Novel genomic loci and pathways influence patterns of structural covariance in the human brain

Junhao Wen[1*], Ilya M. Nasrallah[1,2], Ahmed Abdulkadir[1], Theodore D. Satterthwaite[1,3], Zhijian Yang[1], Guray Erus[1], Timothy Robert-Fitzgerald[4], Ashish Singh[1], Aristeidis Sotiras[5], Aleix Boquet-Pujadas[6], Elizabeth Mamourian[1], Jimit Doshi[1], Yuhan Cui[1], Dhivya Srinivasan[1], Jiong Chen[1], Gyujoon Hwang[1], Mark Bergman[1], Jingxuan Bao[7], Yogasudha Veturi[8], Zhen Zhou[1], Shu Yang[7], Paola Dazzan[9], Rene S. Kahn[10], Hugo G. Schnack[11], Marcus V. Zanetti[12], Eva Meisenzahl[13], Geraldo F. Busatto[12], Benedicto Crespo-Facorro[14], Christos Pantelis[15], Stephen J. Wood[16], Chuanjun Zhuo[17], Russell T. Shinohara[1,4], Ruben C. Gur[3], Raquel E. Gur[3], Nikolaos Koutsouleris[18], Daniel H. Wolf[1,3], Andrew J. Saykin[19], Marylyn D. Ritchie[8], Li Shen[7], Paul M. Thompson[20], Olivier Colliot[21], Katharina Wittfeld[22], Hans J. Grabe[22], Duygu Tosun[23], Murat Bilgel[24], Yang An[24], Daniel S. Marcus[25], Pamela LaMontagne[25], Susan R. Heckbert[26], Thomas R. Austin[26], Lenore J. Launer[27], Mark Espeland[28], Colin L Masters[29], Paul Maruff[29], Jurgen Fripp[30], Sterling C. Johnson[31], John C. Morris[32], Marilyn S. Albert[33], R. Nick Bryan[2], Susan M. Resnick[24], Yong Fan[1], Mohamad Habes[34], David Wolk[1, 35], Haochang Shou[1,4], and Christos Davatzikos[1*], for the iSTAGING, the BLSA, the BIOCARD, the PHENOM, the ADNI studies, and the AI4AD consortium

[1]Artificial Intelligence in Biomedical Imaging Laboratory (AIBIL), Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.
[2]Department of Radiology, University of Pennsylvania, Philadelphia, USA.
[3]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
[4]Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA
[5]Department of Radiology and Institute for Informatics, Washington University School of Medicine, St. Louis, USA
[6]Biomedical Imaging Group, EPFL, Lausanne, Switzerland
[7]Department of Biostatistics, Epidemiology and Informatics University of Pennsylvania Perelman School of Medicine, Philadelphia, USA
[8]Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[9]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[10]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA
[11]Department of Psychiatry, University Medical Center Utrecht, Utrecht, Netherlands
[12]Institute of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil
[13] Department of Psychiatry and Psychtherapy, HHU Düsseldorf, Germany
[14]Hospital Universitario Virgen del Rocio, University of Sevilla-IBIS; IDIVAL-CIBERSAM, Sevilla, Spain
[15]Melbourne Neuropsychiatry Centre, Department of Psychiatry, University of Melbourne and Melbourne Health, Carlton South, Australia
[16]Orygen and the Centre for Youth Mental Health, University of Melbourne; and the Scool of Psychology, University of Birmingham, UK
[17]key Laboratory of Real Tine Tracing of Brain Circuits in Psychiatry and Neurology(RTBCPN-Lab), Nankai University Affiliated Tianjin Fourth Center Hospital; Department of Psychiatry, Tianjin Medical University, Tianjin, China
[18]Department of Psychiatry and Psychotherapy, Ludwig-Maximilian University, Munich, Germany
[19]Radiology and Imaging Sciences, Center for Neuroimaging, Department of Radiology and Imaging Sciences, Indiana Alzheimer's Disease Research Center and the Melvin and Bren Simon Cancer Center, Indiana University School of Medicine, Indianapolis
[20]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
[21]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France

[22]Department of Psychiatry and Psychotherapy, German Center for Neurodegenerative Diseases (DZNE), University Medicine Greifswald, Germany
[23]Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA
[24]Laboratory of Behavioral Neuroscience, National Institute on Aging, NIH, USA
[25]Department of Radiology, Washington University School of Medicine, St. Louis, Missouri, USA
[26]Cardiovascular Health Research Unit, University of Washington, Seattle, WA, USA
[27]Neuroepidemiology Section, Intramural Research Program, National Institute on Aging, Bethesda, Maryland, USA
[28]Sticht Center for Healthy Aging and Alzheimer's Prevention, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA
[29]Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, VIC, Australia
[30]CSIRO Health and Biosecurity, Australian e-Health Research Centre CSIRO, Brisbane, Queensland, Australia
[31]Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA
[32]Knight Alzheimer Disease Research Center, Washington University in St. Louis, St. Louis, MO, USA
[33]Department of Neurology, Johns Hopkins University School of Medicine, USA
[34]Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases, University of Texas Health Science Center at San Antonio, San Antonio, USA
[35]Department of Neurology and Penn Memory Center, University of Pennsylvania, Philadelphia, USA

[*]Corresponding authors:
Junhao Wen, Ph.D. – junhao.wen89@gmail.com
Christos Davatzikos, Ph.D. – Christos.Davatzikos@pennmedicine.upenn.edu
3700 Hamilton Walk, 7th Floor, Philadelphia, PA 19104

Word counts: 4532 words

Abstract

Normal and pathologic neurobiological processes influence brain morphology in coordinated ways that give rise to patterns of structural covariance (PSC) across brain regions and individuals during brain aging and brain diseases. The genetic underpinnings of these patterns remain largely unknown. We apply a stochastic multivariate factorization method to a diverse population of 50,699 individuals (12 studies, 130 sites) and derive data-driven, multi-scale PSCs of regional brain size. PSCs were significantly correlated with 915 genomic loci in the discovery set, 617 of which are novel, and 72% were independently replicated. Key pathways influencing PSCs involved reelin signaling, apoptosis, neurogenesis, and appendage development, while pathways of breast cancer indicate potential interplays between brain metastasis and PSCs associated with neurodegeneration and dementia. Using machine learning, multi-scale PSCs effectively derive imaging signatures of several brain diseases. Our results elucidate new genetic and biological underpinnings that influence structural covariance patterns in the human brain.

Main

Brain structure and function are interrelated via complex networks that operate at multiple scales, ranging from cellular and synaptic processes, such as neural migration, synapse formation, and axon development, to local and broadly connected circuits.[1] Due to a fundamental relationship between activity and structure, many normal and pathologic neurobiological processes, driven by genetic and environmental factors, collectively cause coordinated changes in brain morphology. Structural covariance analyses investigate such coordinated changes by seeking patterns of structural covariation (PSC) across brain regions and individuals.[1] For example, during adolescence, PSCs derived from magnetic resonance imaging (MRI) have been considered to reflect a coordinated cortical remodeling as the brain establishes mature networks of functional specialization.[2] Structural covariance is not only related to normal brain development or aging processes but can also reflect coordinated brain change due to disease. For example, individuals with motor speech dysfunction may develop brain atrophy in Broca's inferior frontal cortex and co-occurring brain atrophy in Wernicke's area of the superior temporal cortex.[3] See **Fig. 1C** for an example.

The human brain develops, matures, and degenerates in coordinated patterns of structural covariance at the macrostructural level of brain morphology.[1] However, the mechanisms underlying structural covariance are still unclear, and their genetic underpinnings are largely unknown. We hypothesized that brain morphology was driven by multiple genes (i.e., polygenic) collectively operating on different brain areas (i.e., pleiotropic), resulting in connected networks covaried by normal aging and various disease-related processes. Along the causal pathway from underlying genetics to brain morphological changes, we sought to elucidate which genetic underpinnings (e.g., genes), biological processes (e.g., neurogenesis), cellular components (e.g., nuclear membrane), molecular functions (e.g., nucleic acid binding), and neuropathological processes (e.g., Alzheimer's disease) might influence the formation, development, and changes of structural covariance patterns in the human brain.

