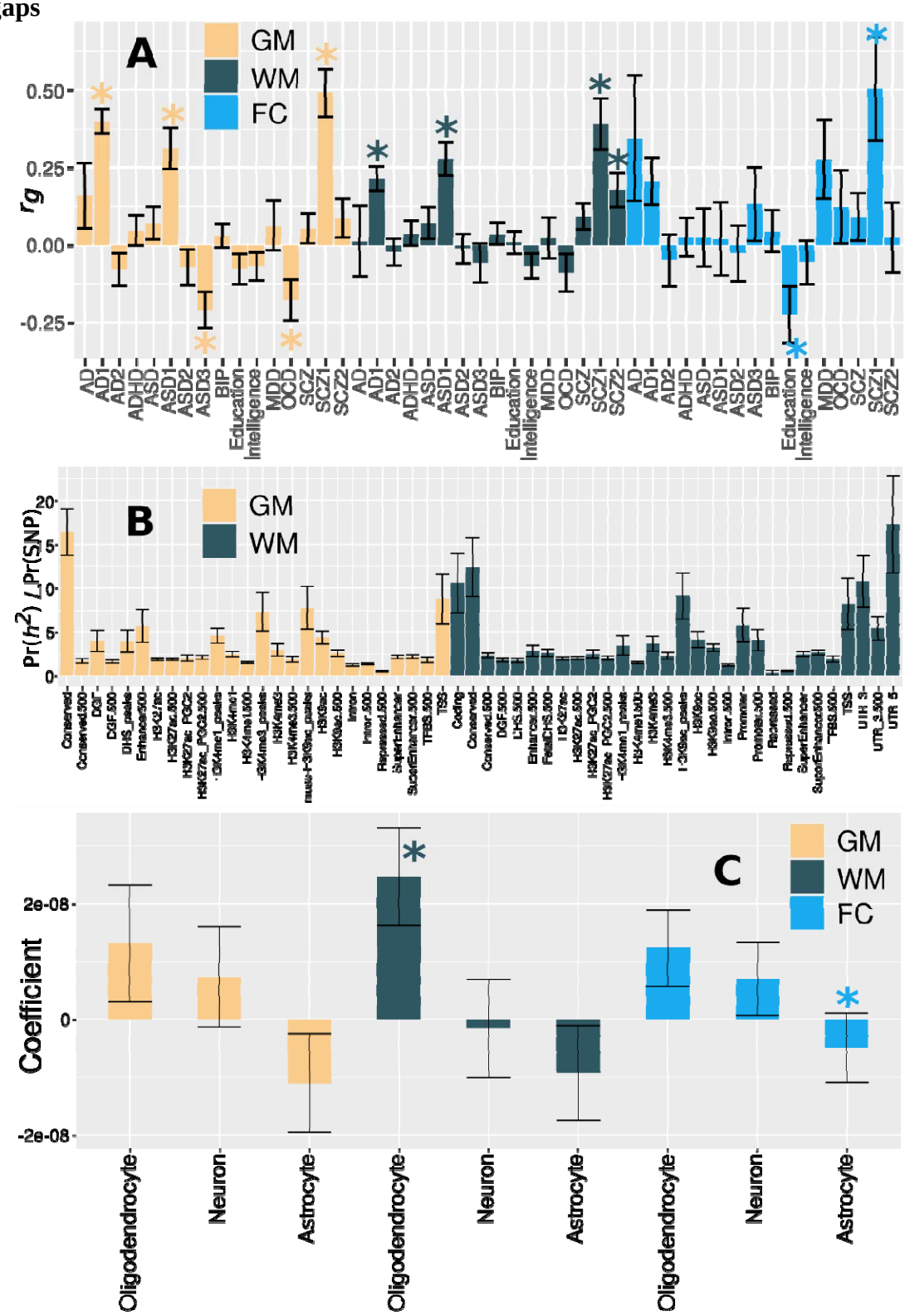
To investigate the enrichment of heritability of genetic variants related to multimodal BAG in the 53 functional categories, we conducted a partitioned heritability analysis[48] (**Method 4E**). Our results revealed that GM and WM-BAG exhibited significant enrichment of heritability across numerous annotated functional categories. Specifically, some categories displayed greater enrichment than others, and we have outlined some in further detail.

For GM-BAG, the regions that are conserved across mammals, as indicated by the label "conserved" in **Fig. 4B**, displayed the most notable enrichment of heritability: approximately 2.61% of SNPs were found to explain 0.43±0.07 of SNP heritability (P-value=5.80×10^{-8}). Additionally, transcription start site (TSS)[49] regions employed 1.82% of SNPs to explain 0.16±0.05 of SNP heritability (P-value=8.05×10^{-3}). TSS initiates the transcription at the 5' end of a gene and is typically embedded within a core promoter that is crucial to the transcription machinery[50]. The heritability enrichment of Histone H3 at lysine 4, as denoted for "H3K4me3_peaks" in **Fig. 4B**, and histone H3 at lysine 9 (H3K9ac)[51] were also found to be large and were known to highlight active gene promoters[52]. For WM-BAG, 5' untranslated

regions (UTR) used 0.54% of SNPs to explain 0.09±0.03 of SNP heritability (P-value=4.24×10^{-3}). The 5' UTR is a crucial region of a messenger RNA (mRNA) located upstream of the initiation codon. It plays a pivotal role in the regulation of transcript translation, with varying mechanisms in viruses, prokaryotes, and eukaryotes.

To further explore the heritability enrichment of multimodal BAG in specific cell types and brain tissues, we conducted an additional analysis (**Fig. 4C**). WM-BAG (P-value=1.69×10^{-3}) exhibited significant heritability enrichment in oligodendrocytes, one type of neuroglial cells. FC-BAG (P-value=1.12×10^{-2}) showed such enrichment in astrocytes, the most prevalent glial cells in the brain. GM-BAG showed no enrichment in any of these cells. Our findings are consistent with the understanding of the molecular and biological characteristics of GM and WM. That is, oligodendrocytes are primarily responsible for forming the lipid-rich myelin structure, whereas astrocytes play a crucial role in various cerebral functions, such as development and homeostasis. In addition, a previous GWAS[14] on WM-IDP has also identified considerable heritability enrichment in glial cells, especially oligodendrocytes. The detailed results for the 53 functional categories and cell-specific analyses are detailed in **Supplementary eTable 5**.

