Susceptibility of brain atrophy to TRIB3 in Alzheimer's disease, evidence from functional prioritization in imaging genetics

Marco Lorenzi, Andre Altmann, Boris Gutman, Selina Wray, Charles Arber, Derrek D Hibar, Neda J Jahanshad, Jonathan Schott, Daniel Alexander, Paul M. Thompson, et al.

Susceptibility of brain atrophy to TRIB3 in Alzheimer's disease: evidence from functional prioritization in imaging-genetics

Marco Lorenzi[+1,2], Andre Altmann[+1], Boris Gutman[3], Selina Wray[4], Charles Arber[4], Derrek P. Hibar[3], Neda Jahanshad[3], Jonathan M. Schott[5], Daniel C. Alexander[6], Paul M. Thompson[3], and Sebastien Ourselin[1]
for the Alzheimer's Disease Neuroimaging Initiative*

[+] Joint first authors

1. Translational Imaging Group, Centre for Medical Image Computing, University College London, London, UK
2. Asclepios Research Project, Université Côte d'Azur, Inria Sophia Antipolis, France.
3. Imaging Genetics Center, University of Southern California, Los Angeles, CA, USA.
4. Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK.
5. Department of Neurodegeneration, Dementia Research Centre, Institute of Neurology, London, UK.
6. Centre of Medical Image Computing, University College London, London, UK.

Corresponding author:
Marco Lorenzi

Mail:
marco.lorenzi@inria.fr

Address:
Asclepios Research Project
Inria, Sophia Antipolis
2004 route des Lucioles BP 93
06 902 SOPHIA ANTIPOLIS Cedex
FRANCE
+33 4 92 38 76 60

Keywords:
GWA, imaging-genetics, genotype, phenotype, Alzheimer's disease, bioinformatics

*Data used in preparation of this article were 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 analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Abstract

The joint modelling of brain imaging information and genetic data is a promising research avenue to highlight the functional role of genes in determining the pathophysiological mechanisms of Alzheimer's disease (AD). However, since genome-wide association (GWA) studies are essentially limited to the exploration of statistical correlations between genetic variants and phenotype, the validation and interpretation of the findings is usually non trivial and prone to false positives.
To address this issue, in this work we investigate the genetic functional mechanisms underlying brain atrophy in AD by studying the involvement of candidate variants in known genetic regulatory functions. This approach, here termed *functional prioritization*, aims at testing the sets of gene-variants identified by high-dimensional multivariate statistical modelling with respect to known biological processes, in order to introduce a biology-driven validation scheme. When applied to the ADNI cohort, the functional prioritization allowed identifying a link between TRIB3 (tribbles pseudokinase 3) and the stereotypical pattern of grey matter loss in AD, which was confirmed in an independent validation sample, and that provides novel evidence about the relation between this gene and known mechanisms of neurodegeneration.

Significance Statement

In this study we employ a novel experimental imaging-genetics approach for investigating the genetic underpinnings of brain atrophy in Alzheimer's disease. We successfully combined state-of-art imaging-genetics methods and experimental gene expression data to uncover novel biology in brain atrophy.

The novel experimental paradigm highlighted a significant role of TRIB3 (tribbles pseudokinase 3) in modulating the typical pattern of Alzheimer's brain pathology. This result corroborates through rigorous data-driven statistical methods evidence emerging from previous studies about the role of TRIB3 in modulating known mechanisms of neurodegeneration, such as neuronal death, cellular homeostasis, and interaction with established genes causing autosomal dominant Alzheimer's disease: APP and PSEN1. The developed integrated statistical-experimental methodology could serve as a roadmap for investigations in other disorders.

\body

Introduction

Alzheimer's disease (AD) is a devastating neurodegenerative disorder and its aetiology still remains largely concealed. In the anticipation of increasing prevalence of AD and other dementias, there is an urgent need for improving the understanding of the disease processes that underlie neurodegeneration. Whilst the knowledge about the genetic and environmental risks underpinning AD is steadily advancing, our understanding of how these factors interact to lead to the complex pathophysiology that results in dementia is less understood.

Advances in imaging technologies have led to non- or minimally-invasive imaging biomarkers that capture various aspects of the disease process including amyloid deposition [1], tau pathology [2], functional decline [3] and neuronal loss [4]. Combining such imaging information with genetic measurements – so called *imaging-genetics* – provides the means for investigating the effect of genetic variation on underlying biological mechanisms [5].