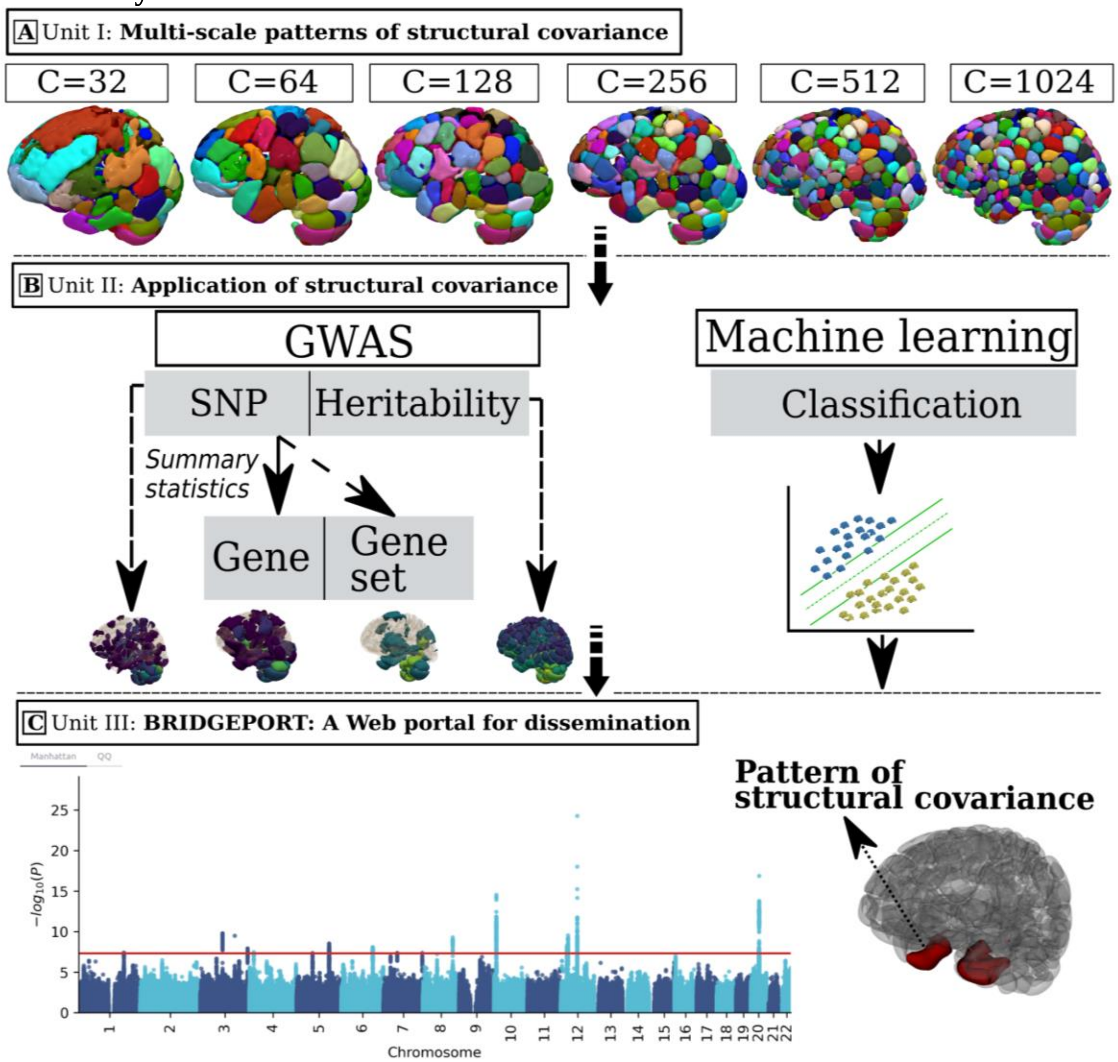
Previous neuroimaging genome-wide association studies (GWAS)[4,5] have partially investigated the abovementioned questions and expanded our understanding of the genetic architecture of the human brain. However, their focus was on conventional neuroanatomical regions of interest (ROI) instead of data-driven PSCs. In brain imaging research, prior studies have applied structural covariance analysis to elucidate underlying coordinated morphological changes in brain aging and various brain diseases,[1] but have had several limitations. They often relied on pre-defined neuroanatomical ROIs to construct inter- and intra-individual structural covariance networks. These *a priori* ROIs might not optimally reflect the molecular-functional characteristics of the brain. In addition, most population-based studies have investigated brain structural covariance within a relatively limited scope, such as within relatively small samples, over a relatively narrow age window (e.g., adolescence[2]), within a single disease (e.g., Parkinson's disease[6]), or within datasets lacking sufficient diversity in cohort characteristics or MRI scanner protocols. These have been imposed, in part, by limitations in both available cohort size and in the algorithmic implementation of structural covariance analysis, which has been computationally restricted to modest sample sizes when investigated at full image resolution. Lastly, prior studies have examined brain structural covariance at a single fixed ROI resolution/scale/granularity. While the optimal scale is unknown and may differ by the question of interest, the highly complex organization of the human brain may demonstrate structural covariance patterns that span multiple scales.[7,8]

We examined structural covariance of regional cortical and subcortical volume in the human brain using MRI from a diverse population of 50,699 people from 12 studies, 130 sites, and 12 countries, comprised of cognitively healthy individuals, as well as participants with various diseases/conditions over their lifespan (ages 5 through 97). In order to enable such a large-scale study, we developed a stochastic adaptation of the orthogonally projective non-negative matrix factorization[9] to derive multi-scale PSCs. This method allowed us to derive PSCs at any desirable scale, which is defined by the number, C, of derived PSCs. Herein we present results from coarse to fine scales corresponding to C = 32, 64, 128, 256, 512, and 1024. We hypothesized that PSCs at multiple scales could delineate the human brain's multi-factorial and multi-faceted morphological landscape and genetic architecture in healthy and diseased individuals. We examined the associations between these multi-scale PSCs and common genetic variants at different levels (N=8,469,833 SNPs). In total, 617 novel genomic loci were identified; key pathways (e.g., neurogenesis and reelin signaling) contributed to shaping structural covariance patterns in the human brain. In addition, we leveraged PSCs at multiple scales to better derive individualized imaging signatures of several diseases than any single scale PSCs using machine learning. All experimental results and the multi-scale PSCs were integrated into the MuSIC (Multi-scale Structural Imaging Covariance) atlas and made publicly accessible through the BRIDGEPORT (BRaIn knowleDGE PORTal) web portal: https://www.cbica.upenn.edu/bridgeport/.

Results

We summarize this work in three units (I to III) outlined in **Fig. 1**. In Unit **I** (**Fig. 1A**), we present the stochastic orthogonally projective non-negative matrix factorization (sopNMF) algorithm (**Method 1**), optimized for large-scale multivariate structural covariance analysis. The sopNMF algorithm decomposes large-scale imaging data through online learning to overcome the memory limitations of opNMF. A subgroup of participants with multiple disease diagnoses and healthy controls (ages 5-97, training population, *N*=4000, **Method 2**) were sampled from the discovery set (*N*=32,440, **Method 2**); their MRI underwent a standard imaging processing pipeline (**Method 3A**). The processed images were then fit to sopNMF to derive the multi-scale PSCs (*N*=2003) from the loadings of the factorization (**Method 1**). We incorporate participants with various disease conditions because previous studies have demonstrated that inter-regional correlated patterns (i.e., depicting a network) show variations in healthy and diseased populations, albeit to a differing degree.[10] Multi-scale PSCs were extracted across the entire population and statistically harmonized[11] (**Method 3B**). Unit **II** (**Fig. 1B**) investigates the harmonized data for 2003 PSCs (13 PSCs have vanished in this process for *C*=1024; see **Method 1**) in two brain structural covariance analyses. Specifically, we performed *i*) GWAS (**Method 4**) that sought to discover associations of PSCs at single nucleotide polymorphism (SNP), gene, or gene set-level; and *ii*) pattern analysis via machine learning (**Method 5**) to derive individualized imaging signatures of several brain diseases and conditions. Unit **III** (**Fig. 1C**) presents BRIDGEPORT, making these massive analytic resources publicly available to the imaging, genomics, and machine learning communities.

Figure 1: **Study workflow**



A) Unit I: the stochastic orthogonally projective non-negative matrix factorization (sopNMF) algorithm was applied to a large, disease-diverse population to derive multi-scale patterns of structural covariance (PSC) at different scales (*C*=32, 64, 128, 256, 512, and 1024; *C* represents the number of PSCs). **B**) Unit II: two types of analyses were performed in this study: Genome-wide association studies (GWAS) relate each of the PSCs (*N*=2003) to common genetic variants; pattern analysis via machine learning demonstrates the utility of the multi-scale PSCs in deriving individualized imaging signatures of various brain pathologies. **C**) Unit III: BRIDGEPORT is a web portal that makes all resources publicly available for dissemination. As an illustration, a Manhattan plot for PSC (C64-3, the third PSC of the C64 atlas) and its 3D brain map are displayed.

Patterns of structural covariance via stochastic orthogonally projective non-negative matrix factorization

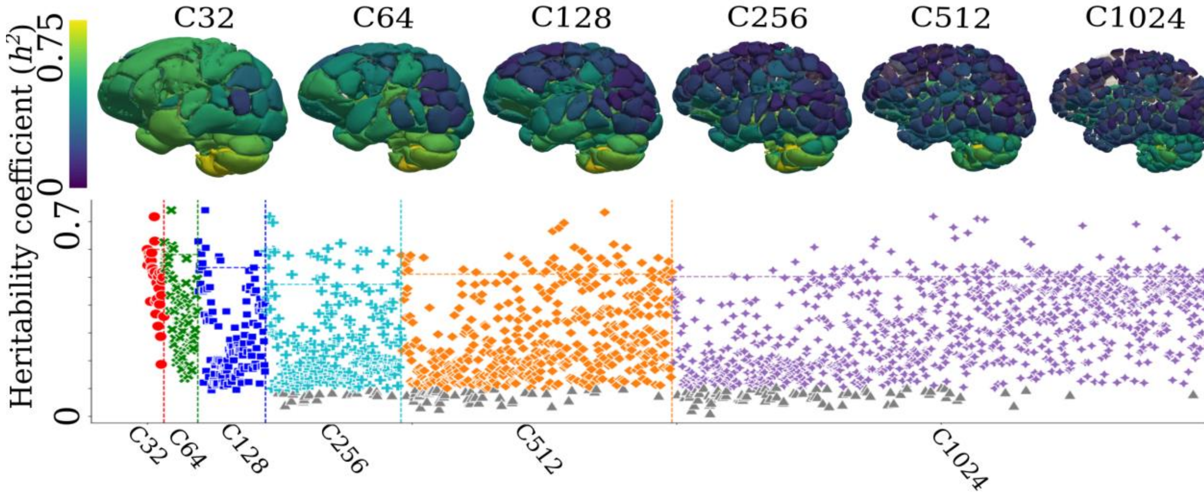
We first validated the sopNMF algorithm by showing that it converged to the global minimum of the factorization problem using the comparison population (N=800, **Method 2**). The sopNMF algorithm achieved similar reconstruction loss and sparsity as opNMF but at reduced memory demand (**Supplementary eFigure 1**). The lower memory requirements of sopNMF made it possible to generate the multi-scale PSCs by jointly factorizing 4000 MRIs in the training population. The results of the algorithm were robust and obtained a high reproducibility index (RI) (**Supplementary eMethod 2**) in reproducibility analyses: split-sample analysis (RI = 0.76 ± 0.27) and split-sex analysis (RI = 0.79 ± 0.27) (**Supplementary eFigure 2**). We then extracted the multi-scale PSCs in the discovery set (N=32,440) and the replication set (N=18,259, **Method 2**) for Unit II. These PSCs succinctly capture underlying neurobiological processes across the lifespan, including the effects of typical aging processes and various brain diseases. In addition, the multi-scale representation constructs a hierarchy of brain structure networks (e.g., PSCs in cerebellum regions), which models the human brain in a multi-scale topology.[7,12]

