



HAL
open science

Genomic loci influence patterns of structural covariance in the human brain

Junhao Wen, Ilya M Nasrallah, Ahmed Abdulkadir, Theodore D Satterthwaite, Zhijian Yang, Guray Erus, Timothy Robert-Fitzgerald, Ashish Singh, Aristeidis Sotiras, Aleix Boquet-Pujadas, et al.

► **To cite this version:**

Junhao Wen, Ilya M Nasrallah, Ahmed Abdulkadir, Theodore D Satterthwaite, Zhijian Yang, et al.. Genomic loci influence patterns of structural covariance in the human brain. Proceedings of the National Academy of Sciences of the United States of America, 2023, 120 (52), 10.1073/pnas.2300842120 . hal-04362321

HAL Id: hal-04362321

<https://hal.science/hal-04362321>

Submitted on 22 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 Genomic loci influence patterns of structural covariance in the human brain

2
3 Junhao Wen^{1,2*}, Ilya M. Nasrallah^{2,3}, Ahmed Abdulkadir², Theodore D. Satterthwaite^{2,4}, Zhijian Yang²,
4 Guray Erus², Timothy Robert-Fitzgerald⁵, Ashish Singh², Aristeidis Sotiras⁶, Aleix Boquet-Pujadas⁷,
5 Elizabeth Mamourian², Jimit Doshi², Yuhua Cui², Dhivya Srinivasan², Ioanna Skampardoni², Jiong
6 Chen², Gyujoon Hwang², Mark Bergman², Jingxuan Bao⁸, Yogasudha Veturi⁹, Zhen Zhou², Shu Yang⁸,
7 Paola Dazzan¹⁰, Rene S. Kahn¹¹, Hugo G. Schnack¹², Marcus V. Zanetti¹³, Eva Meisenzahl¹⁴, Geraldo F.
8 Busatto¹³, Benedicto Crespo-Facorro¹⁵, Christos Pantelis¹⁶, Stephen J. Wood¹⁷, Chuanjun Zhuo¹⁸, Russell
9 T. Shinohara^{2,5}, Ruben C. Gur⁴, Raquel E. Gur⁴, Nikolaos Koutsouleris¹⁹, Daniel H. Wolf^{2,4}, Andrew J.
10 Saykin²⁰, Marylyn D. Ritchie⁹, Li Shen⁸, Paul M. Thompson²¹, Olivier Colliot²², Katharina Wittfeld²³,
11 Hans J. Grabe²³, Duygu Tosun²⁴, Murat Bilgel²⁵, Yang An²⁵, Daniel S. Marcus²⁶, Pamela LaMontagne²⁶,
12 Susan R. Heckbert²⁷, Thomas R. Austin²⁷, Lenore J. Launer²⁸, Mark Espeland²⁹, Colin L Masters³⁰, Paul
13 Maruff³⁰, Jurgen Fripp³¹, Sterling C. Johnson³², John C. Morris³³, Marilyn S. Albert³⁴, R. Nick Bryan³,
14 Susan M. Resnick²⁵, Yong Fan², Mohamad Habes³⁵, David Wolk^{2,36}, Haochang Shou^{2,5}, and Christos
15 Davatzikos^{2*}

16
17 ¹Laboratory of AI and Biomedical Science (LABS), Stevens Neuroimaging and Informatics Institute, Keck School of
18 Medicine of USC, University of Southern California, Los Angeles, California, USA.

19 ²Artificial Intelligence in Biomedical Imaging Laboratory (AIBIL), Center for Biomedical Image Computing and
20 Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA.

21 ³Department of Radiology, University of Pennsylvania, Philadelphia, USA.

22 ⁴Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

23 ⁵Penn Statistics in Imaging and Visualization Center, Department of Biostatistics, Epidemiology, and Informatics,
24 Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

25 ⁶Department of Radiology and Institute for Informatics, Washington University School of Medicine, St. Louis, USA

26 ⁷Biomedical Imaging Group, EPFL, Lausanne, Switzerland

27 ⁸Department of Biostatistics, Epidemiology and Informatics University of Pennsylvania Perelman School of Medicine,
28 Philadelphia, USA

29 ⁹Department of Genetics and Institute for Biomedical Informatics, Perelman School of Medicine, University of
30 Pennsylvania, Philadelphia, PA, USA

31 ¹⁰Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College
32 London, London, UK

33 ¹¹Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA

34 ¹²Department of Psychiatry, University Medical Center Utrecht, Utrecht, Netherlands

35 ¹³Institute of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, Brazil

36 ¹⁴Department of Psychiatry and Psychotherapy, HHU Düsseldorf, Germany

37 ¹⁵Hospital Universitario Virgen del Rocío, University of Sevilla-IBIS; IDIVAL-CIBERSAM, Sevilla, Spain

38 ¹⁶Melbourne Neuropsychiatry Centre, Department of Psychiatry, University of Melbourne and Melbourne Health,
39 Carlton South, Australia

40 ¹⁷Orygen and the Centre for Youth Mental Health, University of Melbourne; and the School of Psychology,
41 University of Birmingham, UK

42 ¹⁸Key Laboratory of Real Time Tracing of Brain Circuits in Psychiatry and Neurology (RTBCPN-Lab), Nankai
43 University Affiliated Tianjin Fourth Center Hospital; Department of Psychiatry, Tianjin Medical University, Tianjin,
44 China

45 ¹⁹Department of Psychiatry and Psychotherapy, Ludwig-Maximilian University, Munich, Germany

46 ²⁰Radiology and Imaging Sciences, Center for Neuroimaging, Department of Radiology and Imaging Sciences,
47 Indiana Alzheimer's Disease Research Center and the Melvin and Bren Simon Cancer Center, Indiana University
48 School of Medicine, Indianapolis

49 ²¹Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of
50 Medicine of USC, University of Southern California, Marina del Rey, California

51 ²²Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la
52 Pitié Salpêtrière, F-75013, Paris, France

53 ²³Department of Psychiatry and Psychotherapy, German Center for Neurodegenerative Diseases (DZNE), University
54 Medicine Greifswald, Germany
55 ²⁴Department of Radiology and Biomedical Imaging, University of California, San Francisco, CA, USA
56 ²⁵Laboratory of Behavioral Neuroscience, National Institute on Aging, NIH, USA
57 ²⁶Department of Radiology, Washington University School of Medicine, St. Louis, Missouri, USA
58 ²⁷Cardiovascular Health Research Unit and Department of Epidemiology, University of Washington, Seattle, WA,
59 USA
60 ²⁸Neuroepidemiology Section, Intramural Research Program, National Institute on Aging, Bethesda, Maryland, USA
61 ²⁹Sticht Center for Healthy Aging and Alzheimer's Prevention, Wake Forest School of Medicine, Winston-Salem,
62 North Carolina, USA
63 ³⁰Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, VIC, Australia
64 ³¹CSIRO Health and Biosecurity, Australian e-Health Research Centre CSIRO, Brisbane, Queensland, Australia
65 ³²Wisconsin Alzheimer's Institute, University of Wisconsin School of Medicine and Public Health, Madison,
66 Wisconsin, USA
67 ³³Knight Alzheimer Disease Research Center, Washington University in St. Louis, St. Louis, MO, USA
68 ³⁴Department of Neurology, Johns Hopkins University School of Medicine, USA
69 ³⁵Glenn Biggs Institute for Alzheimer's & Neurodegenerative Diseases, University of Texas Health Science Center at
70 San Antonio, San Antonio, USA
71 ³⁶Department of Neurology and Penn Memory Center, University of Pennsylvania, Philadelphia, USA

72

73 *Corresponding authors:

74 Junhao Wen, Ph.D. – junhaowe@usc.edu

75 2025 Zonal Ave, Los Angeles, CA 90033, United States

76 Christos Davatzikos, Ph.D. – Christos.Davatzikos@pennmedicine.upenn.edu

77 3700 Hamilton Walk, 7th Floor, Philadelphia, PA 19104, United States

78

79 **Word counts:** 8482 words (Title page + Abstract + Significant statement + Introduction + Results + Discussion +
80 Methods + Acknowledge + Figure/Table legends)

81

82 **Keywords:** Pattern of structural covariance, brain imaging genetics, matrix factorization

83

84 **Abstract**

85 Normal and pathologic neurobiological processes influence brain morphology in coordinated
86 ways that give rise to patterns of structural covariance (PSC) across brain regions and individuals
87 during brain aging and diseases. The genetic underpinnings of these patterns remain largely
88 unknown. We apply a stochastic multivariate factorization method to a diverse population of
89 50,699 individuals (12 studies, 130 sites) and derive data-driven, multi-scale PSCs of regional
90 brain size. PSCs were significantly correlated with 915 genomic loci in the discovery set, 617 of
91 which are novel, and 72% were independently replicated. Key pathways influencing PSCs
92 involve reelin signaling, apoptosis, neurogenesis, and appendage development, while pathways
93 of breast cancer indicate potential interplays between brain metastasis and PSCs associated with
94 neurodegeneration and dementia. Using support vector machines, multi-scale PSCs effectively
95 derive imaging signatures of several brain diseases. Our results elucidate new genetic and
96 biological underpinnings that influence structural covariance patterns in the human brain.

97

98

99 **Significance statement**

100 The coordinated patterns of changes in the human brain throughout life, driven by brain
101 development, aging, and diseases, remain largely unexplored regarding their underlying genetic
102 determinants. This study delineates 2003 multi-scale patterns of structural covariance (PSCs) and
103 identifies 617 novel genomic loci, with the mapped genes enriched in biological pathways
104 implicated in reelin signaling, apoptosis, neurogenesis, and appendage development. Overall, the
105 2003 PSCs provide new genetic insights into understanding human brain morphological changes
106 and demonstrate great potential in predicting various neurologic conditions.

107 **Introduction**

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

122 The human brain develops, matures, and degenerates in coordinated patterns of structural
123 covariance at the macrostructural level of brain morphology.¹ However, the mechanisms
124 underlying structural covariance are still unclear, and their genetic underpinnings are largely
125 unknown. We hypothesized that brain morphology was driven by multiple genes (i.e., polygenic)
126 collectively operating on different brain areas (i.e., pleiotropic), resulting in connected networks
127 covaried by normal aging and various disease-related processes. Along the causal pathway from
128 underlying genetics to brain morphological changes, we sought to elucidate which genetic
129 underpinnings (e.g., genes), biological processes (e.g., neurogenesis), cellular components (e.g.,

130 nuclear membrane), molecular functions (e.g., nucleic acid binding), and neuropathological
131 processes (e.g., Alzheimer's disease) might influence the formation, development, and changes
132 of structural covariance patterns in the human brain.