Genome-wide association studies (GWAS) query millions of single nucleotide polymorphisms (SNPs) individually for their association with either case-control status [6] or disease-specific quantitative phenotypes, e.g., in the case of AD, regional brain volumes [7] or brain amyloid burden [8]. Mass univariate analysis of genetic data is still the predominant method, in virtue of its ease-of-use and well-established theoretical framework, albeit suffering from significant limitations including the requirement for multiple testing, redundancies introduced by linkage disequilibrium (LD) and the lack of analysis of epistatic effects (e.g., SNP-SNP interactions), which

have to be explicitly modeled and searched for exhaustively [9]. Moreover, more than one quantitative phenotype can be derived from the available imaging data, e.g., dozens or hundreds of regional brain volumes, or hundreds of thousands of voxel-level metrics [10]. This potentially large number of genotype-phenotypes features of interest generally complicates the problem of reliably detecting statistical associations, and thus hampers the identification of disease-relevant genetic markers by purely statistical means.

Limitations of classical mass-univariate statistical methods have in recent years been overcome by employing multivariate approaches to data analysis in the context of neuroscience studies [11] and GWAS [12]. Likewise, in imaging-genetics meaningful genotype-phenotype interactions [13] are captured by simultaneously modeling sets of genetic variants that are jointly associated with a given imaging phenotype [14,15,16,17]. Multivariate GWAS have the potential to shed light on the complex genotype-phenotype relationship, and may thus highlight novel links between brain physiology and molecular and biological functions. However, although these methods have proven their ability to identify meaningful SNP combinations associated to brain imaging features, the interpretation and validation of the statistical findings remain very challenging tasks. These problems relate directly to the understanding of the functional role of sets of genetic variants, and to the difficulty of replicating the statistical results in unseen cohorts.

We approach this technical bottleneck by leveraging multivariate approaches to *explore high-dimensional datasets and to generate hypotheses, which are subsequently tested in downstream experiments*. High-quality databases of matched

genotype and gene expression measurements such as GTEx[1] [18] and BRAINEAC[2] [19] facilitate the quantification of effects of SNPs on gene expression in numerous tissues, including various brain tissues. Typically, these databases are used to detail the effect of a genetic variant at the *very end of an analysis pipeline* and to garner evidence for molecular mechanisms of the genetic locus. However, functional information in 'convenience' databases can also be used at an *earlier stage* in the analysis in order to prioritize a few candidate hypotheses with a clear functional mechanism (e.g., expression quantitative trait loci; eQTL) for the validation phase and thus limit the multiple testing burden.

In this work we apply this novel investigative approach to study the genetic functional mechanisms underlying brain atrophy in AD. The framework is comprised of two steps:

i) **Statistical discovery.** Candidate genetic variants are initially identified through data-driven multivariate statistical analysis of the matched imaging and genetics data. This is achieved by modeling the joint covariation between 1.1 million SNPs and the cortical and subcortical atrophy represented by 327,684 cortical and 27,120 sub-cortical thickness values of 639 individuals (either healthy older controls or patients with AD) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort;

ii) **Functional prioritization.** The candidate genetic variants are subsequently screened for functional relevance by querying high-dimensional gene expression databases such as GTEx.

The resulting small set of genetic loci, which are shown to modify gene expression, is then validated in an independent sample of 553 individuals from ADNI diagnosed with mild cognitive impairment (MCI), a proportion of whom progressed to AD.

Compared to previous approaches our work (i) analyses the *whole genome and whole brain* in a hypothesis free fashion, i.e., without preselecting SNPs or brain regions and (ii) uses a *functional prioritization* step in order to select genetic loci for validation in an independent cohort.

Starting from the initial ~1.1 million SNPs, the multivariate statistical analysis allowed the identification of a relatively small number of genetic loci that are statistically associated with the typical pattern of AD brain pathology. The subsequent functional prioritization step ultimately identified a significant role of TRIB3 (tribbles pseudokinase 3), a gene showing important connections to known mechanisms of neurodegenerative diseases. Indeed, although a role for TRIB3 in dementia has not been extensively explored, there are several aspects of TRIB3 function that have relevance to mechanisms related to neuronal death, cellular homeostasis, and of interaction with established AD genes, such as APP and PSEN1.

This study ultimately offers an illustration of the potential of effectively combining multivariate statistical modeling in imaging-genetics with recent instruments available from computational biology, to lead to novel insights on the pathophysiology of neurodegeneration.