Patterns of structural covariance are highly heritable

The multi-scale PSCs are highly heritable ($0.05 < h^2 < 0.78$), showing high SNP-based heritability estimates (h^2) (**Method 4B**) for the discovery set (**Fig. 2**). Specifically, the h^2 estimate was 0.49 ± 0.10, 0.39 ± 0.14, 0.29 ± 0.15, 0.25 ± 0.15, 0.27 ± 0.15, 0.31 ± 0.15 for scales C=32, 64, 128, 256, 512 and 1024 of the PSCs, respectively. The Pearson correlation coefficient between the two independent estimates of h^2 was r = 0.94 (p-value < 10^{-6}, between the discovery and replication sets) in the UK Biobank (UKBB) data. The scatter plot of the two sets of h^2 estimates is shown in **Supplementary eFigure 3**. The h^2 estimates and p-values for all PSCs are detailed in

Supplementary eFile 1 (discovery set) and **eFile 2** (replication set). Our results confirm that brain structure is heritable to a large extent and identify the spatial distribution of the most highly heritable regions of the brain (e.g., subcortical gray matter structures and cerebellum regions).[13]

Figure 2: **Patterns of structural covariance are highly heritable in the human brain.**



Patterns of structural covariance (PSCs) of the human brain are highly heritable. The SNP-based heritability estimates are calculated for the multi-scale PSCs at different scales (*C*). PSCs surviving Bonferroni correction for multiple comparisons are depicted in color in the Manhattan plots (gray otherwise). The heritability estimate (h^2) of each PSC was projected onto the 3D image space to show a statistical map of the brain at each scale *C*. The dotted line indicates the top 10% most heritable PSCs on each scale.

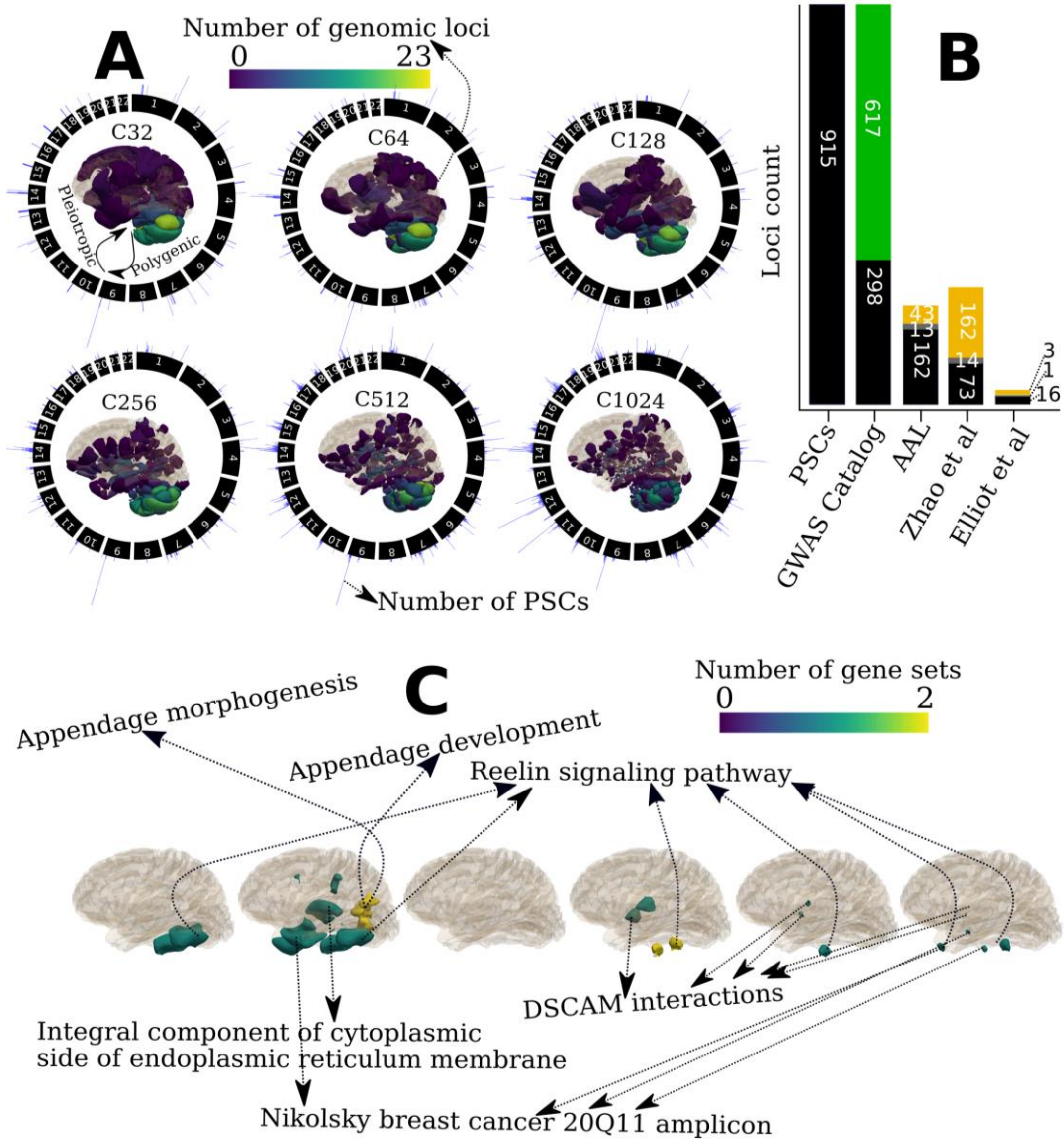
617 novel genomic loci of patterns of structural covariance

We discovered genomic locus-PSC pairwise associations (**Method 4C**, **Supplementary eMethod 5**) within the discovery set and then independently replicated these associations on the replication set. We found that 915 genomic loci had 3791 loci-PSC pairwise significant associations with 924 PSCs after Bonferroni correction (**Method 4G**) for the number of PSCs (p-value threshold per scale: $10.3 > -\log_{10}[\text{p-value}] > 8.8$) (**Supplementary eFile 3**, and **Fig. 3A**). Our results showed that the formation of these PSCs is largely polygenic; the associated SNPs might play a pleiotropic role in shaping these networks.

Compared to previous literature, out of the 915 genomic loci, the multi-scale PSCs identified 617 novel genomic loci not previously associated with any traits or phenotypes in the GWAS Catalog[14] (**Supplementary eFile 4**, **Figure 3B)**. These novel associations might indicate subtle neurobiological processes that are captured thanks to the biologically relevant structural covariance expressed by sopNMF. The multi-scale PSCs identified many novel associations by constraining this comparison to previous neuroimaging GWAS[12,13] using T1w MRI-derived phenotypes (e.g., regions of interest from conventional brain atlases) (**Fig 3B**, **Supplementary eTable 3**, **eFile 5**, **6**, and **7**).

To replicate these genomic loci, our UKBB replication set analysis (**Method 4H**) demonstrated that 3638 (96%) exact genomic locus-PSC associations were replicated at nominal significance ($-\log_{10}$[p-value] > 1.31), 2705 (72%) of which were significant after correction for multiple comparisons (**Method 4G,** $-\log_{10}$[p-value] > 4.27). We present this validation in **Supplementary eFile 8** from the replication set. The summary statistics, Manhattan, and QQ plots derived from the combined population (*N*=33,541) are presented in BRIDGEPORT.

Figure 3: **Patterns of structural covariance highlight novel genomic loci and pathways that shape the human brain.**



A) Patterns of structural covariance (PSC) in the human brain are polygenic: the number of genomic loci of each PSC is projected onto the image space to show a statistical brain map characterized by the number (*C*) of PSCs. In addition, common genetic variants exert pleiotropic effects on the PSCs: circular plots showed the number of associated PSCs (histograms in blue color) of each genomic loci over the entire autosomal chromosomes (1-22). The histogram was plotted for the number of PSCs for each genomic locus in the circular plots. **B**) Novel genomic loci revealed by the multi-scale PSCs compared to previous findings from GWAS Catalog,[14] T1-weighted MRI GWAS,[4,5] and the AAL atlas regions of interest. The green bar indicates the 617 novel genomic loci not previously associated with any clinical traits in GWAS Catalog; the black bar presents the loci identified in other studies that overlap (grey bar for loci in linkage disequilibrium) with the loci from our results; the yellow bar indicates the unique loci in other studies. **C**) Pathway enrichment analysis highlights six unique biological pathways and functional categories (after Bonferroni correction for 16,768 gene sets and the number of PSCs) that might influence the changes of PSCs. DSCAM: Down syndrome cell adhesion molecule.

Gene set enrichment analysis highlights pathways that shape patterns of structural covariance

For gene-level associations (**Method 4D**), we discovered that 164 genes had 2489 gene-PSC pairwise associations with 445 PSCs after Bonferroni correction for the number of genes and PSCs (p-value threshold: $8.6 > -\log_{10}[\text{p-value}] > 7.1$) (**Supplementary eFile 9**).