Figure 4: **Genetic correlation and partitioned heritability enrichment of multimodal brain age gaps**



A) Genetic correlation (r_g) between GM, WM, and FC-BAG and 16 clinical traits, including neurodegenerative diseases (e.g., Alzheimer's disease) and their AI-defined subtypes (e.g., AD1

and AD2[5]), neuropsychiatric disorders (e.g., ASD) and their subtypes (ASD1, 2, and 3[44]), intelligence, and education. **B**) The proportion of heritability enrichment for 53 functional categories[48]. For visualization purposes, we only show the functional categories that survived the correction for multiple comparisons using the FDR method. **C**) Cell type-specific partitioned heritability estimates. We included gene sets from Cahoy et al.[53] for three main cell types (i.e., astrocyte, neuron, and oligodendrocyte). * denotes statistical significance (P-value<0.05) after adjusting for multiple comparisons using the FDR method. Error bars represent the standard error of the estimated parameters. Abbreviation: AD: Alzheimer's Disease; ADHD: Attention-Deficit/Hyperactivity Disorder; ASD: Autism Spectrum Disorder; BIP: Bipolar Disorder; MDD: Major Depressive Disorder; OCD: Obsessive-Compulsive Disorder; SCZ: schizophrenia.

Mendelian randomization for causal inference

We investigated the potential causal effects of several risk factors (i.e., exposure variable) on multimodal BAG (i.e., outcome variable) using a bidirectional two-sample MR approach[54] (**Method 4G**). We hypothesized that several diseases and lifestyle risk factors may contribute to the acceleration or deceleration of human brain aging.

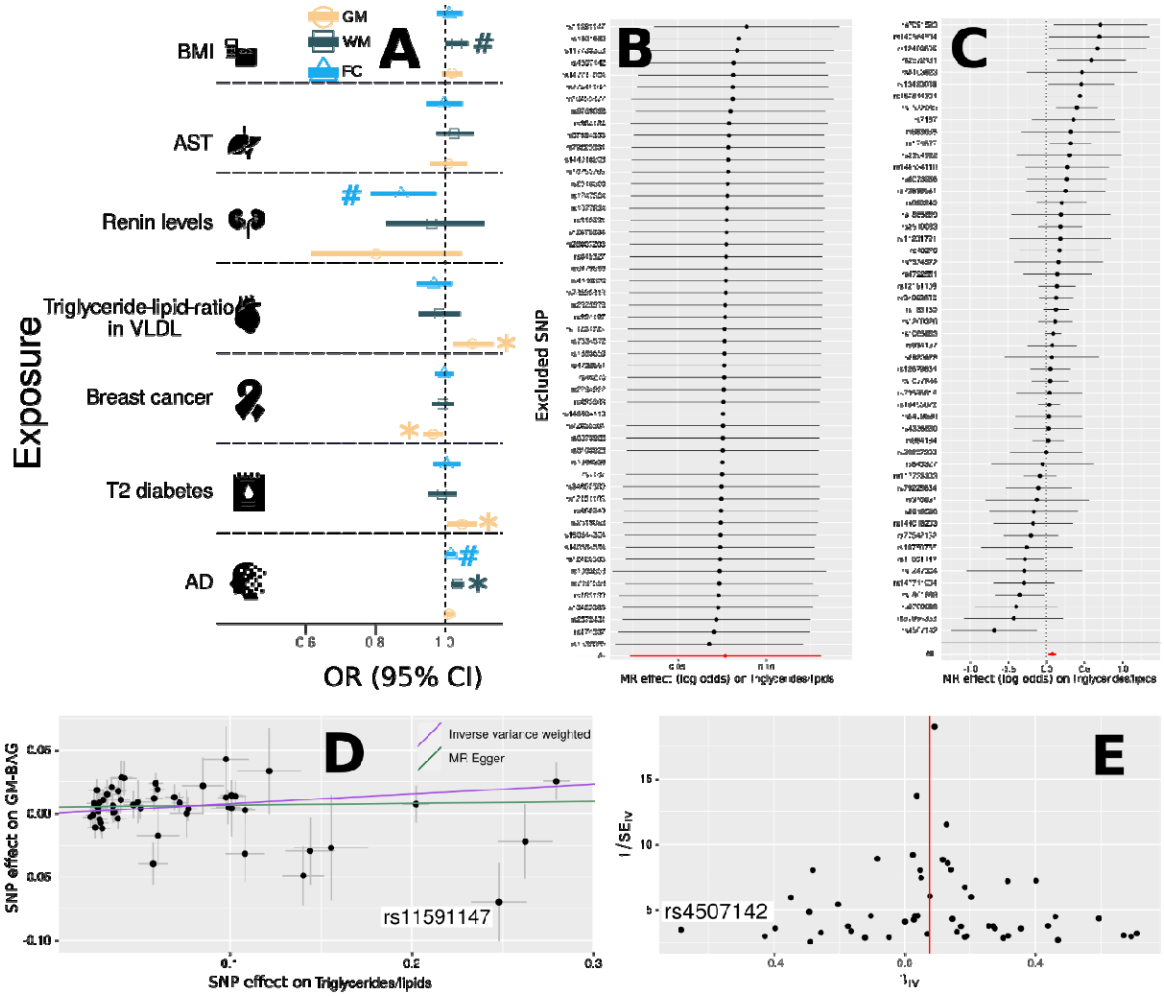
We found putative causal effects of triglyceride-to-lipid ratio in very large very-low-density lipoprotein (VLDL)[55] [risk effect for per unit increase; P-value=5.09×10^{-3}, OR (95% CI) = 1.08 (1.02, 1.13), number of SNPs=52], type 2 diabetes[56] [risk effect; P-value=1.96×10^{-2}, OR (95% CI) = 1.05 (1.01, 1.09), number of SNPs=10], and breast cancer[57] [protect effect; P-value=1.81×10^{-2}, OR (95% CI) = 0.96 (0.93, 0.99), number of SNPs=118] on GM-BAG (i.e., accelerated brain age). We also identified causal effects of AD [risk effect for per unit increase; P-value=7.18×10^{-5}, OR (95% CI) = 1.04 (1.02, 1.05), number of SNPs=13] on WM-BAG (**Fig. 5A**). We subsequently examined the potential inverse causal effects of multimodal BAG (i.e., exposure) on these risk factors (i.e., outcome). However, owing to the restricted power (number of instrumental variables < 6), we did not observe any significant signals (**Supplementary eFigure 2** and **Supplementary eFile 3**).