133 Previous neuroimaging genome-wide association studies (GWAS)^{4,5} have partially
134 investigated the abovementioned questions and expanded our understanding of the genetic
135 architecture of the human brain. However, they focused on conventional neuroanatomical
136 regions of interest (ROI) instead of data-driven PSCs. In brain imaging research, prior studies
137 have applied structural covariance analysis to elucidate underlying coordinated morphological
138 changes in brain aging and various brain diseases,¹ but have had several limitations. They often
139 relied on pre-defined neuroanatomical ROIs to construct inter- and intra-individual structural
140 covariance networks. These *a priori* ROIs might not optimally reflect the molecular-functional
141 characteristics of the brain. In addition, most population-based studies have investigated brain
142 structural covariance within a relatively limited scope, such as within relatively small samples,
143 over a relatively narrow age window (e.g., adolescence²), within a single disease (e.g.,
144 Parkinson's disease⁶), or within datasets lacking sufficient diversity in cohort characteristics or
145 MRI scanner protocols. These have been imposed, in part, by limitations in both available cohort
146 size and in the algorithmic implementation of structural covariance analysis, which has been
147 computationally restricted to modest sample sizes when investigated at full image resolution.
148 Lastly, prior studies have examined brain structural covariance at a single fixed ROI
149 resolution/scale/granularity. While the optimal scale is unknown and may differ by the question
150 of interest, the highly complex organization of the human brain may demonstrate structural
151 covariance patterns that span multiple scales.^{7,8}

152 To address this gap, we modified our previously proposed orthogonally projective non-
153 negative matrix factorization (opNMF⁹) to its stochastic counterpart, sopNMF. This adaptation
154 allowed us to train the model iteratively on large-scale neuroimaging datasets with a pre-defined
155 number of PSCs (C). Non-negative matrix factorization has gained significant attention in
156 neuroimaging due to its ability to reduce complex data into a sparse, part-based brain
157 representation by projection onto a relatively small number of components (the PSCs). NMF has
158 been shown to substantially improve interpretability and reproducibility compared to other
159 unsupervised methods, such as PCA and ICA, thanks to the non-negative constraint that
160 produces parcellation-like decompositions of complex signals. Our opNMF/sopNMF approach
161 imposed an additional orthonormality constraint⁹ (*Equation 1* in **Method 1**), further enhancing
162 sparsity and facilitating clinical interpretability. In our previous work, we applied the opNMF
163 method to 934 youths ages 8–20 to depict the coordinated growth of structural brain networks
164 during adolescence – a period characterized by extensive remodeling of the human cortex to
165 accommodate the rapid expansion of the behavioral repertoire². Remarkably, this study revealed
166 PSCs that exhibited a cortical organization closely aligned with established functional brain
167 networks, such as the well-known 7-network functional parcellation proposed by Yeo et al¹⁰.
168 Notably, this alignment emerged without prior assumptions, was data-driven and hypothesis-
169 free, and potentially reflected underlying neurobiological processes related to brain development
170 and aging. Herein, we used large-scale neuroimaging data to investigate the underlying genetic
171 determinant influencing such changes in structural covariance patterns in the human brain.

172 We examined structural covariance of regional cortical and subcortical volume in the
173 human brain using MRI from a diverse population of 50,699 people from 12 studies, 130 sites,
174 and 12 countries, comprised of cognitively healthy individuals, as well as participants with

175 various diseases/conditions over their lifespan (ages 5 through 97). Herein we present results
176 from coarse to fine scales corresponding to $C = 32, 64, 128, 256, 512, \text{ and } 1024$. We
177 hypothesized that PSCs at multiple scales could delineate the human brain's multi-factorial and
178 multi-faceted morphological landscape and genetic architecture in healthy and diseased
179 individuals. We examined the associations between these multi-scale PSCs and common genetic
180 variants at different levels ($N=8,469,833$ SNPs). In total, 617 novel genomic loci were identified;
181 key pathways (e.g., neurogenesis and reelin signaling) contributed to shaping structural
182 covariance patterns in the human brain. In addition, we leveraged PSCs at multiple scales to
183 better derive individualized imaging signatures of several diseases than any single-scale PSCs
184 using support vector machines. All experimental results and the multi-scale PSCs were
185 integrated into the MuSIC (Multi-scale Structural Imaging Covariance) atlas and made publicly
186 accessible through the BRIDGEPORT (**BR**aIn knowle**DGE** **PORT**al) web portal:
187 <https://www.cbica.upenn.edu/bridgeport/>. **Table 1** provides an overview of the abbreviations
188 used in the present study.

189

Table 1. Abbreviations used in the present study

Item	Abbreviation	Item	Abbreviation
Pattern of structural covariation	PSC	Independent component analysis	ICA
Genome-wide association study	GWAS	BRaIn knowleDGE PORTal	BRIDGEPORT
Orthogonal projective non-negative matrix factorization	opNMF	Multi-scale Structural Imaging Covariance	MuSIC
Stochastic orthogonal projective non-negative matrix factorization	sopNMF	Machine learning	ML
Principal component analysis	PCA	UK Biobank	UKBB
Imaging-based coordinate SysTem for AGing and NeurodeGenerative diseases	iSTAGING	Psychosis Heterogeneity Evaluated via Dimensional Neuroimaging	PHENOM
Single nucleotide polymorphism	SNP	Region of interest	ROI
Magnetic resonance imaging	MRI	Automated anatomical labeling	AAL
MUlti-atlas region Segmentation utilizing Ensembles	MUSE	Alzheimer's disease	AD
Spatial PAtterns for REcognition	SPARE	Support vector machine	SVM

192 **Results**

193 We summarize this work in three units (I to III) outlined in **Fig. 1**. In Unit I (**Fig. 1A**), we
194 present the stochastic orthogonally projective non-negative matrix factorization (sopNMF)
195 algorithm (**Method 1**), optimized for large-scale multivariate structural covariance analysis. The
196 sopNMF algorithm decomposes large-scale imaging data through online learning to overcome
197 the memory limitations of opNMF. A subgroup of participants with multiple disease diagnoses
198 and healthy controls (ages 5-97, training population, $N=4000$, **Method 2**) were sampled from the
199 discovery set ($N=32,440$, **Method 2**); their MRI underwent a standard imaging processing
200 pipeline (**Method 3A**). The processed images were then fit to sopNMF to derive the multi-scale
201 PSCs ($N=2003$) from the loadings of the factorization (**Method 1**). We incorporate participants
202 with various disease conditions because previous studies have demonstrated that inter-regional
203 correlated patterns (i.e., depicting a network) show variations in healthy and diseased
204 populations, albeit to a differing degree.¹¹ Multi-scale PSCs were extracted across the entire
205 population and statistically harmonized¹² (**Method 3B**). Unit II (**Fig. 1B**) investigates the
206 harmonized data for 2003 PSCs (13 PSCs have vanished in this process for $C=1024$; see **Method**
207 **1**) in two brain structural covariance analyses. Specifically, we performed *i*) GWAS (**Method 4**)
208 that sought to discover associations of PSCs at single nucleotide polymorphism (SNP), gene, or
209 gene set-level; and *ii*) pattern analysis via support vector machine (**Method 5**) to derive
210 individualized imaging signatures of several brain diseases and conditions. Unit III (**Fig. 1C**)
211 presents BRIDGEPORT, making these massive analytic resources publicly available to the
212 imaging, genomics, and machine learning communities.

213

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

216 We first validated the sopNMF algorithm by showing that it converged to the global minimum of
217 the factorization problem using the comparison population ($N=800$, **Method 2**). The sopNMF
218 algorithm achieved similar reconstruction loss and sparsity as opNMF but at reduced memory
219 demand (**SI eFigure 1**). The lower memory requirements of sopNMF made it possible to
220 generate multi-scale PSCs by jointly factorizing 4000 MRIs in the training population. The
221 results of the algorithm were robust and obtained a high reproducibility index (RI) (**SI eMethod**
222 **2**) in several reproducibility analyses: split-sample analysis ($RI = 0.76 \pm 0.27$), split-sex analysis
223 ($RI = 0.79 \pm 0.27$), and leave-one-site-out analysis ($RI = 0.65-0.78$ for C32 PSCs) (**SI eFigure 2**).
224 We then extracted the multi-scale PSCs in the discovery set ($N=32,440$) and the replication set
225 ($N=18,259$, **Method 2**) for Unit II. These PSCs succinctly capture underlying neurobiological
226 processes across the lifespan, including the effects of typical aging processes and various brain
227 diseases. In addition, the multi-scale representation constructs a hierarchy of brain structure
228 networks (e.g., PSCs in cerebellum regions), which models the human brain in a multi-scale
229 topology.^{7,13}

230
231 **Patterns of structural covariance are highly heritable**

232 The multi-scale PSCs are highly heritable ($0.05 < h^2 < 0.78$), showing high SNP-based heritability
233 estimates (h^2) (**Method 4B**) for the discovery set (**Fig. 2**). Specifically, the h^2 estimate was
234 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,
235 256, 512 and 1024 of the PSCs, respectively. The Pearson correlation coefficient between the two
236 independent estimates of h^2 was $r = 0.94$ ($p\text{-value} < 10^{-6}$, between the discovery and replication

237 sets) in the UK Biobank (UKBB) data. The scatter plot of the two sets of h^2 estimates is shown in
238 **SI eFigure 3**. The h^2 estimates and p-values for all PSCs are detailed in **SI eFile 1** (discovery set)
239 and **eFile 2** (replication set). Our results confirm that brain structure is heritable to a large extent
240 and identify the spatial distribution of the most highly heritable regions of the brain (e.g.,
241 subcortical gray matter structures and cerebellum regions).¹⁴

242

243 **617 novel genomic loci of patterns of structural covariance**

244 We discovered genomic locus-PSC pairwise associations (**Method 4C**, **SI eMethod 5**) within
245 the discovery set and then independently replicated these associations on the replication set. We
246 found that 915 genomic loci had 3791 loci-PSC pairwise significant associations with 924 PSCs
247 after Bonferroni correction (**Method 4G**) for the number of PSCs (p-value threshold per scale:
248 $10.3 > -\log_{10}[\text{p-value}] > 8.8$) (**SI eFile 3**, and **Fig. 3A**). Our results showed that the formation of
249 these PSCs is largely polygenic; the associated SNPs might play a pleiotropic role in shaping
250 these networks.

251 Compared to previous literature, out of the 915 genomic loci, the multi-scale PSCs
252 identified 617 novel genomic loci not previously associated with any traits or phenotypes in the
253 GWAS Catalog¹⁵ (**SI eFile 4**, **Fig. 3B**, query date: April 5th, 2023). These novel associations
254 might indicate subtle neurobiological processes that are captured thanks to the biologically
255 relevant structural covariance expressed by sopNMF. The multi-scale PSCs identified many
256 novel associations by constraining this comparison to previous neuroimaging GWAS^{12,13} using
257 T1w MRI-derived phenotypes (e.g., regions of interest from conventional brain atlases) (**Fig 3B**,
258 **SI eTable 3**, **eFile 5**, **6**, and **7**).