Results

Model training and estimated components

Figures 2 and 3 show the relevant areas of the identified joint genetic and phenotype variation, respectively, for the first three PLS components through stability selection. The components were very robust (100% reproducible) during the stability selection procedure (Supplementary Methods). The fourth and fifth components did not present

any relevant locations (i.e., all bins have p<0.95) after stability selection for both the genetic modality and for the imaging modality.

Genetic components

The circular Manhattan plot (Circos v0.96 [20]) of Figure 2 shows the PLS weights and the selection frequency for the PLS genotype components, describing the importance of the genetic loci associated to cortical thickness variation for component 1, 2 and 3. The plot shows the probability of a given genetic bin of size 10kb of being relevant in the PLS model, i.e., to contain a SNP that is ranked in the top 10% of the absolute weights of the genotype component. Spatially contiguous loci generally show similar importance values, which is caused by LD of these regions. The genes close to the important loci are listed in the innermost circle depending on their genomic position.

In the genetic components 1 through 3 a total of 118 bins exceeded the selection frequency threshold (61, 50, 7 for component 1,2 and 3, respectively). From these bins 402 (196, 181 and 25) influential SNPs were extracted and annotated with 98 genes through the computational VEP analysis. The extended *APOE* locus comprising *APOE* and *TOMM40* was selected as the highest scoring region in component 1. A total of 3,956 candidate SNP-gene pairs were considered for the GTEx-based eQTL analysis in six tissues. However, a few genes did not show sufficient expression levels in some tissues and these combinations were excluded from the analysis, resulting in 1,598 unique SNP-gene-tissue tests, of those 104 were significant at the Bonferroni corrected p-value threshold (p=3.1e-5) (Table S1) linking to 14 genes (Table S2; Figure S5): *CAPN9*, *CRYL1*, *FAM135B*, *IL10RA*, *IP6K3*, *ITGA1*, *KIN*, *LAMC1*, *LINC00941*, *LYSMD4*, *RBPMS2*, *RP11-181K3.4*, *TM2D1*, and *TRIB3*.

The independent validation of those 14 genes in the MCI cohort confirmed *TRIB3* (p=0.0034) (Table 2). Three additional genes were (close to) nominal significance: *TM2D1* (p=0.053), *LAMC1* (p=0.062), and *RP11-181K3.4* (p=0.053) (Table 2). Of note the top eQTL SNP for *TRIB3* rs4813620 received a p=0.06175 in stage I of a large AD GWAS [6]. However, rs62191440, a SNP in strong LD with rs4813620 (D'=0.8469; r2=0.6559) in the European population [21], received a p-value of 0.00601 (Figure S6) and also constitutes an eQTL for *TRIB3* in various tissues in GTEx including brain tissues cortex and caudate ganglia (Figure S7). Interestingly, when estimating the PLS components on the sub-cohort of 279 training individuals with positive CSF amyloid (Table 1) we identified compatible validation results on the independent testing MCI group. Within this setting, TRIB3 still leads to marginally significant differences (p=0.0134) between progressing and stable MCI, although not significant after correction for multiple comparison (Table S3).

Morphometric components

Figure 3 shows the PLS phenotype components 1 through 3 (top), as well as the associated selection frequency describing the loci of brain atrophy associated with genetic variation (bottom). The selection frequency colors indicate the probability of each cortical mesh points of being relevant in the PLS model, i.e., to be ranked among the top 10% of the absolute weights of the phenotype component.

The first component is mainly associated to the thinning of the cortical mantle, and is localized in temporal and posterior cingulate cortices (Figure 3). The relevant areas at the subcortical level are primarily associated with amygdalae and thalami. The second

component is mostly associated to the thinning of the subcortical areas (hippocampi and amygdalae), and to the cortical thinning of the temporal areas at the cortical level. The third component is similar to component 2, and describes a sub-cortical thickness pattern prevalent in hippocampi, amygdalae, and thalami. At the cortical level, the component is associated with the thinning of frontal cortices, and to isolated spots located in the parahippocampal gyrus.