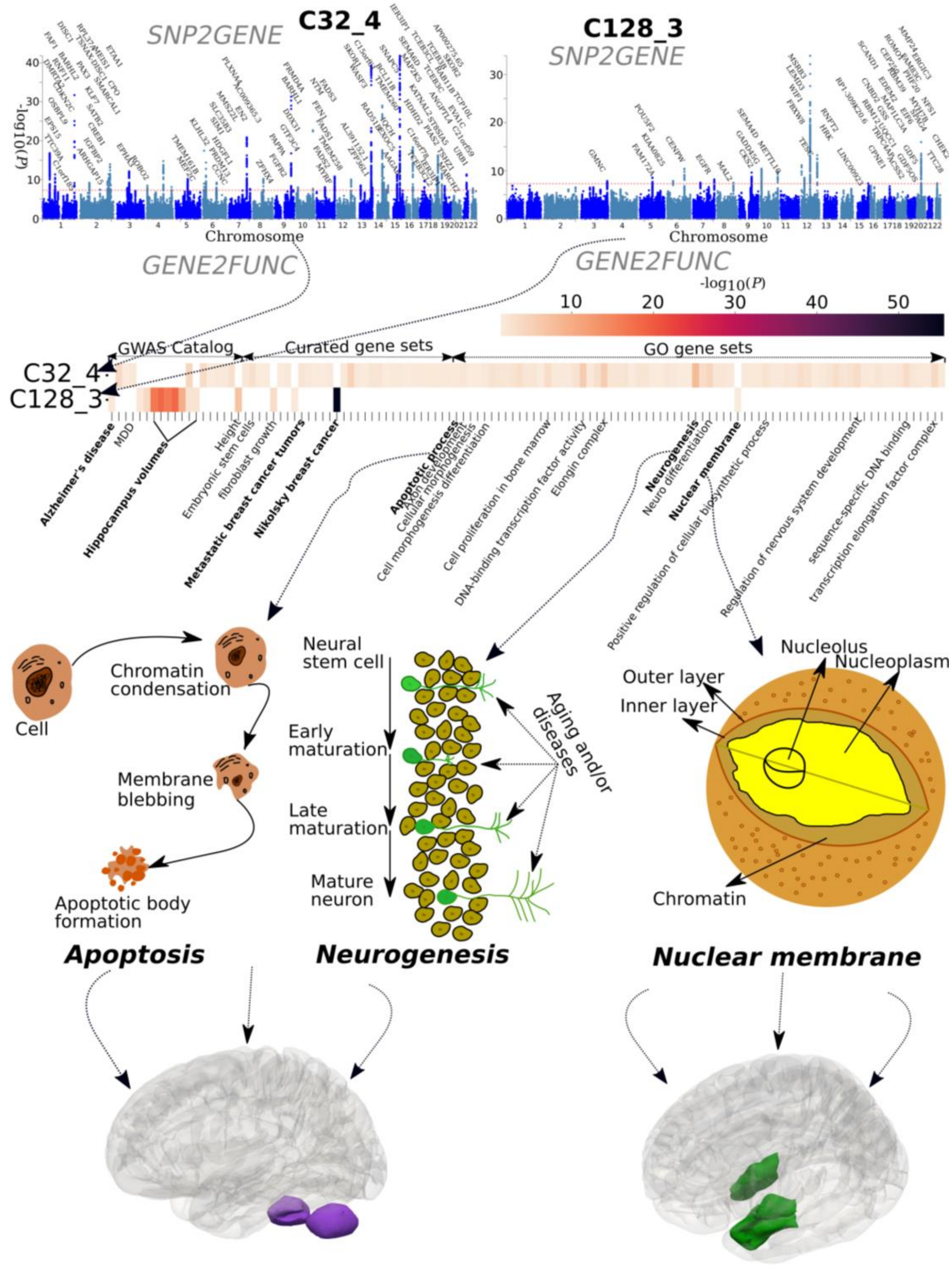
Based on these gene-level p-values, we performed hypothesis-free gene set pathway analysis using MAGMA[15](**Method 4E**): a more stringent correction for multiple comparisons was performed than the prioritized gene set enrichment analysis using *GENE2FUN* from FUMA (**Method 4F**, **Fig. 4**). We identified that six gene set pathways had 18 gene set-PSC pairwise associations with 17 PSCs after Bonferroni correction for the number of gene sets and PSCs (N=16,768 and C from 32 to 1024, p-value threshold: $8.54 > -\log_{10}[\text{p-value}] > 7.03$) (**Fig. 3C**, **Supplementary eFile 10**)**.** These gene sets imply critical biological and molecular pathways that might shape brain morphological changes and development. The reelin signaling pathway exerts vital functions in regulating neuronal migration, dendritic growth, branching, spine formation, synaptogenesis, and synaptic plasticity.[16] The appendage morphogenesis and development pathways indicate how the anatomical structures of appendages are generated, organized and progressed over time, which are often related to the cell adhesion pathway. These pathways elucidate how cells or tissues can be organized to create a complex structure like the human brain.[17] In addition, the integral component of the cytoplasmic side of the endoplasmic reticulum membrane is thought to form a continuous network of tubules and cisternae extending throughout neuronal dendrites and axons.[18] The DSCAM (Down syndrome cell adhesion molecule) pathway likely functions as a cell surface receptor mediating axon pathfinding. Related proteins are involved in hemophilic intercellular interactions.[19] Lastly, Nikolsky et al.[20] defined genes from the breast cancer 20Q11 amplicon pathway were involved in the brain might indicate the brain metastasis of breast cancer, which is usually a late event with deleterious effects on the prognosis.[21] In addition, previous findings[22,23] revealed an inverse relationship between Alzheimer's disease and breast cancer, which might indicate a close genetic relationship between the disease and brain morphological changes mainly affecting the entorhinal cortex and hippocampus (PSC: C128_3 in **Fig. 4**).

Illustrations of genetic loci and pathways forming two patterns of structural covariance

To illustrate how underlying genetic underpinnings might form a specific PSC, we showcased two PSCs: C32_4 for the superior cerebellum and C128_3 for the hippocampus-entorhinal cortex. The two PSCs were highly heritable and polygenic in our GWAS using the entire UKBB data (**Fig. 4,** *N*=33,541). We used the FUMA[24] online platform to perform *SNP2GENE* for annotating the mapped genes and *GENE2FUNC* for prioritized gene set enrichment analyses (**Method 4F**). The superior cerebellum PSC was associated with genomic loci that can be mapped to 85 genes, which were enriched in many biological pathways, including psychiatric disorders, biological processes, molecular functions, and cellular components (e.g., apoptotic process, axon development, cellular morphogenesis, neurogenesis, and neuro differentiation). For example, apoptosis – the regulated cell destruction – is a complicated process that is highly involved in the development and maturation of the human brain and neurodegenerative diseases.[25] Neurogenesis – new neuron formation – is crucial when an embryo develops and continues in specific brain regions throughout the lifespan.[26] All significant results of this prioritized gene set enrichment analysis are presented in **Supplementary eFile 11.**

For the hippocampus-entorhinal cortex PSC, we mapped 45 genes enriched in gene sets defined from GWAS Catalog, including Alzheimer's disease and brain volume derived from hippocampal regions. The hippocampus and medial temporal lobe have been robust hallmarks of Alzheimer's disease.[27] In addition, these genes were enriched in the breast cancer 20Q11 amplicon pathway[20] and the pathway of metastatic breast cancer tumors[28], which might indicate a specific distribution of brain metastases: the vulnerability of medial temporal lobe regions to breast cancer, [21] or highlight an inverse association between Alzheimer's disease and breast cancer.[22] Lastly, the nuclear membrane encloses the cell's nucleus – the chromosomes reside inside – which is critical in cell formation activities related to gene expression and regulation. To further support the overlapping genetic underpinnings between this PSC and Alzheimer's diseases, we calculated the genetic correlation (r_g = -0.28; p-value=0.01) using GWAS summary statistics from the hippocampus-entorhinal cortex PSC (i.e., 33,541 people of European ancestry) and a previous independent study of Alzheimer's disease[29] (i.e., 63,926 people of European ancestry) using LDSC.[30] All significant results of this prioritized gene set enrichment analysis are presented in **Supplementary eFile 12.**

Figure 4: **Illustrations of multiple genetic loci and pathways shaping specific patterns of structural covariance**



We demonstrate how underlying genomic loci and biological pathways might influence the formation, development, and changes of two specific PSCs: the 4th PSC of the C32 PSCs (C32_4) that resides in the superior part of the cerebellum and the 3rd PSC of the C128 PSCs (C128_3) that includes the bilateral hippocampus and entorhinal cortex. We first performed *SNP2GENE* to annotate the mapped genes in the Manhattan plots and then ran *GENE2FUNC* for the prioritized gene set enrichment analysis (**Method 4F**). The mapped genes are used as input genes for prioritized gene set enrichment analyses. The heat map shows the significant gene sets from the GWAS Catalog, curated genes, and gene ontology (GO) that survived the correction for multiple comparisons. We selectively present the schematics for three pathways: apoptosis, neurogenesis, and nuclear membrane function. Several other key pathways are highlighted in bold, and the 3D maps of the two PSCs are presented.

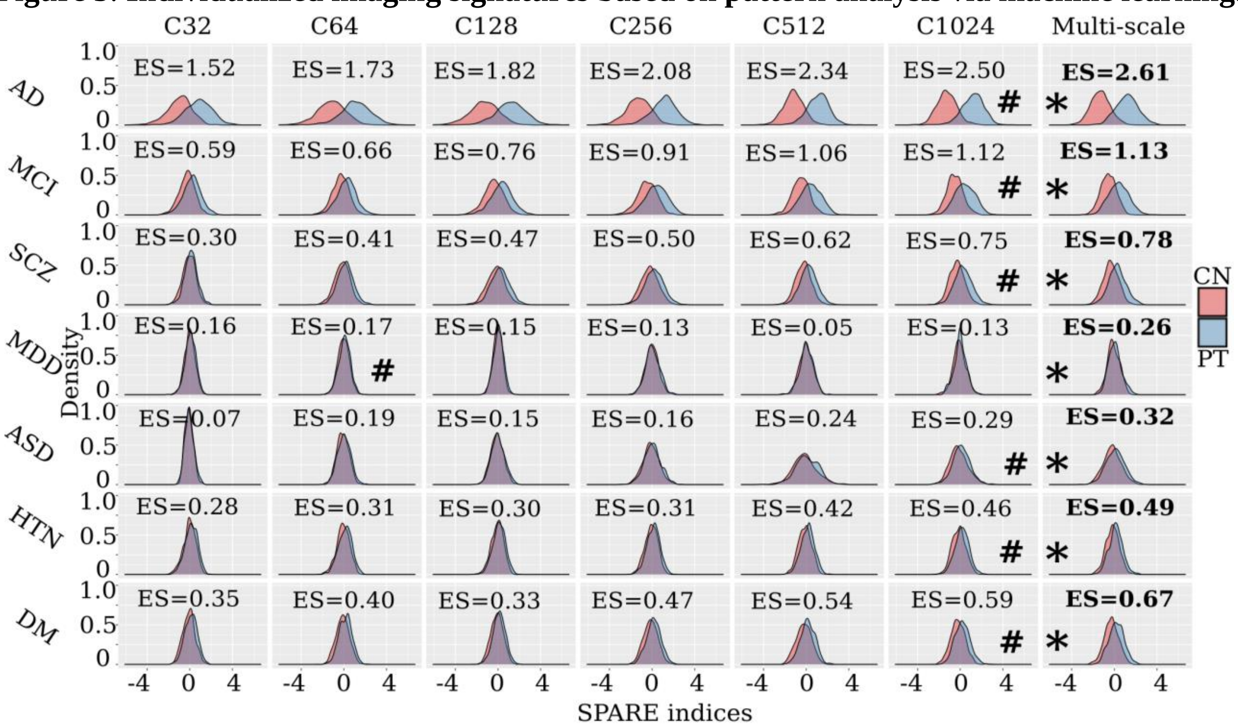
Multi-scale patterns of structural covariance derive disease-related imaging signatures