We performed sensitivity analyses to investigate potential violations of the three IV assumptions (**Method 4G**). To illustrate this, we showed the sensitivity analysis results for the causal effect of the triglyceride-to-lipid in VLDL ratio on GAM-BAG (**Fig. 5B-E**). In a leave-one-out analysis, we found that no single SNP overwhelmingly drove the overall effect (**Fig. 5B**). There was evidence for the presence of slight heterogeneity[58] of the causal effect amongst SNPs (Cochran's Q value=76.06, P-value=5.09×10^{-3}). Some SNPs exerted opposite causal effects compared to the model using all SNPs (**Fig. 5C**). The scatter plot (**Fig. 5D**) indicated one obvious SNP outlier (rs11591147), and the funnel plot showed little asymmetry with only an outlier denoted in **Fig. 5E** (rs4507142). Finally, the MR Egger estimator allows for pleiotropic effects independent of the effect on the exposure of interest (i.e., the InSIDE assumption[59]). Our results showed a small positive intercept ($5.21 \times 10^{-3} \pm 2.87 \times 10^{-3}$, P-value=0.07) and a lower OR [IVW: 1.08 (1.02, 1.13) vs. Egger: 1.01 (0.93, 1.10)], which may indicate the presence of directional horizontal pleiotropy for some SNPs. We present the sensitive check results for other significant exposure variables in **Supplementary eFigure 3**.

To investigate the potential directional pleiotropic effects, we re-analyzed the MR Egger regression by excluding two outliers identified in **Fig. 5D** (rs11591147) and **E** (rs4507142), which led to a slightly increased OR [1.04 (0.96, 1.12)] and a smaller positive intercept of ($4.41 \times 10^{-3} \pm 2.65 \times 10^{-3}$, P-value=0.09). Our findings provide support for the notion that these two outlier SNPs may have a directional pleiotropic effect on GM-BAG. Nevertheless, given the complex nature of brain aging, it is plausible that various other biological pathways may also contribute to human brain aging. For example, the SNP (rs11591147) was largely associated with other blood lipids, such as LDL cholesterol[60], and heart diseases, such as coronary artery disease[61]. Based on our sensitivity analysis and the consistency of the results obtained from all

five MR methods (**Supplementary eFile 3**), we showed evidence for a putative causal effect of the triglyceride-to-lipid ratio in VLDL on GM-BAG.

Figure 5: **Causal inference for 7 traits on multimodal brain age gaps**



A) The causal inference was performed using a two-sample Mendelian Randomization (MR, **Method 4G**) approach for 7 selected exposure variables on three outcome variables (i.e., GM, WM, and FC-BAG). * denotes statistical significance after correcting for multiple comparisons using the FDR method (*N*=7); # denotes the tests passing the nominal significance threshold (P-value=0.05) but did not survive the multiple comparisons. Odds Ratio (OR) and the 95% confidence interval (CI) are presented. **B**) Leave-one-out analysis of the triglyceride-to-lipid ratio on GM-BAG. Each row represents the MR effect (log OR) and the 95% CI by excluding that SNP from the analysis. The red line depicts the IVW estimator using all SNPs. **C**) A forest plot for the single-SNP MR results. Each line represents the MR effect (log OR) for the triglyceride-to-lipid ratio on GM-BAG using only each single SNP; the red line shows the MR effect using all SNPs together. **D**) A scatter plot for the MR effect sizes of the SNP-triglyceride-to-lipid ratio association (x-axis, SD units) and the SNP-GM-BAG associations (y-axis, log OR) with standard error bars. The slopes of the purple and green lines correspond to the causal effect estimated by the IVW estimator and the MR Egger estimator, respectively. We annotated a potential outlier. **E**) A funnel plot showing the relationship between the causal effect of the triglyceride-to-lipid ratio on GM-BAG. Each dot represents MR effect sizes estimated using each SNP as a separate instrument against the inverse of the standard error of the causal estimate. The vertical red line

shows the MR estimates using all SNPs. We annotated a potential outlier. Abbreviation: AD: Alzheimer's Disease; AST: Aspartate Aminotransferase; BMI: Body Mass Index; VLDL: Very Low-Density Lipoprotein; CI: Confidence Interval; OR: Odds Ratio; SD: Standard Deviation; SE: Standard Error.

Discussion

The current study utilizes multimodal MRI and AI to elucidate the genetic architecture of multimodal BAG, shedding light on the genome-wide influence of common genetic variants on human brain aging. Our results unveil the heterogeneity of this intricate biological process and its multifaceted interplay – via associations and causality – with other health conditions and clinical phenotypes. Digitizing multimodal human brain age and deciphering its heterogeneity can aid in patient management, gene-inspired drug development, and potential lifestyle changes and interventions.