259 Our UKBB replication set analysis (**Method 4H**) demonstrated that 3638 (96%) exact
260 genomic locus-PSC associations were replicated at nominal significance ($-\log_{10}[\text{p-value}] > 1.31$),
261 2705 (72%) of which were significant after correction for multiple comparisons (**Method 4G**, -
262 $\log_{10}[\text{p-value}] > 4.27$). We present this validation in **SI eFile 8** from the replication set. The
263 summary statistics, Manhattan, and QQ plots derived from the combined population ($N=33,541$)
264 are presented in BRIDGEPORT. In addition to the abovementioned replication analyses, we also
265 performed several sensitivity analyses (**SI eFigure 4a**). Our findings revealed the robustness of
266 GWAS signals across both the discovery and replication sets, even when considering four
267 additional brain-related covariates. However, the generalizability of these signals was limited in
268 non-European ancestry populations and independent disease-specific populations (**SI eText 1**
269 and **SI eFigure 4**).

270

271 **Gene set enrichment analysis highlights pathways that shape patterns of structural** 272 **covariance**

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

276 Based on these gene-level p-values, we performed hypothesis-free gene set pathway
277 analysis using MAGMA¹⁷(**Method 4E**): a more stringent correction for multiple comparisons
278 was performed than the prioritized gene set enrichment analysis using *GENE2FUN* from FUMA
279 (**Method 4F** and **Fig. 4**). We identified that six gene set pathways had 18 gene set-PSC pairwise
280 associations with 17 PSCs after Bonferroni correction for the number of gene sets and PSCs
281 ($N=16,768$ and C from 32 to 1024, p-value threshold: $8.54 > -\log_{10}[\text{p-value}] > 7.03$) (**Fig. 3C**, **SI**

282 **eFile 10).** These gene sets imply critical biological and molecular pathways that might shape
283 brain morphological changes and development. The reelin signaling pathway regulates neuronal
284 migration, dendritic growth, branching, spine formation, synaptogenesis, and synaptic
285 plasticity.¹⁸ The appendage morphogenesis and development pathways indicate how the
286 anatomical structures of appendages are generated, organized, and progressed over time, often
287 related to the cell adhesion pathway. These pathways elucidate how cells or tissues can be
288 organized to create a complex structure like the human brain.¹⁹ In addition, the integral
289 component of the cytoplasmic side of the endoplasmic reticulum membrane is thought to form a
290 continuous network of tubules and cisternae extending throughout neuronal dendrites and
291 axons.²⁰ The DSCAM (Down syndrome cell adhesion molecule) pathway likely functions as a
292 cell surface receptor mediating axon pathfinding. Related proteins are involved in hemophilic
293 intercellular interactions.²¹ Lastly, Nikolsky et al.²² defined genes from the breast cancer 20Q11
294 amplicon pathway that were involved in the brain might indicate the brain metastasis of breast
295 cancer, which is usually a late event with deleterious effects on the prognosis.²³ In addition,
296 previous findings^{24,25} revealed an inverse relationship between Alzheimer's disease and breast
297 cancer, which might indicate a close genetic relationship between the disease and brain
298 morphological changes mainly affecting the entorhinal cortex and hippocampus (PSC: C128_3
299 in **Fig. 4**).

300

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

302 To illustrate how underlying genetic underpinnings might form a specific PSC, we showcased
303 two PSCs: C32_4 for the superior cerebellum and C128_3 for the hippocampus-entorhinal
304 cortex. The two PSCs were highly heritable and polygenic in our GWAS using the entire UKBB

305 data (**Fig. 4**, $N=33,541$). We used the FUMA²⁶ online platform to perform *SNP2GENE* for
306 annotating the mapped genes and *GENE2FUNC* for prioritized gene set enrichment analyses
307 (**Method 4F**). The superior cerebellum PSC was associated with genomic loci that can be
308 mapped to 85 genes, which were enriched in many biological pathways, including psychiatric
309 disorders, biological processes, molecular functions, and cellular components (e.g., apoptotic
310 process, axon development, cellular morphogenesis, neurogenesis, and neuro differentiation).
311 For example, apoptosis – the regulated cell destruction – is a complicated process that is highly
312 involved in the development and maturation of the human brain and neurodegenerative
313 diseases.²⁷ Neurogenesis – new neuron formation – is crucial when an embryo develops and
314 continues in specific brain regions throughout the lifespan.²⁸ All significant results of this
315 prioritized gene set enrichment analysis are presented in **SI eFile 11**.

316 For the hippocampus-entorhinal cortex PSC, we mapped 45 genes enriched in gene sets
317 defined from GWAS Catalog, including Alzheimer's disease and brain volume derived from
318 hippocampal regions. The hippocampus and medial temporal lobe have been robust hallmarks of
319 Alzheimer's disease.²⁹ In addition, these genes were enriched in the breast cancer 20Q11
320 amplicon pathway²² and the pathway of metastatic breast cancer tumors³⁰, which might indicate
321 a specific distribution of brain metastases: the vulnerability of medial temporal lobe regions to
322 breast cancer,²³ or highlight an inverse association between Alzheimer's disease and breast
323 cancer.²⁴ Lastly, the nuclear membrane encloses the cell's nucleus – the chromosomes reside
324 inside – which is critical in cell formation activities related to gene expression and regulation. To
325 further support the overlapping genetic underpinnings between this PSC and Alzheimer's
326 disease, we calculated the genetic correlation ($r_g = -0.28$; $p\text{-value}=0.01$) using GWAS summary
327 statistics from the hippocampus-entorhinal cortex PSC (i.e., 33,541 people of European ancestry)

328 and a previous independent study of Alzheimer's disease³¹ (i.e., 63,926 people of European
329 ancestry) using LDSC.³² All significant results of this prioritized gene set enrichment analysis
330 are presented in **SI eFile 12**.

331

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

333 We used the multi-scale PSCs from a diverse population to derive imaging signatures that reflect
334 brain development, aging, and the effects of several brain diseases. We investigate the added
335 value of the multi-scale PSCs as building blocks of imaging signatures for several brain diseases
336 and risk conditions using linear support vector machines (SVM) (**Method 5**).³³ The aim is to
337 harness machine learning to drive a clinically interpretable metric for quantifying an individual-
338 level risk to each disease category. To this end, we define the signatures as SPARE-X (Spatial
339 PAtterns for REcognition) indices, where X is the disease. For instance, SPARE-AD captures the
340 degree of expression of an imaging signature of AD-related brain atrophy, which has been shown
341 to offer diagnostic and prognostic value in prior studies.³⁴

342 The most discriminative indices in our samples were SPARE-AD and SPARE-MCI (**Fig.**
343 **5, SI eTable 4a and eFigure 5**). C=1024 achieved the best performance for the single-scale
344 analysis (e.g., AD vs. controls; balanced accuracy: 0.90 ± 0.02 ; Cohen's *d*: 2.50). Multi-scale
345 representations derived imaging signatures that showed the largest effect sizes to classify the
346 patients from the controls (**Fig. 5**) (e.g., AD vs. controls; balanced accuracy: 0.92 ± 0.02 ; Cohen's
347 *d*: 2.61). PSCs obtained better classification performance than both AAL (e.g., AD vs. controls;
348 balanced accuracy: 0.82 ± 0.02 ; Cohen's *d*: 1.81) and voxel-wise regional volumetric maps
349 (RAVENS)³⁵ (e.g., AD vs. controls; balanced accuracy: 0.85 ± 0.02 ; Cohen's *d*: 2.04) (**SI eTable**
350 **4a and eFigure 5**). Our classification results were higher than previous baseline studies^{36,37},

351 which provided an open-source framework to objectively and reproducibly evaluate AD
352 classification. Using the same cross-validation procedure and evaluation metric, they reported
353 the highest balanced accuracy of 0.87 ± 0.02 to classify AD from healthy controls. Notably, our
354 experiments followed good practices, employed rigorous cross-validation procedures, and
355 avoided critical methodological flaws, such as data leakage or double-dipping (refer to critical
356 reviews on this topic elsewhere^{36,38}).

357 To test the robustness of these SPARE indices, we performed leave-one-site-out analyses
358 for SPARE-AD using the combined 2003 PSCs from all scales (**SI eTable 4b**). Overall, holding
359 the ADNI data out as independent test data resulted in a lower balanced accuracy (0.88 ± 0.02)
360 compared to the other cases for AIBL (0.95 ± 0.02) and PENN data (0.95 ± 0.02). The mean
361 balanced accuracy (0.91 ± 0.02) aligns with the nested cross-validated results using the full
362 sample (**Fig. 5**).

363

364 **BRIDGEPORT: bridging knowledge across patterns of structural covariance, genomics,** 365 **and clinical phenotypes**

366 We integrated our experimental results and the MuSIC atlas into the BRIDGEPORT online web
367 portal. This online tool allows researchers to interactively browse the MuSIC atlas in 3D, query
368 our experimental results via variants or PSCs, and download the GWAS summary statistics for
369 further analyses. In addition, we allow users to search via conventional brain anatomical terms
370 (e.g., the right thalamus proper) by automatically annotating traditional anatomic atlas ROIs,
371 specifically from the MUSE atlas³⁹ (**SI eTable 5**), to MuSIC PSCs based on their degree of
372 overlaps (**SI eFigure 6**). Open-source software dedicated to image processing,³⁹ genetic quality

373 check protocols, MuSIC generation with sopNMF, and machine learning³⁶ is also publicly
374 available (see Code Availability for details).

375 **Discussion**

376 The current study investigates patterns of structural covariance in the human brain at multiple
377 scales from a large population of 50,699 people and, importantly, a very diverse cohort allowing
378 us to capture patterns of structural covariance emanating from normal and abnormal brain
379 development and aging, as well as from several brain diseases. Through extensive examination
380 of the genetic architecture of these multi-scale PSCs, we confirmed genetic hits from previous
381 T1-weighted MRI GWAS and, more importantly, identified 617 novel genomic loci and
382 molecular and biological pathways that collectively influence brain morphological changes and
383 development over the lifespan. Using a hypothesis-free, data-driven approach to first derive these
384 PSCs using brain MRIs, we then uncovered their genetic underpinnings and further showed their
385 potential as building blocks to predict various diseases. All experimental results and code are
386 encapsulated and publicly available in BRIDGEPORT for dissemination:
387 <https://www.cbica.upenn.edu/bridgeport/>, to enable various neuroscience studies to investigate
388 these structural covariance patterns in diverse contexts. Together, the current study highlighted
389 the adoption of machine learning methods in brain imaging genomics and deepened our
390 understanding of the genetic architecture of the human brain.