The diversity of the population we used to derive the PSCs allowed us to derive PSCs that reflect brain development, aging, and the effects of several brain diseases. We investigate the added value of the multi-scale PSCs as building blocks of imaging signatures for several brain diseases and risk conditions using machine learning methods (herein, we opted for linear support vector machines (SVM)) (**Method 5**).[31] The aim is to harness machine learning to drive a clinically interpretable metric for quantifying an individual-level risk to each disease category. To this end, we define the signatures as SPARE-X (Spatial PAtterns for REcognition) indices, where X is the disease. For instance, SPARE-AD captures the degree of expression of an imaging signature of AD-related brain atrophy, which has been shown to offer diagnostic and prognostic value in prior studies.[32]

In our samples, the most discriminative indices were SPARE-AD and SPARE-MCI (**Fig. 5**, **Supplementary eTable 4**, **eFigure 4**). C=1024 achieved the best performance for the single-scale analysis (e.g., AD vs. controls; balanced accuracy: 0.90 ± 0.02; Cohen's d: 2.50). Multi-scale representations derived imaging signatures that showed the largest effect sizes to classify the patients from the controls (**Fig. 5**) (e.g., AD vs. controls; balanced accuracy: 0.92 ± 0.02; Cohen's d: 2.61). PSCs obtained better classification performance than both AAL (e.g., AD vs. controls; balanced accuracy: 0.82 ± 0.02; Cohen's d: 1.81) and voxel-wise regional volumetric maps (RAVENS)[33] (e.g., AD vs. controls; balanced accuracy: 0.85 ± 0.02; Cohen's d: 2.04)

(**Supplementary eTable 4** and **eFigure 6**). Our classification results were higher than previous baseline studies,[34,35] which provided an open-source framework to objectively and reproducibly evaluate AD classification using machine learning. Using the same cross-validation procedure and evaluation metric, they reported the highest balanced accuracy of 0.87 ± 0.02 to classify AD from healthy controls. Notably, our machine learning experiments followed good practices, employed rigorous cross-validation procedures, and avoided critical methodological flaws, such as data leakage or double-dipping (see critical reviews on this topic elsewhere[34,36]).

Figure 5: **Individualized imaging signatures based on pattern analysis via machine learning.**



Imaging signatures (SPARE indices) of brain diseases, derived via supervised machine learning models, are more distinctive when formed from multi-scale PSCs than single-scale PSCs. The kernel density estimate plot depicts the distribution of the patient group (blue) in comparison to the healthy control group (red), reflecting the discriminative power of the diagnosis-specific SPARE (imaging signature) indices. We computed Cohen's *d* for each SPARE index between groups to present the effect size of its discrimination power. ***** represents the model with the largest Cohen's *d* for each SPARE index to separate the control vs. patient groups; # represents the model with the best performance with single-scale PSCs. Our results demonstrate that the multi-scale PSCs generally achieve the largest discriminative effect sizes (ES) (**Supplementary**

eTable 4). As a reference, Cohen's d of ≥ 0.2, ≥ 0.5, and ≥ 0.8 respectively refer to small, moderate, and large effect sizes.

BRIDGEPORT: bridging knowledge across patterns of structural covariance, genomics, and clinical phenotypes

We integrated our experimental results and the MuSIC atlas into the BRIDGEPORT online web portal. This online tool allows researchers to interactively browse the MuSIC atlas in 3D, query our experimental results via variants or PSCs, and download the GWAS summary statistics for further analyses. In addition, we allow users to search via conventional brain anatomical terms (e.g., the right thalamus proper) by automatically annotating traditional anatomic atlas ROIs, specifically from the MUSE atlas[37] (**Supplementary eTable 5**), to MuSIC PSCs based on their degree of overlaps (**Supplementary eFigure 5**). Open-source software dedicated to image processing,[37] genetic quality check protocols, MuSIC generation with sopNMF, and machine learning[34] is also publicly available (see Code Availability for details).

Discussion

The current study investigates patterns of structural covariance in the human brain at multiple scales from a large population of 50,699 people and, importantly, a very diverse cohort allowing us to capture patterns of structural covariance emanating from normal and abnormal brain development and aging, as well as from several brain diseases. Through extensive examination of the genetic architecture of these multi-scale PSCs, we confirmed genetic hits from previous T1-weighted MRI GWAS and, more importantly, identified 617 novel genomic loci and molecular and biological pathways that collectively influence brain morphological changes and development over the lifespan. Using a hypothesis-free, data-driven approach, we elucidated that underlying genetics and biological pathways can form multi-scale structural covariance patterns, which can be further used as building blocks to predict various diseases. All experimental results and code are encapsulated and publicly available in BRIDGEPORT for dissemination: https://www.cbica.upenn.edu/bridgeport/, in order to enable various neuroscience studies to investigate these patterns of structural covariance in diverse contexts. Together, the current study highlighted the adoption of machine learning methods in brain imaging genomics and deepened our understanding of the genetic architecture of the human brain.

Our findings reveal new insights into genetic underpinnings that influence patterns of structural covariance in the human brain. Brain morphological development and changes are largely polygenic and heritable, and previous neuroimaging GWAS has not fully uncovered this genetic landscape. In contrast, genetic variants, as well as environmental, aging, and disease effects, exert pleiotropic effects in shaping morphological changes in different brain regions through specific biological pathways. The mechanisms underlying brain structural covariance are not yet fully understood. They may involve an interplay between common underlying genetic factors, shared susceptibility to aging, and various brain pathologies, which affect brain growth or degeneration in coordinated brain morphological changes.[1] Our data-driven, multi-scale PSCs identify the hierarchical structure of the brain under the principle of structural covariance and are associated with genetic factors at different levels, including SNPs, genes, and gene set pathways. These 617 novel genomic loci, as well as those previously identified, collectively shape brain morphological changes through many key biological and molecular pathways. These pathways are widely involved in reelin signaling, apoptotic processes, axonal development, cellular morphogenesis, neurogenesis, and neuro differentiation,[25,26] which may collectively influence the formation of structural covariance patterns in the brain. Strikingly, pathways involved in breast cancer shared overlapping genetic underpinnings evidenced in our MAGMA-based and prioritized (*GENE2FUNC*) gene set enrichment analyses (**Fig. 3C** and **Fig. 4**), which included specific pathways involved in breast cancer and metastatic breast cancer tumors. One previous study showed that common genes might mediate breast cancer metastasis to the brain,[21] and a later study further corroborated that the metastatic spread of breast cancer to other organs (including the brain) accelerated during sleep in both mouse and human models.[38] We further showcased that this brain metastasis of breast cancer might be associated with specific neuropathologic processes, which were captured by PSCs data driven by Alzheimer's disease-related neuropathology. For example, the hippocampus-entorhinal cortex PSC (C128_3, **Fig. 4**) connected the bilateral hippocampus and medial temporal lobe – the salient hallmark of Alzheimer's disease. Our gene set enrichment analysis results further support this claim: the genes were enriched in the gene sets of Alzheimer's disease and breast cancer (**Fig. 4**). Previous research[22,23] also found an inverse association between Alzheimer's disease and breast cancer. In addition, PSCs from the cerebellum were the most genetically influenced brain regions, consistent with previous neuroimaging GWAS.[4,5] The cerebral cortex has been thought to largely contribute to the unique mental abilities of humans. However, the cerebellum may also be associated with a much more comprehensive range of complex cognitive functions and brain diseases than initially thought.[39] Our results confirmed that many genetic substrates might support different molecular pathways, resulting in the cerebellar functional organization, high-order functions, and dysfunctions in various brain disorders.

The current work demonstrates that appropriate machine learning analytics can be used to shed new light on brain imaging genetics. Previous neuroimaging GWAS leveraged multimodal imaging-derived phenotypes from conventional brain atlases[4,5] (e.g., ROIs from the AAL atlas). In contrast, multi-scale PSCs are purely data-driven and likely to reflect the dynamics of underlying normal and pathological neurobiological processes giving rise to structural covariance. The diverse training sample from which the PSCs were derived, including healthy and diseased individuals of a wide age range, enriched the diversity of such neurobiological processes influencing the PSCs. In addition, modeling structural covariance at multiple scales (i.e., multi-scale PSCs) indicated that disease effects could be robustly and complementarily identified across scales (**Fig. 6**), concordant with the paradigm of multi-scale modeling of the brain.[12] Imaging signatures of brain diseases, derived via supervised machine learning models, were consistently more distinctive when formed from multi-scale PSCs compared to single-scale PSCs.