Discussion

In this work we modeled high-dimensional genome-wide SNP data and brain-wide cortical thickness data via joint multivariate statistical modeling and functional prioritization of genes through bioinformatics annotation and a large eQTL database. Our study ultimately identified a link between *TRIB3* (tribbles pseudokinase 3) and the stereotypical pattern of grey matter loss in AD (cortical thinning in temporal and posterior cingulate regions and subcortical atrophy). *TRIB3* is a pseudokinase which acts as a regulator of several signaling pathways. For example it can interact directly with Akt and inhibit the pro-survival pro-survival Akt pathway [22]. *TRIB3* expression is induced during neuronal cell death [23] and recently increased levels of the TRIB3 protein were found in dopaminergic neurons of the substantia nigra pars compacta in patients with Parkinson's disease [24]. *TRIB3* expression is stress induced and increases in response to nerve growth factor (NGF) deprivation; endoplasmatic reticulum (ER) stress, and amino acid deprivation [23]. Although a role for *TRIB3* in dementia has not been extensively explored, there are several aspects of *TRIB3* function that have relevance to known mechanisms of neurodegenerative disease. TRIB3 can interact directly with P62 to modulate autophagic flux [25], an important process in maintaining cellular homeostasis that is

known to be disrupted in neurodegeneration [26]. Knockdown of *TRIB3* modulates PSEN1 stability [25] and a yeast two-hybrid screen identified progranulin as a direct interaction partner of TRIB3 [27]. Intriguingly, it has recently been demonstrated that TRIB3 induces both apoptosis and autophagy in Aβ-induced neuronal death, and silencing of TRIB3 was strongly neuroprotective [28]. These links warrant further investigation for a functional role of *TRIB3* in neuronal death in dementia.

These earlier findings align with our eQTL analysis where carriers of the minor allele show increased *TRIB3* expression (Figure S5), which potentially lowers the threshold to *TRIB3* mediated neuronal cell death. *TRIB3* expression was modulated by the identified SNP in various other tissues including the caudate (Figure S7), a region affected in PD and Huntington's disease. A recent study of *Trib3* expression in mice concluded that "*Trib3* has a pathophysiological role in diabetes" [29]; diabetes itself is a known risk factor for dementia [30] perhaps through shared metabolic processes with AD [31]. Interestingly, one of the three SNPs (rs1555318) selected in the PLS model and attributed to *TRIB3* showed a strong association with type-2 diabetes in stage 1 of a large GWAS (p=4.4e-4; Figure S8) [32]. Other GWAS showed links between *TRIB3* and information processing speed (p=1.7e-7) [33] and AD (p=0.006; [6]). An earlier genetic study on AD in Swedish men found an association in *TRIB3* as well (p=0.044; [34]), which was replicated in a Canadian cohort (p<0.001 ; [35]). Lastly, TRIB3 was reported to physically interact with APP [36] and it shares numerous functional annotations for biological processes regarding lipid metabolism with APOE.

The functional prioritization component of the analysis successfully reduced the set of candidate genetic variants for the independent validation, however, this prioritization

has a shortcoming: it hypothesizes that identified SNPs alter the expression of a nearby gene. Although, this scheme led to the identification of *TRIB3* in the cortical thickness phenotype, it did miss a long-established AD risk gene: *APOE*. SNPs belonging to *APOE* (rs429358 and rs7412) were selected as highest scoring SNPs in component 1. However, none of them was detected as an eQTL and thus *APOE* was excluded from the downstream analysis. Other types of functional prioritizations based on exonic function prediction may have retained *APOE* and other genes in the pipeline. However, SNPs data typically features only a few non-synonymous exonic variants and their high frequency (MAF >5%) renders them unlikely to receive significant 'damaging' scores in these predictions. Thus, for this scenario the use of these function predictions would be limited.

The list of genes we identified contains other interesting candidates. For instance, *IL10RA* (interleukin 10 receptor subunit alpha) is a receptor for interleukin 10 (IL10), a cytokine that controls inflammatory response [37]. Carriers of the minor allele show increased *IL10RA* expression (Figure S5) and *Il10ra* expression is increased in affected brain regions with increasing age and presence of AD pathology in transgenic mouse models of AD (MOUSEAC; [38] Figure S9). Moreover, a link between downregulation of *IL10RA* and *TRIB3* in *TRIB3*-silenced HepG2 cells was reported in [25], along with increased abundance of Presenilin 1, ApoE3, and Clusterin. Finally, blocking IL10 response was recently suggested as a therapeutic mechanism in AD [39]. A gene that showed a statistical trend in the validation sample was TM2D1 (TM2 domain containing 1), which is a beta-amyloid binding protein and may be involved in beta-amyloid-induced apoptosis [40]. Further, *MEF2A* (Myocyte Enhancer Factor 2A), like *APOE*, was filtered out by the functional prioritization.