Genetic architecture of GM-BAG

Our genetic results from GM-BAG substantiate that many diseases, conditions, and clinical phenotypes shared common genetic underpinnings with brain age triggered by macrostructural changes in GM (e.g., brain atrophy). The locus with the highest degree of polygenicity (the top lead SNP rs534114641 at 17q21.31) showed substantial association with the aforementioned traits and was mapped to numerous genes known to be associated with various diseases (**Fig. 2C**). To confirm our findings, this locus was also identified by several previous GM-BAG GWAS[10,62,63]. Among these genes, the *MAPT* gene, known to encode a protein called tau, is a prominent hallmark in AD and is implicated in approximately 30 tauopathies, including progressive supranuclear palsy and frontotemporal lobar degeneration[64]. Our gene-drug-disease network also showed several drugs, such as Semorinemab[42], in active clinical trials currently targeting treatment for AD (**Fig. 3**). The heritability enrichment of GM-BAG was high in several functional categories, with conserved regions being the most prominent. The observed higher heritability enrichment in conserved regions compared to coding regions[65] supports the long-standing hypothesis regarding the functional significance of

conserved sequences, although the precise role of many highly conserved non-coding DNA sequences remains unclear[66]. Remarkably, the genetic correlation results of GM-BAG support that disease subtypes serve as a more robust endophenotypic instrument than the "one-for-all" unitary diagnosis, providing significant clinical implications for re-investigations of brain health and disease[67–71].

The elevated triglyceride-to-lipid ratio in VLDL, which serve as an established biomarker for cardiovascular diseases[72], is causally associated with higher GM-BAG (accelerated brain age). Therefore, lifestyle interventions that target this biomarker might hold promise as an effective strategy to enhance overall brain health. In addition, we revealed that the unit-increased likelihood of type 2 diabetes has a causal effect on increased GM-BAG. Research has shown that normal brain aging is accelerated by approximately 26% in patients with progressive type 2 diabetes compared with healthy controls[73]. The protective causal effect of breast cancer on GM-BAG is intriguing in light of existing literature adversely linking breast cancer to brain metastasis[74] and chemotherapy-induced cognitive impairments, commonly known as "chemo brain".

Genetic architecture of WM-BAG

The genetic architecture of WM-BAG exhibits strong correlations with cancer-related traits, AD, and physical measures such as BMI, among others. Our phenome-wide association results largely confirm the enrichment of these traits in previous literature. In particular, the *DNAJC1* gene, annotated from the most polygenic locus on chromosome 10 (top lead SNP: rs564819152), encodes a protein called heat shock protein 40 (Hsp40) and plays a role in protein folding and the response to cellular stress. This gene has been implicated in various cancer types, such as breast cancer, renal cancer, and melanoma (**Supplementary eFigure 4**). In addition,

several FDA-approved drug drugs have been developed based on these WM-BAG genes for different types of cancer (**Fig. 3**). These findings provide novel insights into the genetic underpinnings of WM-BAG and their potential relevance to cancer.

Remarkably, AD was causally associated with WM-BAG. Our Mendelian randomization analysis confirmed the abundant association evidenced by the phenome-wide associations (**Fig. 2B**). Dementia, such as AD, is undeniably a significant factor that contributes to the decline of the aging brain. There is a growing body of evidence suggesting that Alzheimer's Disease (AD) is not solely a gray matter disease and that significant microstructural changes can be observed in WM before the onset of cognitive decline[75]. We also identified a nominal causal significance of BMI [risk effect; P-value=4.73×10^{-2}, OR (95% CI) = 1.03 (1.00, 1.07)] on WM-BAG. These findings underscore the potential of lifestyle interventions and medications currently being tested in clinical trials for AD to improve brain health.

Genetic architecture of FC-BAG

The genetic signals for FC-BAG were weaker than those observed for GM and WM-BAG, which is consistent with the age prediction performance. A novel genomic locus on chromosome 17q21.31 harbored an independent variant (rs1204329), which was previously linked to insomnia[76]. The top lead SNP, rs5877290, associated with this locus is a delins mutation type; no known association with any human disease or gene mapping has been established for this SNP. The genetic basis of FC-BAG covaries with educational performance and schizophrenia subtypes. Specifically, parental education has been linked to cognitive ability, and researchers have identified a functional connectivity biomarker between the right rostral prefrontal cortex and occipital cortex that mediates the transmission of maternal education to

offspring's performance IQ[77]. On the other hand, schizophrenia is a highly heritable mental disorder that exhibits functional dysconnectivity throughout the brain[78].

In summary, this study offers a comprehensive depiction of the genetic architecture and heterogeneity of multimodal human brain age, potentially opening new avenues for drug repurposing research and identifying modifiable protective and risk factors that can ameliorate human brain health.

Methods

Method 1: Study populations

UKBB is a population-based study of more than 50,000 people recruited between 2006 and 2010 from Great Britain. The current study focused on the imaging-genomics population participants who underwent the MRI scan and genome sequencing (genotype array data and the imputed genotype data). The UKBB study has ethical approval, and the ethics committee is detailed here: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/governance/ethics-advisory-committee. The study design, phenotype and genetic data availability, and quality check have been published and detailed elsewhere[23]. The current work is under application number 35148. **Table 1** shows the study characteristics of the present work.

To objectively train the machine learning model and compare the performance of the multimodal BAG, we defined the following datasets:

- *Training/validation/test dataset*: To objectively compare the age prediction performance of different MRI modalities and machine learning models, we randomly sub-sampled 500 (250 females) participants within a ten-year range from 44 to 84 years old, resulting in the same 4000 participants (excluding all major diseased participants, such as AD, MDD, etc.) for GM, WM, and FC-IDP. This dataset was used to train machine learning models.
- *Independent test dataset*: The rest of the population for each MRI modality was set as independent test datasets – not touched until we finalized the training procedure.

The GM-IDP includes 119 GM regional volumes from the MUSE atlas. For WM-IDP, we studied different features: i) 48 FA values; ii) 109 TBSS-based[79] values from FA, MD, ODI, and NDI; 192 skeleton-based values from FA, MD, ODI, and NDI.