MuSIC – with the strengths of being data-driven, multi-scale, and disease-effect informative – contributes to the century-old quest for a "universal" atlas in brain cartography[40] and is highly complementary to previously proposed brain atlases. For instance, Chen and colleagues[41] used a semi-automated fuzzy clustering technique with MRI data from 406 twins and parcellated the cortical surface area into a genetic covariance-informative brain atlas; MuSIC was data-driven by structural covariance. Glasser and colleagues[42] adopted a semi-automated parcellation procedure to create a multimodal cortex atlas from 210 healthy individuals. Although this method was successful in integrating multimodal information from cortical folding, myelination, and functional connectivity, this semi-automatic approach requires significant resources, some with limited resolution. MuSIC allows flexible, multiple scales for delineating macroscopic brain topology; the inclusion of patient samples exposes the model to sources of variability that may not be visible in healthy controls. Another pioneering endeavor is the Allen Brain Atlas project,[43] whose overarching goals of mapping the human brain to gene expression data via existing conventional atlases, identifying local gene expression patterns across the brain in a few individuals, and deepening our understanding of the human brain's differential genetic architecture, are complementary to ours – characterizing the global genetic architecture of the human brain, emphasizing pathogenic variability and morphological heterogeneity.

Bridging knowledge across the brain imaging, genomics, and machine learning communities is another pivotal contribution of this work. BRIDGEPORT provides a platform to lower the entry barrier for whole-brain genetic-structural analyses, foster interdisciplinary communication, and advocate for research reproducibility.[34,44–47] The current study demonstrates the broad applicability of this large-scale, multi-omics platform across a spectrum of neurodegenerative and neuropsychiatric diseases.

Methods

Method 1: Structural covariance patterns via stochastic orthogonally projective non-negative matrix factorization

The sopNMF algorithm is a stochastic approximation built and extended based on opNMF[9,48]. We consider a dataset of n MR images and d voxels per image. We represent the data as a matrix $\boldsymbol{X}$ where each column corresponds to a flattened image: $\boldsymbol{X} = [x_1, x_2, \dots, x_n], \boldsymbol{X} \in \mathbb{R}_{\geq 0}^{d \times n}$. The sopNMF algorithm factorizes $\boldsymbol{X}$ into two low-rank (r) matrices $\boldsymbol{W} \in \mathbb{R}_{\geq 0}^{d \times r}$ and $\boldsymbol{H} \in \mathbb{R}_{\geq 0}^{r \times n}$ under the constraints of non-negativity and column-orthonormality. Using the Frobenius norm, the loss of this factorization problem can be formulated as

$$\|\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}\|_F^2$$

$$\text{subject to } \boldsymbol{H} = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}\,, \boldsymbol{W} \geq 0 \text{ and } \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W} = \boldsymbol{I} \quad (1)$$

where $\boldsymbol{I}$ stands for the identity matrix. The columns $w_i \in \mathbb{R}^d$, $\|w_i\|^2 = 1, \forall\, i \in \{1..r\}$ of the so-called component matrix $\boldsymbol{W} = [w_1, w_2, \dots, w_r]$ are part-based representations promoting sparsity in data in this lower-dimensional subspace. From this perspective, the loading coefficient matrix $\boldsymbol{H}$ represents the importance (weights) of each of the features above for a given image. Instead of optimizing the non-convex problem in a batch learning paradigm (i.e., reading all images into memory) as opNMF,[9] sopNMF subsamples the number of images at each iteration, thereby significantly reducing its memory demand, by randomly drawing data batches $\boldsymbol{X_b} \in \mathbb{R}_{\geq 0}^{d \times b}$ of $b \leq n$ images (b is the batch size); this is done without replacement so that all data goes through the model once ($\lceil n/b \rceil$). In this case, the updating rule can be rewritten as

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t \frac{\left(\boldsymbol{X_b}\boldsymbol{X_b^T}\boldsymbol{W}\right)_t}{(\boldsymbol{W}\boldsymbol{W}^T\boldsymbol{X_b}\boldsymbol{X_b^T}\boldsymbol{W})_t} \quad (2)$$

We calculate the loss on the entire dataset at the end of each epoch (i.e., the loss is incremental across all batches) with the following expression

$$\sum_{i=1}^{\lceil n/b \rceil} \left\| \boldsymbol{X}_{\boldsymbol{b\_i}} - \boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{bi}} \right\|_F^2 \quad (3)$$

We evaluated the training loss and the sparsity of $\boldsymbol{W}$ at the end of each iteration. Moreover, early stopping was implemented to improve training efficiency and alleviate overfitting. We summarize the sopNMF algorithm in **Supplementary Algorithm 1**. An empirical comparison between sopNMF and opNMF is detailed in **Supplementary eMethod 1**.

We applied sopNMF to the training population (*N*=4000). After the algorithm converged, the component matrix $\boldsymbol{W}$ was sparse and the loading coefficient matrix was used as the multi-scale PSCs. To build the MuSIC atlas, we clustered each voxel (row-wise) into one of the r features/PSCs as follows:

$$\boldsymbol{M}_j = \text{argmax}_k(\boldsymbol{W}_{j,k}) \; (4)$$

where $\boldsymbol{M}$ is a *d*-dimensional vector and $j \in \{1..d\}$. The *j*-th element of $\boldsymbol{M}$ equals k if $\boldsymbol{W}_{j,k}$ is the maximum value of the *j*-th row. We finally projected the vector $\boldsymbol{M} \in \mathbb{R}^d_{\geq 0}$ into the original image space to visualize each PSC of the MuSIC atlas (**Fig. 1**). Of note, 13 PSCs have vanished in this process for *C*=1024: all 0 for these 13 vectors.

Method 2: Study population

We consolidated a large-scale multimodal consortium (*N*=50,699) consisting of imaging, cognition, and genetic data from 12 studies, 130 sites, and 12 countries (**Supplementary eTable 1**): the Alzheimer's Disease Neuroimaging Initiative[49] (ADNI) (*N*=1765); the UK Biobank[50] (UKBB) (*N*=39,564); the Australian Imaging, Biomarker, and Lifestyle study of aging[51] (AIBL)

(*N*=830); the Biomarkers of Cognitive Decline Among Normal Individuals in the Johns Hopkins University[52] (BIOCARD) (*N*=288); the Baltimore Longitudinal Study of Aging[53,54] (BLSA) (*N*=1114); the Coronary Artery Risk Development in Young Adults[55] (CARDIA) (*N*=892); the Open Access Series of Imaging Studies[56] (OASIS) (*N*=983), PENN (*N*=807); the Women's Health Initiative Memory Study[57] (WHIMS) (*N*=995), the Wisconsin Registry for Alzheimer's Prevention[58] (WRAP) (*N*=116); the Psychosis Heterogeneity (evaluated) via dimEnsional NeurOiMaging[59] (PHENOM) (*N*=2125); and the Autism Brain Imaging Data Exchange[60] (ABIDE) (*N*=1220). All studies were approved by the corresponding Institutional Review Boards. Each participant consented to be part of the imaging, cognition, and/or genetic biobanks. Data from the UKBB for this project pertains to application 35148.

We present the demographic information of the population under study in **Supplementary eTable 1**. This large-scale consortium reflects the diversity of MRI scans over different races, disease conditions, and ages over the lifespan. To be concise, we defined four populations or data sets per analysis across the paper:

- Discovery set*:* It consists of a multi-disease and lifespan population that includes participants from all 12 studies (*N*=32,440). Note that this population does not contain the entire UKBB population but only our first download (July 2017, *N*=21,305).
- Replication set: We held out 18,259 participants from the UKBB dataset to replicate the GWAS results. We took these data from our second download of the UKBB dataset (November 2021, *N*=18,259).
- Training population: We randomly drew 250 patients (PT), including AD, MCI, SCZ, ASD, MDD, HTN (hypertension), DM (diabetes mellitus), and 250 healthy controls (CN) per decade from the discovery set, ensuring that the PT and CN groups have

similar sex, study and age distributions. The resulting set of 4000 imaging data was used to generate the MuSIC atlas with the sopNMF algorithm. The rationale is to maximize variability across a balanced sample of multiple diseases or risk conditions, age, and study protocols rather than overfit the entire data by including all images in training.

- Comparison population: To validate sopNMF by comparison to the original opNMF algorithm, we randomly subsampled 800 participants from the training population (100 per decade for balanced CN and PT). For this scale of sample size, opNMF can load all images into memory for batch learning.[61]

Method 3: Image processing and statistical harmonization

(A): Image processing. Images that passed the quality check (**Supplementary eMethod 4**) were first corrected for magnetic field intensity inhomogeneity.[62] Voxel-wise regional volumetric maps (RAVENS)[33] for each tissue volume were then generated by using a registration method to spatially align the skull-stripped images to a template in MNI-space.[63] We applied sopNMF to the RAVENS maps to derive MuSIC.

(B): Statistical harmonization of MuSIC PSCs: We applied MuSIC to the entire population (N=50,699) to extract the multi-scale PSCs. Specifically, MuSIC was applied to each individual's RAVENS gray matter map to extract the sum of brain volume in each PSC. Subsequently, the PSCs were statistically harmonized by an extensively validated approach, i.e., ComBat-GAM [11] (**Supplementary eMethod 3**). After harmonization, the PSCs were considerably normally distributed (skewness = 0.11 ± 0.17, and kurtosis = 0.67 ± 0.68) (**Supplementary eFigure 6A** and **B**). To alleviate the potential violation of normal distribution in downstream statistical learning, we quantile-transformed all PSCs. In agreement with the literature,[64,65] males were found to have larger brain volumes than females on average (**Supplementary eFigure 6C**). The AAL ROIs underwent the same statistical harmonization procedure.

Method 4: Genetic analyses

Genetic analyses were restricted to the discovery and replication set from UKBB (**Method 2**). We processed the array genotyping and imputed genetic data (SNPs). The two data sets went through a "best-practice" imaging-genetics quality check (QC) protocol (**Method 4A**) and were restricted to participants of European ancestry. This resulted in 18,052 participants and 8,430,655 SNPs for the discovery set, and 15,243 participants and 8,470,709 SNPs for the replication set. We reperformed the genetic QC and genetic analyses for the combined populations for BRIDGEPORT, resulting in 33,541 participants and 8,469,833 SNPs. **Method 4G** details the correction for multiple comparisons throughout our analyses.