391 Our findings reveal new insights into genetic underpinnings that influence structural
392 covariance patterns in the human brain. Brain morphological development and changes are
393 largely polygenic and heritable, and previous neuroimaging GWAS has not fully uncovered this
394 genetic landscape. In contrast, genetic variants, as well as environmental, aging, and disease
395 effects, exert pleiotropic effects in shaping morphological changes in different brain regions
396 through specific biological pathways. The mechanisms underlying brain structural covariance are
397 not yet fully understood. They may involve an interplay between common underlying genetic

398 factors, shared susceptibility to aging, and various brain pathologies, which affect brain growth
399 or degeneration in coordinated brain morphological changes.¹ Our data-driven, multi-scale PSCs
400 identify the hierarchical structure of the brain under the principle of structural covariance and are
401 associated with genetic factors at different levels, including SNPs, genes, and gene set pathways.
402 These 617 novel genomic loci, as well as those previously identified, collectively shape brain
403 morphological changes through many key biological and molecular pathways. These pathways
404 are widely involved in reelin signaling, apoptotic processes, axonal development, cellular
405 morphogenesis, neurogenesis, and neuro differentiation,^{27,28} which may collectively influence the
406 formation of structural covariance patterns in the brain. Strikingly, pathways involved in breast
407 cancer shared overlapping genetic underpinnings evidenced in our MAGMA-based and
408 prioritized (*GENE2FUNC*) gene set enrichment analyses (**Fig. 3C** and **Fig. 4**), which included
409 specific pathways involved in breast cancer and metastatic breast cancer tumors. One previous
410 study showed that common genes might mediate breast cancer metastasis to the brain,²³ and a
411 later study further corroborated that the metastatic spread of breast cancer to other organs
412 (including the brain) accelerated during sleep in both mouse and human models.⁴⁰ We further
413 showcased that this brain metastasis of breast cancer might be associated with specific
414 neuropathologic processes, which were captured by PSCs data driven by Alzheimer's disease-
415 related neuropathology. For example, the hippocampus-entorhinal cortex PSC (C128_3, **Fig. 4**)
416 connected the bilateral hippocampus and medial temporal lobe – the salient hallmark of
417 Alzheimer's disease. Our gene set enrichment analysis results further support this claim: the
418 genes were enriched in the gene sets of Alzheimer's disease and breast cancer (**Fig. 4**). Previous
419 research^{24,25} also found an inverse association between Alzheimer's disease and breast cancer. In
420 addition, PSCs from the cerebellum were the most genetically influenced brain regions,

421 consistent with previous neuroimaging GWAS.^{4,5} The cerebral cortex has been thought to largely
422 contribute to the unique mental abilities of humans. However, the cerebellum may also be
423 associated with a much more comprehensive range of complex cognitive functions and brain
424 diseases than initially thought.⁴¹ Our results confirmed that many genetic substrates might
425 support different molecular pathways, resulting in cerebellar functional organization, high-order
426 functions, and dysfunctions in various brain disorders.

427 The current work demonstrates that appropriate machine learning analytics can be used to
428 shed new light on brain imaging genetics. Previous neuroimaging GWAS leveraged multimodal
429 imaging-derived phenotypes from conventional brain atlases^{4,5} (e.g., the AAL atlas). In contrast,
430 multi-scale PSCs are purely data-driven and likely to reflect the dynamics of underlying normal
431 and pathological neurobiological processes giving rise to structural covariance. The diverse
432 training sample from which the PSCs were derived, including healthy and diseased individuals of
433 a wide age range, enriched the diversity of such neurobiological processes influencing the PSCs.
434 In addition, modeling structural covariance at multiple scales (i.e., multi-scale PSCs) indicated
435 that disease effects could be robustly and complementarily identified across scales (**Fig. 5**),
436 concordant with the paradigm of multi-scale brain modeling.¹³ Imaging signatures of brain
437 diseases, derived via supervised machine learning models, were consistently more distinctive
438 when formed from multi-scale PSCs than single-scale PSCs. Multivariate learning techniques
439 have gained significant prominence in neuroimaging and have recently attracted considerable
440 attention in the domain of imaging genomics. These methods have proven valuable for analyzing
441 complex and high-dimensional data, facilitating the exploration of relationships between imaging
442 features and genetic factors. For instance, the MOSTest, a multivariate GWAS approach,
443 preserves correlation structure among phenotypes via permutation on each SNP and derives a

444 genotype vector for testing the association across all phenotypes⁴². A separate study by Soheili-
445 Nezhad et al. demonstrated that genetic components obtained through PCA or ICA applied to
446 neuroimaging GWAS summary statistics exhibited greater reproducibility than raw univariate
447 GWAS effect sizes⁴³. A recent study utilized a CNN-based autoencoder to discover new
448 phenotypes and identify numerous novel genetic signals⁴⁴. Despite the effectiveness of these
449 multivariate approaches in GWAS, they typically conduct phenotype engineering before
450 performing GWAS without explicitly incorporating imaging genetic associations during the
451 modeling process. Yang et al. recently conducted a study that employed generative adversarial
452 networks (termed GeneSGAN⁴⁵) to integrate imaging and genetic variations within the modeling
453 framework to address this limitation. By incorporating both modalities, their approach aimed to
454 capture the complexity and heterogeneity of disease manifestations.

455 MuSIC – with the strengths of being data-driven, multi-scale, and disease-effect
456 informative – contributes to the century-old quest for a "universal" atlas in brain cartography⁴⁶
457 and is highly complementary to previously proposed brain atlases. For instance, Chen and
458 colleagues⁴⁷ used a semi-automated fuzzy clustering technique with MRI data from 406 twins
459 and parcellated the cortical surface area into a genetic covariance-informative brain atlas; MuSIC
460 was data-driven by structural covariance. Glasser and colleagues⁴⁸ adopted a semi-automated
461 parcellation procedure to create a multimodal cortex atlas from 210 healthy individuals.
462 Although this method successfully integrates multimodal information from cortical folding,
463 myelination, and functional connectivity, this semi-automatic approach requires significant
464 resources, some with limited resolution. MuSIC allows flexible, multiple scales for delineating
465 macroscopic brain topology; including patient samples exposes the model to sources of
466 variability that may not be visible in healthy controls. Another pioneering endeavor is the Allen

467 Brain Atlas project,⁴⁹ whose overarching goals of mapping the human brain to gene expression
468 data via existing conventional atlases, identifying local gene expression patterns across the brain
469 in a few individuals, and deepening our understanding of the human brain's differential genetic
470 architecture, are complementary to ours – characterizing the global genetic architecture of the
471 human brain, emphasizing pathogenic variability and morphological heterogeneity.

472 Bridging knowledge across the brain imaging, genomics, and machine learning
473 communities is another pivotal contribution of this work. BRIDGEPORT provides a platform to
474 lower the entry barrier for whole-brain genetic-structural analyses, foster interdisciplinary
475 communication, and advocate for research reproducibility.^{36,50-53} The current study demonstrates
476 the broad applicability of this large-scale, multi-omics platform across a spectrum of
477 neurodegenerative and neuropsychiatric diseases.

478 The present study has certain limitations. Firstly, the sopNMF method utilized in brain
479 parcellation considers only imaging structural covariance and overlooks the genetic determinants
480 contributing to forming these structural networks, as indicated by our GWAS findings.
481 Consequently, further investigations are needed to integrate imaging and genetics into brain
482 parcellation. Additionally, it is important to note that our GWAS analyses primarily involved
483 participants of European ancestry. To enhance genetic findings for underrepresented ethnic
484 groups, future studies should prioritize the inclusion of diverse ancestral backgrounds, thereby
485 promoting a more comprehensive understanding of the genetic underpinnings across different
486 populations.

487 **Methods**

488 **Method 1: Structural covariance patterns via stochastic orthogonally projective non-** 489 **negative matrix factorization**

490 The sopNMF algorithm is a stochastic approximation built and extended based on opNMF^{9,54}.

491 We consider a dataset of n MR images and d voxels per image. We represent the data as a
492 matrix \mathbf{X} where each column corresponds to a flattened image: $\mathbf{X} = [x_1, x_2, \dots, x_n]$, $\mathbf{X} \in \mathbb{R}_{\geq 0}^{d \times n}$.

493 The sopNMF algorithm factorizes \mathbf{X} into two low-rank (r) matrices $\mathbf{W} \in \mathbb{R}_{\geq 0}^{d \times r}$ and $\mathbf{H} \in \mathbb{R}_{\geq 0}^{r \times n}$
494 under the constraints of non-negativity and column-orthonormality. Using the Frobenius norm,
495 the loss of this factorization problem can be formulated as

$$496 \quad \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2$$

$$497 \quad \text{subject to } \mathbf{H} = \mathbf{W}^T \mathbf{X}, \mathbf{W} \geq 0 \text{ and } \mathbf{W}^T \mathbf{W} = \mathbf{I} \quad (1)$$

498 where \mathbf{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-
499 called component matrix $\mathbf{W} = [w_1, w_2, \dots, w_r]$ are part-based representations promoting sparsity
500 in data in this lower-dimensional subspace. From this perspective, the loading coefficient matrix
501 \mathbf{H} represents the importance (weights) of each feature above for a given image. Instead of
502 optimizing the non-convex problem in a batch learning paradigm (i.e., reading all images into
503 memory) as opNMF,⁹ sopNMF subsamples the number of images at each iteration, thereby
504 significantly reducing its memory demand by randomly drawing data batches $\mathbf{X}_b \in \mathbb{R}_{\geq 0}^{d \times b}$ of $b \leq$
505 n images (b is the batch size; $b=32$ was used in the current analyses); this is done without
506 replacement so that all data goes through the model once ($\lceil n/b \rceil$). In this case, the updating rule
507 can be rewritten as

$$508 \quad \mathbf{W}_{t+1} = \mathbf{W}_t \frac{(\mathbf{X}_b \mathbf{X}_b^T \mathbf{W})_t}{(\mathbf{W} \mathbf{W}^T \mathbf{X}_b \mathbf{X}_b^T \mathbf{W})_t} \quad (2)$$

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

$$511 \quad \sum_{i=1}^{\lfloor n/b \rfloor} \|\mathbf{X}_{b_i} - \mathbf{W}\mathbf{W}^T \mathbf{X}_{b_i}\|_F^2 \quad (3)$$

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

516 We applied sopNMF to the training population ($N=4000$). The component matrix \mathbf{W} was
517 sparse after the algorithm converged with a pre-defined maximum number of epochs (100 by
518 default) with an early stopping criterion. To build the MuSIC atlas, we clustered each voxel
519 (row-wise) into one of the r features/PSCs as follows:

$$520 \quad \mathbf{M}_j = \operatorname{argmax}_k(\mathbf{W}_{j,k}) \quad (4)$$

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

526

527 **Method 2: Study population**

528 We consolidated a large-scale multimodal consortium ($N=50,699$) consisting of imaging,
529 cognition, and genetic data from 12 studies, 130 sites, and 12 countries. We present the detailed
530 demographic information of the population under study in **SI eTable 1**. All individual studies

531 were approved by their local corresponding Institutional Review Boards (IRB) (**SI eText 2**). This
532 large-scale consortium reflects the diversity of MRI scans over different races, disease
533 conditions, and ages over the lifespan. To be concise, we defined four populations or data sets
534 per analysis across the paper: *i) discovery set, ii) replication set, iii) training population, and iv)*
535 *comparison population* (refer to **SI eText 3** for details).