Table 1. Study characteristics.
The current table presents participants of all ancestries. We constrained participants with only European ancestry for downstream genetic analysis but not the age prediction tasks. * For age and sex, since the number of participants differed for the different MRI modalities, we report statistics for the overlapping population of the three modalities: 35,261 participants for the entire population, 4000 participants for the training/validation/test dataset, and 31,261 participants for the independent test dataset. We also show the number of participants for the GWAS of GM, WM, and FC-BAG. In total, we include 42,089 participants that had at least one image scan.

Population	T1w MRI	dMRI	rsfMRI	Age (year)*	Sex /female*
Total	36,304	39,661	36,858	63.64 (45.00, 81.00)	18,700/53%
Training/validation/test	4000	4000	4000	63.47 (46.00, 81.00)	2000/50%
Independent test	32,304	35,661	32,858	63.66 (45.00, 81.00)	16,700/53%
GWAS	31,557	31,749	32,017	NA	NA

Method 2: Multimodal brain age prediction using machine learning models

GM, WM, and FC-IDP (details of image processing are presented in **Method 3** below) were fit into four machine learning models (linear and non-linear) to predict brain age. Specifically, we used support vector regression (SVR), lasso regression, multilayer perceptron (MLP), and a five-layer neural network (NN: three linear layers and one Relu layer).

To objectively and reproducibly train and compare the performance of age prediction for different machine learning models[26], we adopted a nested CV procedure and included an independent test dataset. Specifically, the outer loop CV was performed for 100 repeated random splits: 80% of the data were used for training. The remaining 20% was used for validation/testing in the inner loop with a 10-fold CV. In addition, we concealed an independent test dataset[80] – not touched for testing until we finished fine-tuning the machine learning models (e.g., hyperparameters for SVR and neural networks). To present the results, we showed the mean and

empirical standard deviation of MAE of the age prediction task instead of performing any statistical test (e.g., a two-sample t-test) because no unbiased variance estimate exists for complex CV procedures (refer to notes from Nadeau and Benjio[81]).

Method 3: Image processing

(A): T1-weighted MRI processing: The imaging quality check is detailed in **Supplementary eMethod 2**. All images were first corrected for magnetic field intensity inhomogeneity.[82] A deep learning-based skull stripping algorithm was applied to remove extra-cranial material. In total, 145 anatomical regions of interest (ROIs) were generated in gray matter (GM, 119 ROIs), white matter (WM, 20 ROIs), and ventricles (6 ROIs) using a multi atlas label fusion method.[83] The 119 GM ROIs were fit to the four machine learning models to derive the GM-BAG.

(B): Diffusion MRI processing: UKBB has processed dMRI data and released several WM tract-based metrics for the Diffusion Tensor Imaging (DTI) model (single-shell dMRI) and Neurite Orientation Dispersion and Density Imaging (NODDI[84]) model (multi-shell dMRI). The Eddy[85] tool corrected raw images for eddy currents, head motion, and outlier slices. The mean values of FA, MD, ODI, and NDI were extracted from the 48 WM tracts of the "ICBM-DTI-81 white-matter labels" atlas[86], resulting in 192 WM-IDP (category code:134). In addition, a tract-skeleton (TBSS)[79] and probabilistic tractography analysis[87] were employed to derive weighted-mean measures within the 27 major WM tracts, referred to as the 108 TBSS WM-IDP (category code: 135). Finally, to overcome the overfitting problem due to the high correlations between FA, MD, ODI, and NDI within the same WM tract, we also used only the 48 FA WM-IDP to fit the models for brain age prediction.

(C): Resting-state functional MRI processing: For FC-IDP, we used the 21 × 21 resting-state functional connectivity (full correlation) matrices (data-field code: 25750) from UKBB[88,89]. UKBB processed rsfMRI data and released 25 whole-brain spatial independent component analysis (ICA)-derived components[90]; four components were removed due to artifactual components. This resulted in 210 FC-IDP quantifying pairwise correlations of the ICA-derived components. Details of dMRI and rsfMRI processing are documented here: https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/brain_mri.pdf.

Method 4: Genetic analyses

Imputed genotype data were quality-checked for downstream analyses. Our quality check pipeline (see below**)** resulted in 33,541 European ancestry participants and 8,469,833 SNPs. After merging with the multimodal MRI populations, we included 31,557 European participants for GM, 31,749 participants for WM-BAG, and 32,017 participants for FC-BAG GWAS. Details of the protocol are described elsewhere[15,69], and documented in our BRIDGEPORT web portal: https://www.cbica.upenn.edu/bridgeport/data/pdf/BIGS_genetic_protocol.pdf.

(A): Genome-wide association analysis: For GWAS, we ran a linear regression using Plink[91] for GM, WM, and FC-BAG, controlling for confounders of age (at imaging), status in the training or independent test dataset, age-squared, sex, age-sex interaction, age-squared-sex interaction, ICV, and the first 40 genetic principal components, following a previous pioneer study[13]. We adopted the genome-wide P-value threshold (5×10^{-8}) and annotated independent

genetic signals considering linkage disequilibrium (see below). We then estimated the SNP-based heritability using GCTA[40] with the same covariates in GWAS.

(B): Annotation of genomic loci and genes: The annotation of genomic loci and mapped genes was performed via FUMA (https://fuma.ctglab.nl/, version: v1.3.8). For the annotation of genomic loci, we first defined lead SNPs (correlation $r^2 \leq 0.1$, distance < 250 kilobases) and assigned them to a genomic locus (non-overlapping); the lead SNP with the lowest P-value (i.e., the top lead SNP) was used to represent the genomic locus. For gene mappings, three different strategies were considered. First, positional mapping assigns the SNP to its physically nearby genes (a 10 kb window by default). Second, eQTL mapping annotates SNPs to genes based on eQTL associations. Finally, chromatin interaction mapping annotates SNPs to genes when there is a significant chromatin interaction between the disease-associated regions and nearby or distant genes.[92] The definition of top lead SNP, lead SNP, independent significant SNP, and candidate SNP can be found in **Supplementary eMethod 1**.