(A): Genetic data quality check protocol. First, we excluded related individuals (up to 2nd-degree) from the complete UKBB sample (*N*=488,377) using the KING software for family relationship inference.[66] We then removed duplicated variants from all 22 autosomal chromosomes. We also excluded individuals for whom either imaging or genetic data were not available. Individuals whose genetically-identified sex did not match their self-acknowledged sex were removed. Other excluding criteria were: i) individuals with more than 3% of missing genotypes; ii) variants with minor allele frequency (MAF) of less than 1%; iii) variants with larger than 3% missing genotyping rate; iv) variants that failed the Hardy-Weinberg test at 1×10^{-10}. To adjust for population stratification,[67] we derived the first 40 genetic principle components (PC) using the FlashPCA software[68]. The genetic pipeline was described elsewhere,[69] and documented online: https://www.cbica.upenn.edu/bridgeport/data/pdf/BIGS_genetic_protocol.pdf.

(B): Heritability estimates and genome-wide association analysis. We estimated the SNP-based heritability explained by all autosomal genetic variants using GCTA-GREML.[70] We adjusted for confounders of age (at imaging), age-squared, sex, age-sex interaction, age-squared-sex interaction, ICV, and the first 40 genetic principal components (PC). One-side likelihood ratio tests were performed to derive the heritability estimates. In GWAS, we performed a linear regression for each PSC and included the same covariates as in the heritability estimates using PLINK.[71]

(C): Identification of novel genomic loci. Using PLINK, we clumped the GWAS summary statistics based on their linkage disequilibrium to identify the genomic loci (see **Supplementary eMethod 5** for the definition of the index, candidate, independent significant, lead SNP, and genomic locus). In particular, the threshold for significance was set to 5×10^{-8} (*clump-p1*) for the index SNPs and 0.05 (*clump-p2*) for the **candidate SNPs**. The threshold for linkage disequilibrium-based clumping was set to 0.60 (*clump-r2*) for **independent significant SNPs** and 0.10 for lead SNPs. The linkage disequilibrium physical-distance threshold was 250 kilobases (*clump-kb*). **Genomic loci** take into account linkage disequilibrium (within 250 kilobases) when interpreting the association results. The GWASRAPIDD[72] software was then used to query the genomic loci for any previously-reported associations with clinical phenotypes documented in the NHGRI-EBI GWAS Catalog[14] (p-value < 1.0×10^{-5}, default inclusion value of

GWAS Catalog). We defined a genomic locus as **novel** when it was not present in GWAS Catalog.

(D): Gene-level associations with MAGMA. We performed gene-level association analysis using MAGMA.[15] First, gene annotation was performed to map the SNPs (reference variant location from Phase 3 of 1,000 Genomes for European ancestry) to genes (human genome Build 37) according to their physical positions. The second step was to perform the gene analysis based on the GWAS summary statistics to obtain gene-level p-values between the pairwise of 2003 PSCs and the 18,097 protein-encoding genes containing valid SNPs.

(E): Hypothesis-free gene set enrichment analysis with MAGMA. Using the gene-level association p-values, we performed gene set enrichment analysis using MAGMA. Gene sets were obtained from Molecular Signatures Database (MsigDB, v7.5.1),[73] including 6366 curated gene sets and 10,402 Gene Ontology (GO) terms. All other parameters were set by default for MAGMA. This hypothesis-free analysis resulted in a more stringent correction for multiple comparisons (i.e., by the total number of tested genes and the number of PSCs) than the FUMA prioritized gene set enrichment analysis (see below F).

(F): FUMA analyses for the illustrations of specific PSCs. In *SNP2GENE*, three different methods were used to map the SNPs to genes. First, positional mapping maps SNPs to genes if the SNPs are physically located inside a gene (a 10 kb window by default). Second, expression quantitative trait loci (eQTL) mapping maps SNPs to genes showing a significant eQTL association. Lastly, chromatin interaction mapping maps SNPs to genes when there is a significant chromatin interaction between the disease-associated regions and nearby or distant genes.[24] In addition, *GENE2FUNC* studies the expression of prioritized genes and tests for enrichment of the set of genes in pre-defined pathways. We used the mapped genes as prioritized genes. The background genes were specified as all genes in FUMA, and all other parameters were set by defaults. We only reported gene sets with adjusted p-value < 0.05.

(G): Correction for multiple comparisons. We practiced a conservative procedure to control for the multiple comparisons. In the case of GWAS, we chose the default genome-wide significant threshold (5.0×10^{-8}, and 0.05 for all other analyses) and independently adjusted for multiple comparisons (Bonferroni methods) at each scale by the number of PSCs. We corrected the p-values for the number of phenotypes (*N*=6) for genetic correlation analyses. For heritability estimates, we adjusted the p-values for the number of PSCs at each scale. For gene analyses, we controlled for both the number of PSCs at each scale and the number of genes. We adopted these strategies per analysis to correct the multiple comparisons because PSCs of different scales are likely hierarchical and correlated – avoiding the potential of "overcorrection".

(H): Replication analysis for genome-wide association studies. We performed GWAS by fitting the same linear regressing models as the discovery set. Also, following the same procedure for consistency, we corrected for the multiple comparisons using the Bonferroni method. In particular, we corrected it for the number of genomic loci (*N*=915) found in the discovery set with a nominal p-value of 0.05, which thereby resulted in a stringent test with an equivalent p-value threshold of 3.1×10^{-5} (i.e., ($-\log_{10}$[p-value] = 4.27). We performed a replication for the 915 genomic loci, but, in reality, SNPs in linkage disequilibrium with the genomic loci are likely highly significant.

Method 5: Pattern analysis via machine learning for individualized imaging signatures

SPARE-AD captures the degree of expression of an imaging signature of AD, and prior studies have shown its diagnostic and prognostic values.[32] We generalized the SPARE imaging signature to multiple diseases (SPARE-X, X represents disease diagnoses). Following our reproducible open-source framework[35], we performed nested cross-validation (**Supplementary eMethod 6**) for the machine learning models and derived imaging signatures to quantify individualized disease vulnerability.

SPARE indices. MuSIC PSCs were fit into a linear support vector machine (SVM) to derive SPARE-AD, MCI, SCZ, DM, HTN, MDD, and ASD. Specifically, the SVM aims to classify the patient group (e.g., AD) from the control group and outputs a decision function that indicates how close each participant is to the hyperplane. The samples selected for each task are presented in **Supplementary eTable 2**.

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

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 GWAS summary statistics used in the genetic correlation analyses were fetched from the GWAS Catalog platform (https://www.ebi.ac.uk/gwas), although each study provided the original links; The GWAS Catalog platform was used to query if the SNPs identified by MuSIC were previously reported.

Code Availability

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

- sopNMF: https://pypi.org/project/sopnmf/, MuSIC, and sopNMF (developed for this study)
- BRIDGEPORT: https://www.cbica.upenn.edu/bridgeport/, (developed for this study)
- BIGS: https://www.cbica.upenn.edu/bridgeport/data/pdf/BIGS_genetic_protocol.pdf, genetic processing protocol (developed for this study)
- MLNI: https://pypi.org/project/mlni/, machine learning (developed for this study)
- MUSE: https://www.med.upenn.edu/sbia/muse.html, image preprocessing
- Clinica: http://www.clinica.run/, machine learning and image preprocessing
- PLINK: https://www.cog-genomics.org/plink/, GWAS
- GCTA: https://yanglab.westlake.edu.cn/software/gcta/#Overview, heritability estimates
- LDSC: https://github.com/bulik/ldsc, genetic correlation estimates
- MAGMA: https://ctg.cncr.nl/software/magma, gene analysis
- GWASRAPIDD: https://rmagno.eu/gwasrapidd/articles/gwasrapidd.html, GWAS Catalog query
- MsigDB: https://www.gsea-msigdb.org/gsea/msigdb/, gene sets database

Competing Interests

DAW served as Site PI for studies by Biogen, Merck, and Eli Lilly/Avid. He has received consulting fees from GE Healthcare and Neuronix. He is on the DSMB for a trial sponsored by Functional Neuromodulation. AJS receives support from multiple NIH grants (P30 AG010133, P30 AG072976, R01 AG019771, R01 AG057739, U01 AG024904, R01 LM013463, R01 AG068193, T32 AG071444, and U01 AG068057 and U01 AG072177). He has also received support from Avid Radiopharmaceuticals, a subsidiary of Eli Lilly (in-kind contribution of PET tracer precursor); Bayer Oncology (Scientific Advisory Board); Eisai (Scientific Advisory Board); Siemens Medical Solutions USA, Inc. (Dementia Advisory Board); Springer-Nature Publishing (Editorial Office Support as Editor-in-Chief, Brain Imaging, and Behavior). OC reports having received consulting fees from AskBio (2020) and Therapanacea (2022), having received payments for writing a lay audience short paper from Expression Santé (2019), and that his laboratory has received grants (paid to the institution) from Qynapse (2017-present). Members of his laboratory have co-supervised a Ph.D. thesis with myBrainTechnologies (2016-2019) and with Qynapse (2017-present). OC's spouse is an employee and holds stock options of myBrainTechnologies (2015-present). OC has a patent registered at the International Bureau of the World Intellectual Property Organization (PCT/IB2016/0526993, Schiratti J-B, Allassonniere S, Colliot O, Durrleman S, A method for determining the temporal progression of a biological phenomenon and associated methods and devices) (2017). ME receives support from multiple NIH grants, the Alzheimer's Association, and the Alzheimer's Therapeutic Research Institute. MZ serves as a consultant and/or speaker for the following pharmaceutical companies: Eurofarma, Lundbeck, Abbott, Greencare, Myralis, and Elleven Healthcare.