536

537 **Method 3: Image processing and statistical harmonization**

538 **(A): Image processing.** Images that passed the quality check (**SI eMethod 4**) were first
539 corrected for magnetic field intensity inhomogeneity.⁵⁵ Voxel-wise regional volumetric maps
540 (RAVENS)³⁵ for each tissue volume were then generated by using a registration method to
541 spatially align the skull-stripped images to a template in MNI-space.⁵⁶ We applied sopNMF to
542 the RAVENS maps to derive MuSIC.

543

544 **(B): Statistical harmonization of MuSIC PSCs:** We applied MuSIC to the entire population
545 ($N=50,699$) to extract the multi-scale PSCs. Specifically, MuSIC was applied to each individual's
546 RAVENS gray matter map to extract the sum of brain volume in each PSC. Subsequently, the
547 PSCs were statistically harmonized by an extensively validated approach, i.e., ComBat-GAM¹²
548 (**SI eMethod 3**) to account for site-related differences in the imaging data. After harmonization,
549 the PSCs were normally distributed (skewness = 0.11 ± 0.17 , and kurtosis = 0.67 ± 0.68) (**SI**
550 **eFigure 7A and B**). To alleviate the potential violation of normal distribution in downstream
551 statistical learning, we quantile-transformed all PSCs. In agreement with the literature,^{57,58} males
552 were found to have larger brain volumes than females on average (**SI eFigure 7C**). Overall, the

553 Combat-GAM model slightly improved data normality across sites (**SI eFigure 7E-H**). The
554 AAL ROIs underwent the same statistical harmonization procedure.

555

556 **Method 4: Genetic analyses**

557 Genetic analyses were restricted to the discovery and replication set from UKBB (**Method 2**).

558 We processed the array genotyping and imputed genetic data (SNPs). The two data sets went

559 through a "best-practice" imaging-genetics quality check (QC) protocol (**Method 4A**) and were

560 restricted to participants of European ancestry. This resulted in 18,052 participants and 8,430,655

561 SNPs for the discovery set and 15,243 participants and 8,470,709 SNPs for the replication set.

562 We reperformed the genetic QC and genetic analyses for the combined populations for

563 BRIDGEPORT, resulting in 33,541 participants and 8,469,833 SNPs. **Method 4G** details the

564 correction for multiple comparisons throughout our analyses.

565

566 **(A): Genetic data quality check protocol.** First, we excluded related individuals (up to 2nd-

567 degree) from the complete UKBB sample ($N=488,377$) using the KING software for family

568 relationship inference.⁵⁹ We then removed duplicated variants from all 22 autosomal

569 chromosomes. We also excluded individuals for whom either imaging or genetic data were not

570 available. Individuals whose genetically identified sex did not match their self-acknowledged sex

571 were removed. Other excluding criteria were: i) individuals with more than 3% of missing

572 genotypes; ii) variants with minor allele frequency (MAF) of less than 1%; iii) variants with larger

573 than 3% missing genotyping rate; iv) variants that failed the Hardy-Weinberg test at 1×10^{-10} . To

574 adjust for population stratification,⁶⁰ we derived the first 40 genetic principle components (PC)

575 using the FlashPCA software⁶¹. The genetic pipeline was also described elsewhere⁶².

576

577 **(B): Heritability estimates and genome-wide association analysis.** We estimated the SNP-
578 based heritability explained by all autosomal genetic variants using GCTA-GREML.⁶³ We
579 adjusted for confounders of age (at imaging), age-squared, sex, age-sex interaction, age-squared-
580 sex interaction, ICV, and the first 40 genetic principal components (PC), guided by a previous
581 neuroimaging GWAS⁴. In addition, Elliot et al.⁵ investigated more than 200 confounders in
582 another study. Therefore, our sensitivity analyses included four additional imaging-related
583 covariates (i.e., brain positions and head motion). One-side likelihood ratio tests were performed
584 to derive the heritability estimates. In GWAS, we performed a linear regression for each PSC
585 and included the same covariates as in the heritability estimates using PLINK.⁶⁴

586

587 **(C): Identification of novel genomic loci.** Using PLINK, we clumped the GWAS summary
588 statistics based on their linkage disequilibrium to identify the genomic loci (see **SI eMethod 5**
589 for the definition of the index, candidate, independent significant, lead SNP, and genomic locus).
590 In particular, the threshold for significance was set to 5×10^{-8} (*clump-p1*) for the index SNPs and
591 0.05 (*clump-p2*) for the **candidate SNPs**. The threshold for linkage disequilibrium-based
592 clumping was set to 0.60 (*clump-r2*) for **independent significant SNPs** and 0.10 for lead SNPs.
593 The linkage disequilibrium physical-distance threshold was 250 kilobases (*clump-kb*). **Genomic**
594 **loci** consider linkage disequilibrium (within 250 kilobases) when interpreting the association
595 results. The GWASRAPIDD⁶⁵ package (version: 0.99.14) was then used to query the genomic
596 loci for any previously-reported associations with clinical phenotypes documented in the
597 NHGRI-EBI GWAS Catalog¹⁵ (p-value $< 1.0 \times 10^{-5}$, default inclusion value of GWAS Catalog).

598 We defined a genomic locus as **novel** when it was not present in GWAS Catalog (query date:
599 April 5th, 2023).

600

601 **(D): Gene-level associations with MAGMA.** We performed gene-level association analysis
602 using MAGMA.¹⁷ First, gene annotation was performed to map the SNPs (reference variant
603 location from Phase 3 of 1,000 Genomes for European ancestry) to genes (human genome Build
604 37) according to their physical positions. The second step was to perform the gene analysis based
605 on the GWAS summary statistics to obtain gene-level p-values between the pairwise 2003 PSCs
606 and the 18,097 protein-encoding genes containing valid SNPs.

607

608 **(E): Hypothesis-free gene set enrichment analysis with MAGMA.** Using the gene-level
609 association p-values, we performed gene set enrichment analysis using MAGMA. Gene sets
610 were obtained from Molecular Signatures Database (MsigDB, v7.5.1),⁶⁶ including 6366 curated
611 gene sets and 10,402 Gene Ontology (GO) terms. All other parameters were set by default for
612 MAGMA. This hypothesis-free analysis resulted in a more stringent correction for multiple
613 comparisons (i.e., by the total number of tested genes and PSCs) than the FUMA-prioritized
614 gene set enrichment analysis (see below F).

615

616 **(F): FUMA analyses for the illustrations of specific PSCs.** In *SNP2GENE*, three different
617 methods were used to map the SNPs to genes. First, positional mapping maps SNPs to genes if
618 the SNPs are physically located inside a gene (a 10 kb window by default). Second, expression
619 quantitative trait loci (eQTL) mapping maps SNPs to genes showing a significant eQTL
620 association. Lastly, chromatin interaction mapping maps SNPs to genes when there is a

621 significant chromatin interaction between the disease-associated regions and nearby or distant
622 genes.²⁶ In addition, *GENE2FUNC* studies the expression of prioritized genes and tests for the
623 enrichment of the set of genes in pre-defined pathways. We used the mapped genes as prioritized
624 genes. The background genes were specified as all genes in FUMA, and all other parameters
625 were set by default. We only reported gene sets with adjusted p-value < 0.05.

626

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

636

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

643 genomic loci, but, in reality, SNPs in linkage disequilibrium with the genomic loci are likely
644 highly significant.

645

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

647 SPARE-AD captures the degree of expression of an imaging signature of AD, and prior studies
648 have shown its diagnostic and prognostic values.³⁴ Here, we extended the concept of the SPARE
649 imaging signature to multiple diseases (SPARE-X, X represents disease diagnoses). Following
650 our reproducible open-source framework³⁷, we performed nested cross-validation (**SI eMethod**
651 **6**) for the machine learning models and derived imaging signatures to quantify individualized
652 disease vulnerability.

653 **SPARE indices.** MuSIC PSCs were fit into a linear support vector machine (SVM) to derive
654 SPARE-AD, MCI, SCZ, DM, HTN, MDD, and ASD. Specifically, the SVM aims to classify the
655 patient group (e.g., AD) from the control group and outputs a continuous variable (i.e., the
656 SPARE indices), which indicates the proximity of each participant to the hyperplane in either the
657 patient or control space. We compared the classification performance using different sets of
658 features: i) the single-scale PSC from 32 to 1024, ii) the multi-scale PSCs by combining all
659 features (with and without feature selections embedded in the CV); iii) the ROIs from the AAL
660 atlas; and iv) voxel-wise RAVENS maps. The samples selected for each task are presented in **SI**
661 **eTable 2.**

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

665 **Data Availability**

666 The GWAS summary statistics corresponding to this study are publicly available on the
667 BRIDGEPORT web portal (<https://www.cbica.upenn.edu/bridgeport/>) and the MEDICINE web
668 portal (<http://labs.loni.usc.edu/medicine/>).