(C): Phenome-wide association query for genomic loci associated with other traits: We manually queried the significant independent SNPs within each locus in the GWAS Catalog to determine their previously identified associations with other traits. For these associated traits, we further mapped them into several high-level categories for visualization purposes (**Fig. 2B**).

(D): Genetic correlation: We used LDSC[93] to estimate the pairwise genetic correlation (r_g) between GM, WM, and FC-BAG and the pre-selected traits (**Supplementary eTable 3**) by using the precomputed LD scores from the 1000 Genomes of European ancestry. The following pre-

selected traits were included: Alzheimer's Disease (AD), Autism Spectrum Disorder (ASD), Attention-Deficit/Hyperactivity Disorder (ADHD), Obsessive-Compulsive Disorder (OCD), Major Depressive Disorder (MDD), Bipolar Disorder (BIP), schizophrenia (SCZ), education and intelligence, as well as several neuroanatomical subtypes for AD (AD1 and AD2[94]), ASD (ASD1, ASD2, and ASD3[44]), and SCZ (SCZ1 and SCZ2[95]) – serving as more robust endophenotypes than the disease diagnoses themselves – derived from machine learning. To ensure the suitability of the GWAS summary statistics, we first checked that the selected study's population was predominantly of European ancestry; we then guaranteed that the SNP-based heritability thereof and excluded the studies with spurious low h^2 (<0.05). It is noteworthy that LDSC corrects for sample overlaps to provide an unbiased estimation of genetic correlation[96].

(E): Partitioned heritability estimate: Partitioned heritability analysis estimates the percentage of heritability enrichment explained by annotated genome regions[48]. First, the partitioned heritability was estimated for 53 main functional categories. The 53 functional categories are not specific to any cell type, including coding, UTR, promoter and intronic regions, etc. Details of the 53 categories are described elsewhere[48]. Subsequently, cell type-specific partitioned heritability was estimated using gene sets from Cahoy et al.[53] for three main cell types (i.e., astrocyte, neuron, and oligodendrocyte), gene sets from multi-tissue chromatin (ROADMAP[97] and ENTEx[98]) data, and gene sets from multi-tissue gene expression (GTEx[99]) data.

(F): Gene-drug-disease network construction: We curated data from the Drug Bank database (v.5.1.9)[100] and the Therapeutic Target Database (updated by September 29th, 2021) to construct a gene-drug-disease network. Specifically, we constrained the target to human organisms and

included all drugs with active statuses (e.g., patented, approved, etc.) but excluded inactive ones (e.g., terminated or discontinued at any phase). To represent the disease, we mapped the identified drugs to the Anatomical Therapeutic Chemical (ATC) classification system for the Drugbank database and the International Classification of Diseases (ICD-11) for the Therapeutic Target Database.

(G): Two-sample Mendelian Randomization: We investigated whether these traits previously associated with our genomic loci were a cause or a consequence of GM, WM, and FC-BAG using a bidirectional two-sample MR approach. GM, WM, and FC-BAG are the outcome/exposure variables in the forward/inverse MR, respectively. We applied five different MR methods using the TwoSampleMR R package[54], and reported the results of IVW in the main text and the four others (e.g., Egger estimator) in the supplementary.

We selected exposure variables from different categories, including neurodegenerative diseases (e.g., AD), biomarkers from the liver (e.g., AST), cardiovascular diseases (e.g., the triglyceride-to-lipid ratio in VLDL), and lifestyle risk factors (e.g., BMI), hypothesizing that different diseases and lifestyle factors may lead to the risk/protection of accelerated/ decelerated brain age. After a specific selection procedure, we also manually checked these papers and excluded studies that used the UKBB data in the discovery samples, resulting in 7 exposure variables. The characteristics of selected studies for the IVs are presented in **Supplementary eTable 6**.

We performed several sensitivity analyses. First, a heterogeneity test was performed to inform the violation of the IV assumptions. Horizontal pleiotropy was estimated to navigate the violation of the IV's exclusivity assumption[58] using a funnel plot, single-SNP MR approaches,

and MR Egger estimator[101]. Moreover, the leave-one-out analysis excluded one instrument (SNP) at a time and assessed the sensitivity of the results to individual SNP.

Data Availability

The GWAS summary statistics corresponding to this study are publicly available on the BRIDGEPORT web portal (https://www.cbica.upenn.edu/bridgeport/), the FUMA online platform (https://fuma.ctglab.nl/), the GWAS Catalog platform (https://www.ebi.ac.uk/gwas/home), and the IEU GWAS database (https://gwas.mrcieu.ac.uk/).

Code Availability

The software and resources used in this study are all publicly available:

- MLNI: https://anbai106.github.io/mlni/, brain age prediction
- BIGS: https://www.cbica.upenn.edu/bridgeport/data/pdf/BIGS_genetic_protocol.pdf, genetic processing protocol
- BRIDGEPORT: https://www.cbica.upenn.edu/bridgeport/, web portal for dissemination
- MUSE: https://www.med.upenn.edu/sbia/muse.html, image preprocessing
- PLINK: https://www.cog-genomics.org/plink/, GWAS
- FUMA: https://fuma.ctglab.nl/, gene mapping, genomic locus annotation
- GCTA: https://yanglab.westlake.edu.cn/software/gcta/#Overview, heritability estimates
- LDSC: https://github.com/bulik/ldsc, genetic correlation, and partitioned heritability
- TwoSampleMR: https://mrcieu.github.io/TwoSampleMR/index.html, MR

Competing Interests

None

Authors' contributions

Dr. Wen has full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Wen

Acquisition, analysis, or interpretation of data: Wen

Drafting of the manuscript: Wen

Critical revision of the manuscript for important intellectual content: all authors