Authors' contributions

Dr. Wen takes full responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Wen, Nasrallah, Davatzikos

Acquisition, analysis, or interpretation of data: Wen, Nasrallah, Davatzikos, Abdulkadi, Satterthwaite, Dazzan, Kahn, Schnack, Zanetti, Meisenzahl, Busatto, Crespo-Facorro, Pantelis, Wood, Zhuo, Koutsouleris, Wittfeld, Grabe, Marcus, LaMontagne, Heckbert, Austin, Launer, Espeland, Masters, Maruff, Fripp, Johnson, Morris, Albert, Resnick, Saykin, Thompson, Li, Wolf, Raquel Gur, Ruben Gur, Shinohara, Tosun-Turgut, Fan, Shou, Erus, Wolk

Drafting of the manuscript: Wen, Nasrallah, Davatzikos

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

Statistical and genetic analysis: Wen

Study supervision: Davatzikos

References

1. Alexander-Bloch, A., Giedd, J. N. & Bullmore, E. Imaging structural covariance between human brain regions. *Nat Rev Neurosci* **14**, 322–336 (2013).
2. Sotiras, A. *et al.* Patterns of coordinated cortical remodeling during adolescence and their associations with functional specialization and evolutionary expansion. *Proc Natl Acad Sci USA* **114**, 3527–3532 (2017).
3. Blank, S. C., Scott, S. K., Murphy, K., Warburton, E. & Wise, R. J. S. Speech production: Wernicke, Broca and beyond. *Brain* **125**, 1829–1838 (2002).
4. 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).
5. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* **562**, 210–216 (2018).
6. Vignando, M. *et al.* Mapping brain structural differences and neuroreceptor correlates in Parkinson's disease visual hallucinations. *Nat Commun* **13**, 519 (2022).
7. Bassett, D. S. & Siebenhühner, F. Multiscale Network Organization in the Human Brain. in *Multiscale Analysis and Nonlinear Dynamics* 179–204 (John Wiley & Sons, Ltd, 2013). doi:10.1002/9783527671632.ch07.
8. Schaefer, A. *et al.* Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex* **28**, 3095–3114 (2018).
9. Sotiras, A., Resnick, S. M. & Davatzikos, C. Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization. *NeuroImage* **108**, 1–16 (2015).

10. Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L. & Greicius, M. D. Neurodegenerative diseases target large-scale human brain networks. *Neuron* **62**, 42–52 (2009).
11. Pomponio, R. *et al.* Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. *Neuroimage* **208**, 116450 (2020).
12. Betzel, R. F. & Bassett, D. S. Multi-scale brain networks. *NeuroImage* **160**, 73–83 (2017).
13. Roshchupkin, G. V. *et al.* Heritability of the shape of subcortical brain structures in the general population. *Nat Commun* **7**, 13738 (2016).
14. 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).
15. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).
16. Jossin, Y. Reelin Functions, Mechanisms of Action and Signaling Pathways During Brain Development and Maturation. *Biomolecules* **10**, E964 (2020).
17. Gilbert, S. F. Morphogenesis and Cell Adhesion. *Developmental Biology. 6th edition* (2000).
18. Wu, Y. *et al.* Contacts between the endoplasmic reticulum and other membranes in neurons. *Proc Natl Acad Sci U S A* **114**, E4859–E4867 (2017).
19. Ly, A. *et al.* DSCAM Is a Netrin Receptor that Collaborates with DCC in Mediating Turning Responses to Netrin-1. *Cell* **133**, 1241–1254 (2008).
20. Nikolsky, Y. *et al.* Genome-wide functional synergy between amplified and mutated genes in human breast cancer. *Cancer Res* **68**, 9532–9540 (2008).

21. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459**, 1005–1009 (2009).
22. Lanni, C., Masi, M., Racchi, M. & Govoni, S. Cancer and Alzheimer's disease inverse relationship: an age-associated diverging derailment of shared pathways. *Mol Psychiatry* **26**, 280–295 (2021).
23. Shafi, O. Inverse relationship between Alzheimer's disease and cancer, and other factors contributing to Alzheimer's disease: a systematic review. *BMC Neurol* **16**, 236 (2016).
24. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
25. Yuan, J. & Yankner, B. A. Apoptosis in the nervous system. *Nature* **407**, 802–809 (2000).
26. Steiner, E., Tata, M. & Frisén, J. A fresh look at adult neurogenesis. *Nat Med* **25**, 542–543 (2019).
27. de Flores, R. *et al.* Medial Temporal Lobe Networks in Alzheimer's Disease: Structural and Molecular Vulnerabilities. *J Neurosci* **42**, 2131–2141 (2022).
28. Ginestier, C. *et al.* Prognosis and gene expression profiling of 20q13-amplified breast cancers. *Clin Cancer Res* **12**, 4533–4544 (2006).
29. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet* **51**, 414–430 (2019).
30. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
31. Davatzikos, C. Machine learning in neuroimaging: Progress and challenges. *NeuroImage* **197**, 652–656 (2019).

32. Davatzikos, C., Xu, F., An, Y., Fan, Y. & Resnick, S. M. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain* **132**, 2026–2035 (2009).
33. Davatzikos, C., Genc, A., Xu, D. & Resnick, S. M. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. *Neuroimage* **14**, 1361–1369 (2001).
34. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis* **63**, 101694 (2020).
35. 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).
36. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12**, 535–540 (2009).
37. 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).
38. Diamantopoulou, Z. *et al.* The metastatic spread of breast cancer accelerates during sleep. *Nature* **607**, 156–162 (2022).
39. Barton, R. A. & Venditti, C. Rapid Evolution of the Cerebellum in Humans and Other Great Apes. *Curr Biol* **27**, 1249–1250 (2017).
40. Eickhoff, S. B., Yeo, B. T. T. & Genon, S. Imaging-based parcellations of the human brain. *Nat Rev Neurosci* **19**, 672–686 (2018).

41. Chen, C.-H. *et al.* Hierarchical Genetic Organization of Human Cortical Surface Area. *Science* **335**, 1634–1636 (2012).

42. Glasser, M. F. *et al.* A multimodal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).

43. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res* **41**, D996–D1008 (2013).

44. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat Hum Behav* **1**, 1–9 (2017).

45. Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).

46. Routier, A. *et al.* Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies. *Frontiers in Neuroinformatics* **15**, 39 (2021).

47. Wen, J. *et al.* Reproducible Evaluation of Diffusion MRI Features for Automatic Classification of Patients with Alzheimer's Disease. *Neuroinformatics* **19**, 57–78 (2021).

48. Zhirong Yang & Oja, E. Linear and Nonlinear Projective Nonnegative Matrix Factorization. *IEEE Trans. Neural Netw.* **21**, 734–749 (2010).

49. Petersen, R. C. *et al.* Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology* **74**, 201–209 (2010).

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

51. Ellis, K. A. *et al.* The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int. Psychogeriatr.* **21**, 672–687 (2009).

52. Soldan, A. *et al.* Relationship of medial temporal lobe atrophy, APOE genotype, and cognitive reserve in preclinical Alzheimer's disease. *Hum Brain Mapp* **36**, 2826–2841 (2015).
53. Resnick, S. M., Pham, D. L., Kraut, M. A., Zonderman, A. B. & Davatzikos, C. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *J Neurosci* **23**, 3295–3301 (2003).
54. Resnick, S. M. *et al.* One-year age changes in MRI brain volumes in older adults. *Cereb Cortex* **10**, 464–472 (2000).
55. Friedman, G. D. *et al.* CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J Clin Epidemiol* **41**, 1105–1116 (1988).
56. LaMontagne, P. J. *et al.* OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. 2019.12.13.19014902 Preprint at https://doi.org/10.1101/2019.12.13.19014902 (2019).
57. Resnick, S. M. *et al.* Postmenopausal hormone therapy and regional brain volumes: the WHIMS-MRI Study. *Neurology* **72**, 135–142 (2009).
58. Johnson, S. C. *et al.* The Wisconsin Registry for Alzheimer's Prevention: A review of findings and current directions. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **10**, 130–142 (2018).
59. Chand, G. B. *et al.* Two distinct neuroanatomical subtypes of schizophrenia revealed using machine learning. *Brain* **143**, 1027–1038 (2020).
60. Di Martino, A. *et al.* Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data* **4**, 170010 (2017).

61. Shalev-Shwartz, S., Singer, Y. & Ng, A. Y. Online and batch learning of pseudo-metrics. in *Proceedings of the twenty-first international conference on Machine learning* 94 (Association for Computing Machinery, 2004). doi:10.1145/1015330.1015376.
62. Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
63. Ou, Y., Sotiras, A., Paragios, N. & Davatzikos, C. DRAMMS: Deformable Registration via Attribute Matching and Mutual-Saliency Weighting. *Med Image Anal* **15**, 622–639 (2011).
64. Coupé, P., Catheline, G., Lanuza, E., Manjón, J. V. & Initiative, for the A. D. N. Towards a unified analysis of brain maturation and aging across the entire lifespan: A MRI analysis. *Human Brain Mapping* **38**, 5501–5518 (2017).
65. Bethlehem, R. a. I. *et al. Brain charts for the human lifespan*. 2021.06.08.447489 https://www.biorxiv.org/content/10.1101/2021.06.08.447489v1 (2021) doi:10.1101/2021.06.08.447489.
66. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
67. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459–463 (2010).
68. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
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. Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. *PNAS* **113**, E4579–E4580 (2016).

71. 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).

72. Magno, R. & Maia, A.-T. gwasrapidd: an R package to query, download and wrangle GWAS catalog data. *Bioinformatics* **36**, 649–650 (2020).

73. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

Acknowledgments

The iSTAGING consortium is a multi-institutional effort funded by NIA by RF1 AG054409. The Baltimore Longitudinal Study of Aging neuroimaging study is funded by the Intramural Research Program, National Institute on Aging, National Institutes of Health and by HHSN271201600059C. The BIOCARD study is in part supported by NIH grant U19-AG033655. The PHENOM study is funded by NIA grant R01MH112070 and by the PRONIA project as funded by the European Union 7th Framework Program grant 602152. Other supporting funds are 5U01AG068057, 1U24AG074855, and S10OD023495. Data were provided [in part] by OASIS OASIS-3: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly. OC has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6). This research has been conducted using the UK Biobank Resource under Application Number 35148. Data used in preparation of this article were in part obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Dr. Wen and Dr. Davatzikos had full access to all the data in the study. They took responsibility for the integrity of the data and the accuracy of the data analysis.