669 **Code Availability**

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

- 671 • sopNMF: <https://pypi.org/project/sopnmf/>, MuSIC, and sopNMF (developed for this
672 study)
- 673 • BRIDGEPORT: <https://www.cbica.upenn.edu/bridgeport/>, (developed for this study)
- 674 • MLNI: <https://pypi.org/project/mlni/>, machine learning (developed for this study)
- 675 • MUSE: <https://www.med.upenn.edu/sbia/muse.html>, image preprocessing
- 676 • PLINK: <https://www.cog-genomics.org/plink/>, GWAS
- 677 • GCTA: <https://yanglab.westlake.edu.cn/software/gcta/#Overview>, heritability estimates
- 678 • LDSC: <https://github.com/bulik/ldsc>, genetic correlation estimates
- 679 • MAGMA: <https://ctg.cncr.nl/software/magma>, gene analysis
- 680 • GWASRAPIDD: <https://rmagno.eu/gwasrapidd/articles/gwasrapidd.html>, GWAS
681 Catalog query
- 682 • MsigDB: <https://www.gsea-msigdb.org/gsea/msigdb/>, gene sets database

683

684 **Competing Interests**

685 DAW served as Site PI for studies by Biogen, Merck, and Eli Lilly/Avid. He has received
686 consulting fees from GE Healthcare and Neuronix. He is on the DSMB for a trial sponsored by
687 Functional Neuromodulation. AJS receives support from multiple NIH grants (P30 AG010133,
688 P30 AG072976, R01 AG019771, R01 AG057739, U01 AG024904, R01 LM013463, R01
689 AG068193, T32 AG071444, and U01 AG068057 and U01 AG072177). He has also received
690 support from Avid Radiopharmaceuticals, a subsidiary of Eli Lilly (in-kind contribution of PET
691 tracer precursor); Bayer Oncology (Scientific Advisory Board); Eisai (Scientific Advisory
692 Board); Siemens Medical Solutions USA, Inc. (Dementia Advisory Board); Springer-Nature
693 Publishing (Editorial Office Support as Editor-in-Chief, Brain Imaging, and Behavior). OC
694 reports having received consulting fees from AskBio (2020) and Therapanacea (2022), having
695 received payments for writing a lay audience short paper from Expression Santé (2019), and that
696 his laboratory has received grants (paid to the institution) from Qynapse (2017-present).
697 Members of his laboratory have co-supervised a Ph.D. thesis with myBrainTechnologies (2016-
698 2019) and with Qynapse (2017-present). OC's spouse is an employee and holds stock options of
699 myBrainTechnologies (2015-present). OC has a patent registered at the International Bureau of
700 the World Intellectual Property Organization (PCT/IB2016/0526993, Schiratti J-B, Allasonniere
701 S, Colliot O, Durrleman S, A method for determining the temporal progression of a biological
702 phenomenon and associated methods and devices) (2017). ME receives support from multiple
703 NIH grants, the Alzheimer's Association, and the Alzheimer's Therapeutic Research Institute.
704 MZ serves as a consultant and/or speaker for the following pharmaceutical companies:
705 Eurofarma, Lundbeck, Abbott, Greencare, Myralis, and Elleven Healthcare.

706

707 **Authors' contributions**

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

709 *Study concept and design:* Wen, Davatzikos

710 *Acquisition, analysis, or interpretation of data:* Wen, Nasrallah, Davatzikos, Abdulkadi,

711 Satterthwaite, Dazzan, Kahn, Schnack, Zanetti, Meisenzahl, Busatto, Crespo-Facorro, Pantelis,

712 Wood, Zhuo, Koutsouleris, Wittfeld, Grabe, Marcus, LaMontagne, Heckbert, Austin, Launer,

713 Espeland, Masters, Maruff, Fripp, Johnson, Morris, Albert, Resnick, Saykin, Thompson, Li,

714 Wolf, Raquel Gur, Ruben Gur, Shinohara, Tosun-Turgut, Fan, Shou, Erus, Wolk

715 *Drafting of the manuscript:* Wen, Nasrallah, Davatzikos

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

717 *Statistical and genetic analysis:* Wen

718 **References**

- 719 1. Alexander-Bloch, A., Giedd, J. N. & Bullmore, E. Imaging structural co-variance between
720 human brain regions. *Nat Rev Neurosci* **14**, 322–336 (2013).
- 721 2. Sotiras, A. *et al.* Patterns of coordinated cortical remodeling during adolescence and their
722 associations with functional specialization and evolutionary expansion. *Proc Natl Acad Sci*
723 *USA* **114**, 3527–3532 (2017).
- 724 3. Blank, S. C., Scott, S. K., Murphy, K., Warburton, E. & Wise, R. J. S. Speech production:
725 Wernicke, Broca and beyond. *Brain* **125**, 1829–1838 (2002).
- 726 4. Zhao, B. *et al.* Genome-wide association analysis of 19,629 individuals identifies variants
727 influencing regional brain volumes and refines their genetic co-architecture with cognitive
728 and mental health traits. *Nat Genet* **51**, 1637–1644 (2019).
- 729 5. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK
730 Biobank. *Nature* **562**, 210–216 (2018).
- 731 6. Vignando, M. *et al.* Mapping brain structural differences and neuroreceptor correlates in
732 Parkinson’s disease visual hallucinations. *Nat Commun* **13**, 519 (2022).
- 733 7. Bassett, D. S. & Siebenhühner, F. Multiscale Network Organization in the Human Brain. in
734 *Multiscale Analysis and Nonlinear Dynamics* 179–204 (John Wiley & Sons, Ltd, 2013).
735 doi:10.1002/9783527671632.ch07.
- 736 8. Schaefer, A. *et al.* Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic
737 Functional Connectivity MRI. *Cerebral Cortex* **28**, 3095–3114 (2018).
- 738 9. Sotiras, A., Resnick, S. M. & Davatzikos, C. Finding imaging patterns of structural
739 covariance via Non-Negative Matrix Factorization. *NeuroImage* **108**, 1–16 (2015).

- 740 10. Thomas Yeo, B. T. *et al.* The organization of the human cerebral cortex estimated by
741 intrinsic functional connectivity. *Journal of Neurophysiology* **106**, 1125–1165 (2011).
- 742 11. Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L. & Greicius, M. D.
743 Neurodegenerative diseases target large-scale human brain networks. *Neuron* **62**, 42–52
744 (2009).
- 745 12. Pomponio, R. *et al.* Harmonization of large MRI datasets for the analysis of brain imaging
746 patterns throughout the lifespan. *Neuroimage* **208**, 116450 (2020).
- 747 13. Betzel, R. F. & Bassett, D. S. Multi-scale brain networks. *NeuroImage* **160**, 73–83 (2017).
- 748 14. Roshchupkin, G. V. *et al.* Heritability of the shape of subcortical brain structures in the
749 general population. *Nat Commun* **7**, 13738 (2016).
- 750 15. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association
751 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012
752 (2019).
- 753 16. Wen, J. *et al.* Genetic, clinical underpinnings of subtle early brain change along
754 Alzheimer’s dimensions. 2022.09.16.508329 Preprint at
755 <https://doi.org/10.1101/2022.09.16.508329> (2022).
- 756 17. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-
757 Set Analysis of GWAS Data. *PLOS Computational Biology* **11**, e1004219 (2015).
- 758 18. Jossin, Y. Reelin Functions, Mechanisms of Action and Signaling Pathways During Brain
759 Development and Maturation. *Biomolecules* **10**, E964 (2020).
- 760 19. Gilbert, S. F. Morphogenesis and Cell Adhesion. *Developmental Biology*. 6th edition
761 (2000).

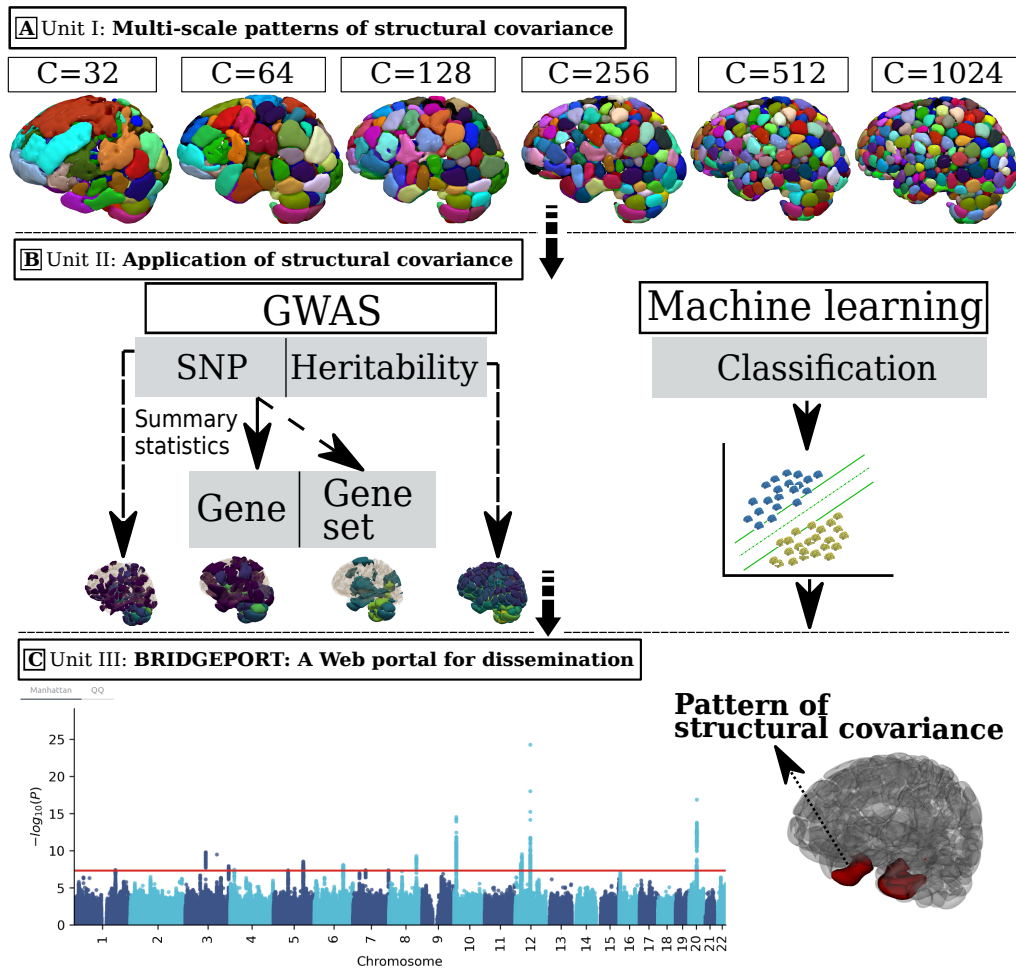
- 762 20. Wu, Y. *et al.* Contacts between the endoplasmic reticulum and other membranes in
763 neurons. *Proc Natl Acad Sci U S A* **114**, E4859–E4867 (2017).
- 764 21. Ly, A. *et al.* DSCAM Is a Netrin Receptor that Collaborates with DCC in Mediating
765 Turning Responses to Netrin-1. *Cell* **133**, 1241–1254 (2008).
- 766 22. Nikolsky, Y. *et al.* Genome-wide functional synergy between amplified and mutated genes
767 in human breast cancer. *Cancer Res* **68**, 9532–9540 (2008).
- 768 23. Bos, P. D. *et al.* Genes that mediate breast cancer metastasis to the brain. *Nature* **459**,
769 1005–1009 (2009).
- 770 24. Lanni, C., Masi, M., Racchi, M. & Govoni, S. Cancer and Alzheimer’s disease inverse
771 relationship: an age-associated diverging derailment of shared pathways. *Mol Psychiatry*
772 **26**, 280–295 (2021).
- 773 25. Shafi, O. Inverse relationship between Alzheimer’s disease and cancer, and other factors
774 contributing to Alzheimer’s disease: a systematic review. *BMC Neurol* **16**, 236 (2016).
- 775 26. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and
776 annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
- 777 27. Yuan, J. & Yankner, B. A. Apoptosis in the nervous system. *Nature* **407**, 802–809 (2000).
- 778 28. Steiner, E., Tata, M. & Frisén, J. A fresh look at adult neurogenesis. *Nat Med* **25**, 542–543
779 (2019).
- 780 29. de Flores, R. *et al.* Medial Temporal Lobe Networks in Alzheimer’s Disease: Structural and
781 Molecular Vulnerabilities. *J Neurosci* **42**, 2131–2141 (2022).
- 782 30. Ginestier, C. *et al.* Prognosis and gene expression profiling of 20q13-amplified breast
783 cancers. *Clin Cancer Res* **12**, 4533–4544 (2006).