Statistical and genetic analysis: Wen

References

1. Hodson, R. The brain. *Nature* **571**, S1–S1 (2019).
2. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. *Nat Med* **28**, 31–38 (2022).
3. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-Inspired Artificial Intelligence. *Neuron* **95**, 245–258 (2017).
4. Lee, J. *et al.* Deep learning-based brain age prediction in normal aging and dementia. *Nat Aging* **2**, 412–424 (2022).
5. Wen, J. *et al.* Genetic, clinical underpinnings of subtle early brain change along Alzheimer's dimensions. 2022.09.16.508329 Preprint at https://doi.org/10.1101/2022.09.16.508329 (2022).
6. Hollon, T. *et al.* Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging. *Nat Med* 1–5 (2023) doi:10.1038/s41591-023-02252-4.
7. Cole, J. H., Marioni, R. E., Harris, S. E. & Deary, I. J. Brain age and other bodily 'ages': implications for neuropsychiatry. *Mol Psychiatry* **24**, 266–281 (2019).
8. Jones, D. T., Lee, J. & Topol, E. J. Digitising brain age. *The Lancet* **400**, 988 (2022).
9. Tian, Y. E. *et al.* Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nat Med* 1–11 (2023) doi:10.1038/s41591-023-02296-6.
10. Ning, K. *et al.* Improving brain age estimates with deep learning leads to identification of novel genetic factors associated with brain aging. *Neurobiology of Aging* **105**, 199–204 (2021).

11. Shen, L. & Thompson, P. M. Brain Imaging Genomics: Integrated Analysis and Machine Learning. *Proceedings of the IEEE* **108**, 125–162 (2020).

12. Smith, S. M. *et al.* An expanded set of genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nat Neurosci* **24**, 737–745 (2021).

13. Zhao, B. *et al.* Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. *Nat Genet* **51**, 1637–1644 (2019).

14. Zhao, B. *et al.* Common genetic variation influencing human white matter microstructure. *Science* **372**, (2021).

15. Wen, J. *et al.* Novel genomic loci and pathways influence patterns of structural covariance in the human brain. 2022.07.20.22277727 Preprint at https://doi.org/10.1101/2022.07.20.22277727 (2022).

16. Grasby, K. L. *et al.* The genetic architecture of the human cerebral cortex. *Science* **367**, eaay6690 (2020).

17. Brouwer, R. M. *et al.* Genetic variants associated with longitudinal changes in brain structure across the lifespan. *Nat Neurosci* **25**, 421–432 (2022).

18. Zhao, B. *et al.* Common variants contribute to intrinsic human brain functional networks. *Nat Genet* **54**, 508–517 (2022).

19. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).

20. Smith, S. M. *et al.* Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *eLife* **9**, e52677.

21. Kaufmann, T. *et al*. Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat Neurosci* **22**, 1617–1623 (2019).

22. Kendler, K. & Neale, M. Endophenotype: a conceptual analysis. *Mol Psychiatry* **15**, 789–797 (2010).

23. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

24. Emdin, C. A., Khera, A. V. & Kathiresan, S. Mendelian Randomization. *JAMA* **318**, 1925–1926 (2017).

25. Varoquaux, G. *et al.* Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *NeuroImage* **145**, 166–179 (2017).

26. Samper-González, J. *et al.* Reproducible evaluation of classification methods in Alzheimer's disease: Framework and application to MRI and PET data. *NeuroImage* **183**, 504–521 (2018).

27. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

28. More, S. *et al.* Brain-age prediction: A systematic comparison of machine learning workflows. *NeuroImage* **270**, 119947 (2023).

29. de Lange, A.-M. G. & Cole, J. H. Commentary: Correction procedures in brain-age prediction. *Neuroimage Clin* **26**, 102229 (2020).

30. Jun, G. *et al.* A novel Alzheimer disease locus located near the gene encoding tau protein. *Mol Psychiatry* **21**, 108–117 (2016).

31. Nagel, M. *et al.* Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet* **50**, 920–927 (2018).

32. Desikan, R. S. *et al.* Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. *Mol Psychiatry* **20**, 1588–1595 (2015).

33. Hamza, T. H. *et al.* Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet* **42**, 781–785 (2010).

34. Kuchenbaecker, K. B. *et al.* Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nat Genet* **47**, 164–171 (2015).

35. de la Fuente, J., Davies, G., Grotzinger, A. D., Tucker-Drob, E. M. & Deary, I. J. A general dimension of genetic sharing across diverse cognitive traits inferred from molecular data. *Nat Hum Behav* **5**, 49–58 (2021).

36. Georges, A. *et al.* Genetic investigation of fibromuscular dysplasia identifies risk loci and shared genetics with common cardiovascular diseases. *Nat Commun* **12**, 6031 (2021).

37. Nagel, M., Speed, D., van der Sluis, S. & Østergaard, S. D. Genome-wide association study of the sensitivity to environmental stress and adversity neuroticism cluster. *Acta Psychiatrica Scandinavica* **141**, 476–478 (2020).

38. Russell, G. & Lightman, S. The human stress response. *Nat Rev Endocrinol* **15**, 525–534 (2019).

39. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012 (2019).

40. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am J Hum Genet* **88**, 76–82 (2011).

41. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat Rev Drug Discov* **1**, 727–730 (2002).

42. Antibody-Mediated Targeting of Tau In Vivo Does Not Require Effector Function and Microglial Engagement - PubMed. https://pubmed.ncbi.nlm.nih.gov/27475227/.

43. Wilcock, G. K. *et al.* Potential of Low Dose Leuco-Methylthioninium Bis(Hydromethanesulphonate) (LMTM) Monotherapy for Treatment of Mild Alzheimer's Disease: Cohort Analysis as Modified Primary Outcome in a Phase III Clinical Trial. *J Alzheimers Dis* **61**, 435–457.

44. Hwang, G. *et al.* Assessment of Neuroanatomical Endophenotypes of Autism Spectrum Disorder and Association With Characteristics of Individuals With Schizophrenia and the General Population. *JAMA Psychiatry* (2023) doi:10.1001/jamapsychiatry.2023.0409.