- 784 31. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new
785 risk loci and implicates A β , tau, immunity and lipid processing. *Nat Genet* **51**, 414–430
786 (2019).
- 787 32. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from
788 polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).
- 789 33. Davatzikos, C. Machine learning in neuroimaging: Progress and challenges. *NeuroImage*
790 **197**, 652–656 (2019).
- 791 34. Davatzikos, C., Xu, F., An, Y., Fan, Y. & Resnick, S. M. Longitudinal progression of
792 Alzheimer’s-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain*
793 **132**, 2026–2035 (2009).
- 794 35. Davatzikos, C., Genc, A., Xu, D. & Resnick, S. M. Voxel-based morphometry using the
795 RAVENS maps: methods and validation using simulated longitudinal atrophy. *Neuroimage*
796 **14**, 1361–1369 (2001).
- 797 36. Wen, J. *et al.* Convolutional neural networks for classification of Alzheimer’s disease:
798 Overview and reproducible evaluation. *Medical Image Analysis* **63**, 101694 (2020).
- 799 37. Samper-González, J. *et al.* Reproducible evaluation of classification methods in
800 Alzheimer’s disease: Framework and application to MRI and PET data. *NeuroImage* **183**,
801 504–521 (2018).
- 802 38. Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F. & Baker, C. I. Circular analysis in
803 systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* **12**, 535–540 (2009).
- 804 39. Doshi, J. *et al.* MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration
805 algorithms and parameters, and locally optimal atlas selection. *Neuroimage* **127**, 186–195
806 (2016).

- 807 40. Diamantopoulou, Z. *et al.* The metastatic spread of breast cancer accelerates during sleep.
808 *Nature* **607**, 156–162 (2022).
- 809 41. Barton, R. A. & Venditti, C. Rapid Evolution of the Cerebellum in Humans and Other
810 Great Apes. *Curr Biol* **27**, 1249–1250 (2017).
- 811 42. van der Meer, D. *et al.* Understanding the genetic determinants of the brain with MOSTest.
812 *Nat Commun* **11**, 3512 (2020).
- 813 43. Soheili-Nezhad, S., Beckmann, C. F. & Sprooten, E. Reproducibility of Principal and
814 Independent Genomic Components of Brain Structure and Function. 2022.07.13.499912
815 Preprint at <https://doi.org/10.1101/2022.07.13.499912> (2022).
- 816 44. Patel, K. *et al.* New phenotype discovery method by unsupervised deep representation
817 learning empowers genetic association studies of brain imaging. 2022.12.10.22283302
818 Preprint at <https://doi.org/10.1101/2022.12.10.22283302> (2022).
- 819 45. Yang, Z. *et al.* Gene-SGAN: a method for discovering disease subtypes with imaging and
820 genetic signatures via multi-view weakly-supervised deep clustering. *ArXiv*
821 arXiv:2301.10772v1 (2023).
- 822 46. Eickhoff, S. B., Yeo, B. T. T. & Genon, S. Imaging-based parcellations of the human brain.
823 *Nat Rev Neurosci* **19**, 672–686 (2018).
- 824 47. Chen, C.-H. *et al.* Hierarchical Genetic Organization of Human Cortical Surface Area.
825 *Science* **335**, 1634–1636 (2012).
- 826 48. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–
827 178 (2016).
- 828 49. Sunkin, S. M. *et al.* Allen Brain Atlas: an integrated spatio-temporal portal for exploring the
829 central nervous system. *Nucleic Acids Res* **41**, D996–D1008 (2013).

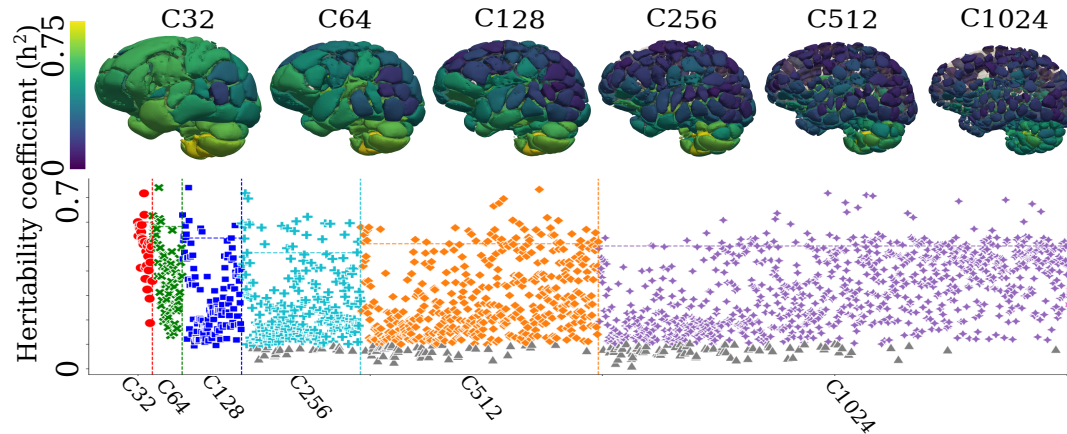
- 830 50. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat Hum Behav* **1**, 1–9 (2017).
- 831 51. Poldrack, R. A. *et al.* Scanning the horizon: towards transparent and reproducible
832 neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
- 833 52. Routier, A. *et al.* Clinica: An Open-Source Software Platform for Reproducible Clinical
834 Neuroscience Studies. *Frontiers in Neuroinformatics* **15**, 39 (2021).
- 835 53. Wen, J. *et al.* Reproducible Evaluation of Diffusion MRI Features for Automatic
836 Classification of Patients with Alzheimer’s Disease. *Neuroinformatics* **19**, 57–78 (2021).
- 837 54. Zhirong Yang & Oja, E. Linear and Nonlinear Projective Nonnegative Matrix
838 Factorization. *IEEE Trans. Neural Netw.* **21**, 734–749 (2010).
- 839 55. Tustison, N. J. *et al.* N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* **29**,
840 1310–1320 (2010).
- 841 56. Ou, Y., Sotiras, A., Paragios, N. & Davatzikos, C. DRAMMS: Deformable Registration via
842 Attribute Matching and Mutual-Saliency Weighting. *Med Image Anal* **15**, 622–639 (2011).
- 843 57. Coupé, P., Catheline, G., Lanuza, E., Manjón, J. V. & Initiative, for the A. D. N. Towards a
844 unified analysis of brain maturation and aging across the entire lifespan: A MRI analysis.
845 *Human Brain Mapping* **38**, 5501–5518 (2017).
- 846 58. Bethlehem, R. a. I. *et al.* *Brain charts for the human lifespan*. 2021.06.08.447489
847 <https://www.biorxiv.org/content/10.1101/2021.06.08.447489v1> (2021)
848 doi:10.1101/2021.06.08.447489.
- 849 59. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies.
850 *Bioinformatics* **26**, 2867–2873 (2010).
- 851 60. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population
852 stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459–463 (2010).

- 853 61. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-
854 scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
- 855 62. Wen, J. *et al.* Characterizing Heterogeneity in Neuroimaging, Cognition, Clinical
856 Symptoms, and Genetics Among Patients With Late-Life Depression. *JAMA Psychiatry*
857 (2022) doi:10.1001/jamapsychiatry.2022.0020.
- 858 63. Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. GCTA-GREML
859 accounts for linkage disequilibrium when estimating genetic variance from genome-wide
860 SNPs. *PNAS* **113**, E4579–E4580 (2016).
- 861 64. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based
862 Linkage Analyses. *Am J Hum Genet* **81**, 559–575 (2007).
- 863 65. Magno, R. & Maia, A.-T. gwasrapidd: an R package to query, download and wrangle
864 GWAS catalog data. *Bioinformatics* **36**, 649–650 (2020).
- 865 66. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for
866 interpreting genome-wide expression profiles. *Proceedings of the National Academy of*
867 *Sciences* **102**, 15545–15550 (2005).
- 868



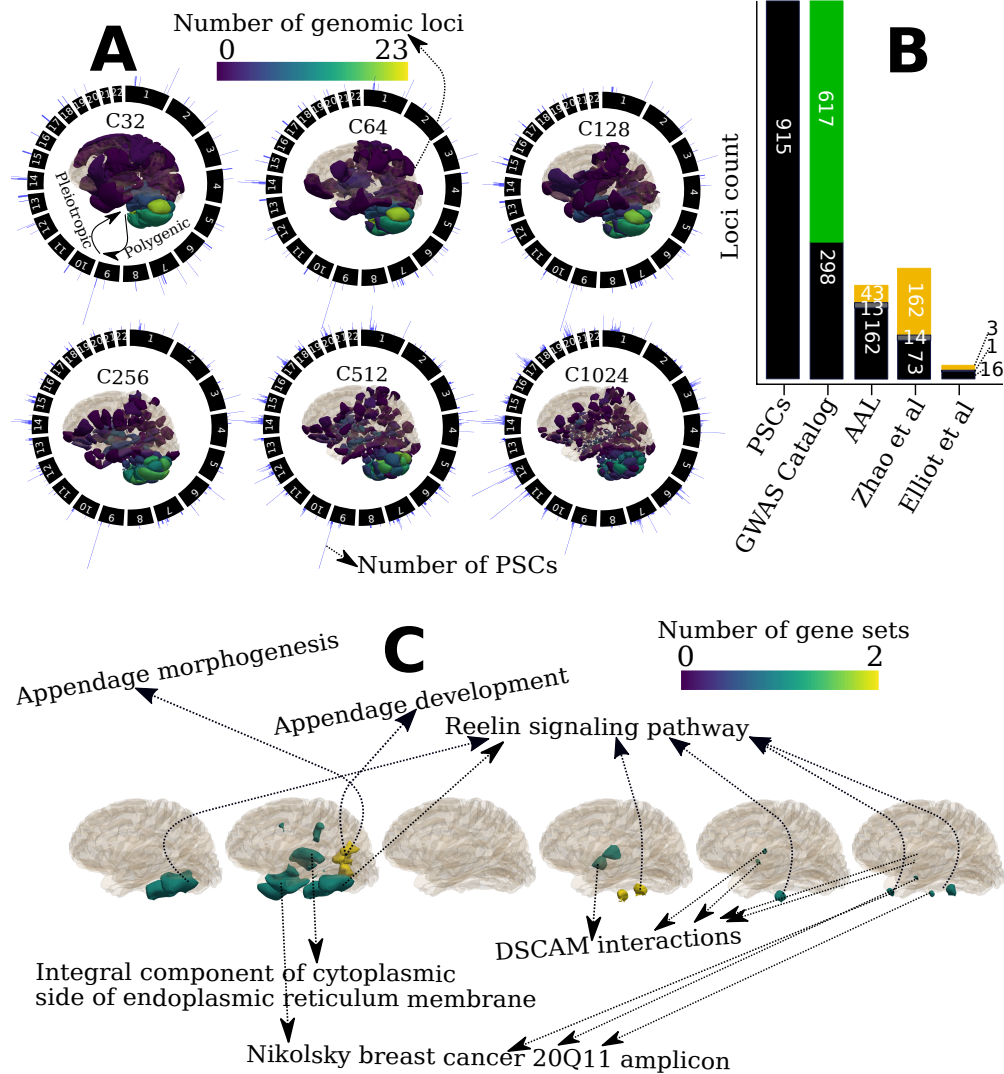
870
871 **Figure 1: Study workflow**

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



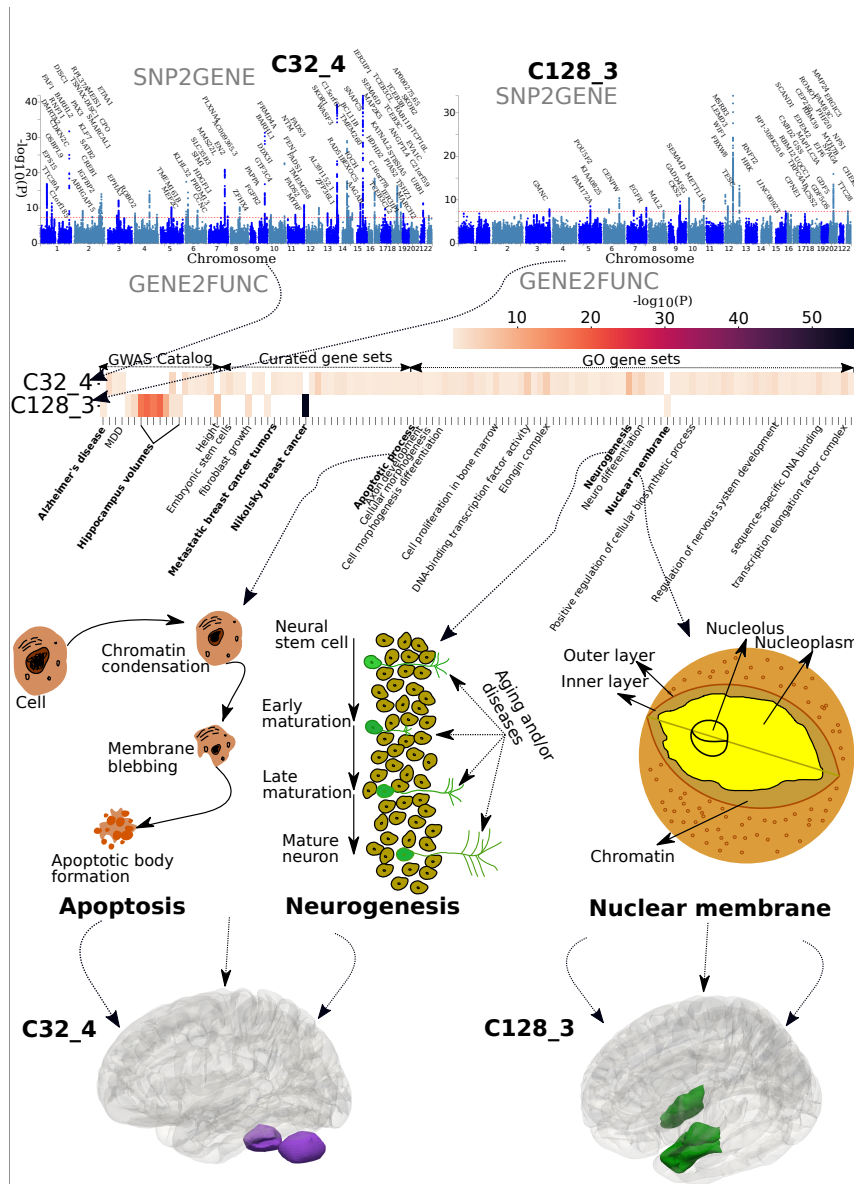
882
 883
 884
 885
 886
 887
 888
 889
 890

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). Each PSC's heritability estimate (h^2) was projected onto the 3D image space to show a statistical map of the brain at each scale C . The dotted line indicates each scale's top 10% of most heritable PSCs.



891 **Figure 3: Patterns of structural covariance highlight novel genomic loci and pathways that**
 892 **shape the human brain.**
 893

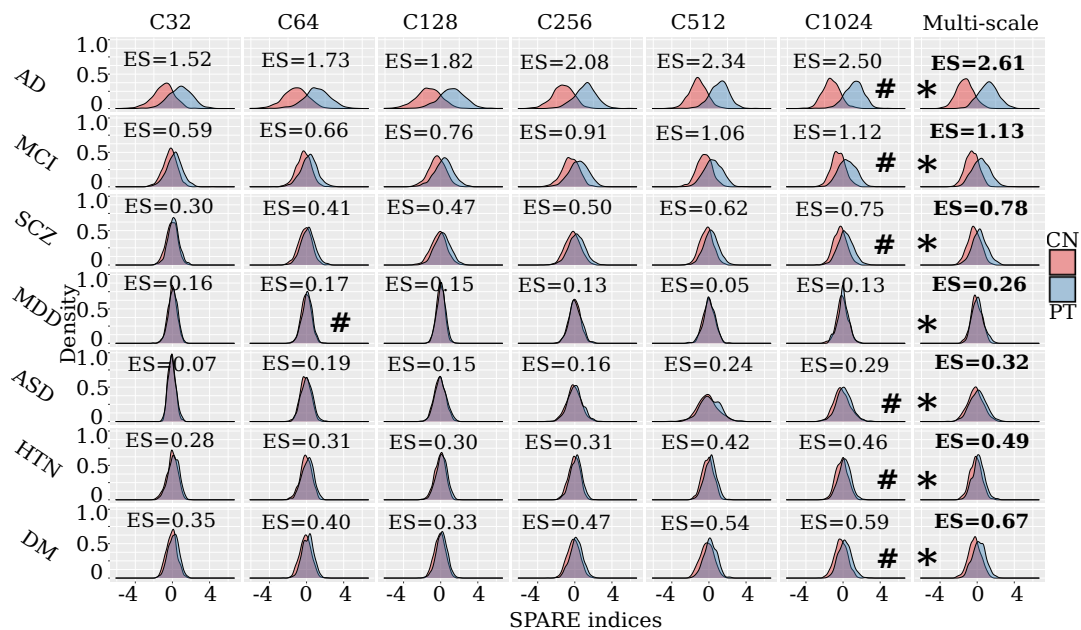
894 **A)** Patterns of structural covariance (PSC) in the human brain are polygenic: the number of
 895 genomic loci of each PSC is projected onto the image space to show a statistical brain map
 896 characterized by the number (*C*) of PSCs. In addition, common genetic variants exert pleiotropic
 897 effects on the PSCs: circular plots showed the number of associated PSCs (histograms in blue
 898 color) of each genomic loci over the entire autosomal chromosome (1-22). The histogram was
 899 plotted for the number of PSCs for each genomic locus in the circular plots. **B)** Novel genomic
 900 loci revealed by the multi-scale PSCs compared to previous findings from the GWAS Catalog,¹⁵
 901 T1-weighted MRI GWAS^{4,5}, and the AAL atlas regions of interest. The green bar indicates the
 902 617 novel genomic loci not previously associated with any clinical traits in GWAS Catalog; the
 903 black bar presents the loci identified in other studies that overlap (grey bar for loci in linkage
 904 disequilibrium) with the loci from our results; the yellow bar indicates the unique loci in other
 905 studies. **C)** Pathway enrichment analysis highlights six unique biological pathways and
 906 functional categories (after Bonferroni correction for 16,768 gene sets and the number of PSCs)
 907 that might influence the changes of PSCs. DSCAM: Down syndrome cell adhesion molecule.



909
 910
 911
 912
 913
 914
 915
 916
 917
 918
 919
 920
 921
 922
 923

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 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.



924
 925 **Figure 5: Individualized imaging signatures based on pattern analysis via machine learning.**
 926 Imaging signatures (SPARE indices) of brain diseases, derived via supervised machine learning
 927 models, are more distinctive when formed from multi-scale PSCs than single-scale PSCs. The
 928 kernel density estimate plot depicts the distribution of the patient group (blue) in comparison to
 929 the healthy control group (red), reflecting the discriminative power of the diagnosis-specific
 930 SPARE (imaging signature) indices. We computed Cohen's d for each SPARE index between
 931 groups to present the effect size of its discrimination power. * represents the model with the
 932 largest Cohen's d for each SPARE index to separate the control vs. patient groups; # represents
 933 the model with the best performance with single-scale PSCs. Our results demonstrate that the
 934 multi-scale PSCs generally achieve the largest discriminative effect sizes (ES) (SI eTable 4a).
 935 As a reference, Cohen's d of ≥ 0.2 , ≥ 0.5 , and ≥ 0.8 , respectively, refer to small, moderate, and
 936 large effect sizes.
 937

938 **Acknowledgments**

939 We acknowledge the contribution from the iSTAGING, the BLSA, the BIOCARD, the
940 PHENOM, the ADNI studies, and the AI4AD consortium. The initial funding package for WJ as
941 an Assistant Professor of Neurology, provided by Stevens Neuroimaging and Informatics
942 Institute, Keck School of Medicine of USC, University of Southern California, supports the
943 present study. The iSTAGING consortium is a multi-institutional effort funded by NIA by RF1
944 AG054409. The PHENOM study is funded by NIA grant R01MH112070 and by the PRONIA
945 project as funded by the European Union 7th Framework Program grant 602152. The Baltimore
946 Longitudinal Study of Aging neuroimaging study is funded by the Intramural Research Program,
947 National Institute on Aging, National Institutes of Health, and by HHSN271201600059C. This
948 research has been conducted using the UK Biobank Resource under Application Number 35148.
949 The research leading to these results has received funding from the French government under
950 management of Agence Nationale de la Recherche as part of the ``Investissements d'avenir"
951 program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAHU-
952 0006 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6)".