45. Chand, G. B. *et al.* Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning. *Brain* **143**, 1027–1038 (2020).

46. International Obsessive Compulsive Disorder Foundation Genetics Collaborative (IOCDF-GC) and OCD Collaborative Genetics Association Studies (OCGAS). Revealing the complex genetic architecture of obsessive-compulsive disorder using meta-analysis. *Mol Psychiatry* **23**, 1181–1188 (2018).

47. Rietveld, C. A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).

48. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–1235 (2015).

49. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827–841 (2013).

50. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**, 621–637 (2018).

51. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* **45**, 124–130 (2013).

52. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

53. Cahoy, J. D. *et al.* A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *J. Neurosci.* **28**, 264–278 (2008).

54. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

55. Borges, M. C. *et al.* Circulating Fatty Acids and Risk of Coronary Heart Disease and Stroke: Individual Participant Data Meta-Analysis in Up to 16 126 Participants. *J Am Heart Assoc* **9**, e013131 (2020).

56. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* **44**, 981–990 (2012).

57. K, M. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, (2017).

58. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med* **36**, 1783–1802 (2017).

59. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol* **32**, 377–389 (2017).

60. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat Genet* **50**, 1514–1523 (2018).

61. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet* **49**, 1385–1391 (2017).

62. Jonsson, B. A. *et al.* Brain age prediction using deep learning uncovers associated sequence variants. *Nat Commun* **10**, 5409 (2019).

63. Ning, K., Zhao, L., Matloff, W., Sun, F. & Toga, A. W. Association of relative brain age with tobacco smoking, alcohol consumption, and genetic variants. *Sci Rep* **10**, 10 (2020).

64. Horie, K. *et al.* CSF tau microtubule-binding region identifies pathological changes in primary tauopathies. *Nat Med* **28**, 2547–2554 (2022).

65. Gusev, A. *et al.* Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *The American Journal of Human Genetics* **95**, 535–552 (2014).

66. Stamatoyannopoulos, J. A. What does our genome encode? *Genome Res* **22**, 1602–1611 (2012).

67. Yang, Z. *et al.* A deep learning framework identifies dimensional representations of Alzheimer's Disease from brain structure. *Nat Commun* **12**, 7065 (2021).

68. Wen, J. *et al.* Multi-scale semi-supervised clustering of brain images: Deriving disease subtypes. *Med Image Anal* **75**, 102304 (2021).

69. Wen, J. *et al.* Characterizing Heterogeneity in Neuroimaging, Cognition, Clinical Symptoms, and Genetics Among Patients With Late-Life Depression. *JAMA Psychiatry* (2022) doi:10.1001/jamapsychiatry.2022.0020.

70. Vogel, J. W. *et al.* Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med* **27**, 871–881 (2021).

71. Young, A. L. *et al.* Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun* **9**, 4273 (2018).

72. Nordestgaard, B. G. & Varbo, A. Triglycerides and cardiovascular disease. *Lancet* **384**, 626–635 (2014).

73. Antal, B. *et al.* Type 2 diabetes mellitus accelerates brain aging and cognitive decline: Complementary findings from UK Biobank and meta-analyses. *Elife* **11**, e73138 (2022).

74. Wu, A. M. L. *et al.* Aging and CNS Myeloid Cell Depletion Attenuate Breast Cancer Brain Metastasis. *Clinical Cancer Research* **27**, 4422–4434 (2021).

75. Sachdev, P. S., Zhuang, L., Braidy, N. & Wen, W. Is Alzheimer's a disease of the white matter? *Curr Opin Psychiatry* **26**, 244–251 (2013).

76. Watanabe, K. *et al.* Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways. *Nat Genet* **54**, 1125–1132 (2022).

77. Cermakova, P. *et al.* Parental education, cognition and functional connectivity of the salience network. *Sci Rep* **13**, 2761 (2023).

78. Cao, H., Zhou, H. & Cannon, T. D. Functional connectome-wide associations of schizophrenia polygenic risk. *Mol Psychiatry* **26**, 2553–2561 (2021).

79. Smith, S. M. *et al.* Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage* **31**, 1487–1505 (2006).

80. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis* **63**, 101694 (2020).

81. Inference for the Generalization Error | SpringerLink. https://link.springer.com/article/10.1023/A:1024068626366.

82. Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).

83. Doshi, J. *et al.* MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. *Neuroimage* **127**, 186–195 (2016).

84. Zhang, H., Schneider, T., Wheeler-Kingshott, C. A. & Alexander, D. C. NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage* **61**, 1000–1016 (2012).

85. Smith, S. M. *et al.* Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23 Suppl 1**, S208-19 (2004).

86. Mori, S., Wakana, S., Nagae-Poetscher, L. & van Zijl, P. *MRI Atlas of Human White Matter*. (Elsevier, 2005).

87. Wakana, S. *et al.* Reproducibility of quantitative tractography methods applied to cerebral white matter. *NeuroImage* **36**, 630–644 (2007).

88. Alfaro-Almagro, F. *et al.* Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage* **166**, (2018).

89. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience* **19**, 1523–1536 (2016).

90. Beckmann, C. F. & Smith, S. M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Transactions on Medical Imaging* **23**, 137–152 (2004).

91. Purcell, S. *et al*. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).

92. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

93. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).

94. Wen, J. *et al*. Genetic, clinical underpinnings of subtle early brain change along Alzheimer's dimensions. 2022.09.16.508329 Preprint at https://doi.org/10.1101/2022.09.16.508329 (2022).

95. Chand, G. B. *et al.* Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning. *Brain* **143**, 1027–1038 (2020).

96. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236–1241 (2015).

97. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045–1048 (2010).

98. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

99. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

100. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074–D1082 (2018).

101. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512–525 (2015).

Acknowledgments

The iSTAGING consortium is a multi-institutional effort funded by NIA by RF1 AG054409. This research has been conducted using the UK Biobank Resource under Application Number 35148.