However, *MEF2A* is a paralog of *MEF2C*, which is an established AD gene [6]. Noteworthy, bins covering *MEF2C* only barely missed the selection threshold in component 2 for further analysis (max p=0.926; Figure 2).

Methodological Considerations

The experimental setting proposed in this study is based on the investigation of potential genetic candidates in the AD and healthy training population, and on their testing in the MCI cohort. This experimental choice was motivated by clinical and practical considerations.

From the clinical point of view, although we cannot exclude that the imaging-genetics association patterns could be modulated by state-specific factors throughout the development of the disease [41], the heterogeneity of the MCI label is likely to lead to the inclusion in the discovery dataset of individuals with non-AD pathologies. Thus, including MCIs in the discovery cohort bears the risk of diluting the gene finding (especially considering the relatively low sample size of the study cohort). Likewise, GWAS in AD carried out to date focus on comparing CT and AD. Moreover, the paradigm proposed in this study is rather conservative since it explores associations present throughout the progression of the pathology, i.e., associations were discovered by comparing CT and AD subjects and validated on disease progression in the intermediate MCI cohort. This consideration, while being more conservative, may play in favor of the robustness of the reported results. From a practical point of view, the proposed scheme allowed the validation of the model on a clinically relevant testing cohort by taking advantage of the full sample available in the ADNI dataset. Splitting the available AD and CT subjects into discovery and validation cohort,

would have dramatically reduced the sample size, thus increasing the uncertainty of the PLS findings.

Concerning the number of components analyzed in the PLS model, we limited the study to the exploration of the first five eigen-modes. As shown in the experimental results, the stability of PLS parameters of the high-order components was generally quite low and did not lead to any significant results after permutation testing. For this reason, we believe that extending the analysis to higher-order components (e.g., components six to ten) would not change the proposed analysis and subsequent results.

The relevance assessment procedure proposed in this study relies on the choice of statistical significance thresholds, such as the 10% cutoff on the magnitude of the PLS weights, and $p<0.05$ for the selection frequency over the 1,000,000 folds. These thresholds were not optimized to maximize specific statistical outcome (e.g. the ratio between true and false positives). Indeed, the optimization of these parameters may lead to important methodological issues such as overfitting and selection bias [42], and ultimately lead to poor generalization of the statistical findings. This is particularly true in the challenging setting proposed in this work, characterized by large dimensions and low sample size. For this reason, we chose to use standard cutoffs for significance assessment as a compromise between minimizing this important source of bias while still identifying meaningful genotype and phenotype features. Furthermore, we believe that the ultimate approach to assess the validity of the findings is through testing on genuinely independent data, such as on the MCI cohort proposed in this study.

Conclusions

This study illustrates the potential of effectively combining multivariate statistical modeling in imaging-genetics with recent instruments available from computational biology, to lead to novel insights on the disease pathophysiology. Thanks to the ever-growing data-driven knowledge based on the vast quantities of information now available to the research community, the paradigm proposed in this study may represent a promising avenue for linking imaging-genetics findings to the current knowledge on functional genetics mechanisms involved in neurodegeneration.

Materials and methods

Study Participants

Data used in the preparation of this article were obtained from the ADNI database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI is to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org. This research mainly involves further processing of previously collected personal data. We have explicit authorization for the use of the ADNI dataset, and we have signed the relevant papers guaranteeing that we abide to the ethics standards. The ADNI protocol details on page 30-31 the informed consent for imaging data (section d.5.d) and the procedures to maintain confidentiality of the data (section D.5.e).

We selected genotype and phenotype data available in the ADNI-1/GO/2 datasets for 1,192 subjects. Summary socio-demographic, clinical and genetic information are available in Table 1. At time of study entry subjects were diagnosed as healthy individuals (N=401), MCI (N=553) or AD (N=238). A total of 212 (38.3%) MCI patients subsequently converted to AD over the course of the study (6 years). All participants were non-Hispanic Caucasian, with a prevalence of males across the considered groups. AD and MCI groups show significant cognitive decline measured by MMSE and ADAS-COG as compared to the healthy individuals ($p<1e-2$, two sample t-test for group-wise comparison). There was also a significant increase of individuals with pathological levels of $A\beta_{1-42}$ in the CSF ($A\beta_{1-42}$ <192pg/ml) across the clinical groups, with proportions ranging from 43% for healthy individuals to 93% for AD patients ($p<1e-2$). Similarly, we observed a higher prevalence of APOE4 carriers in AD and progressing MCI individuals when compared to healthy and MCI stable groups.

In what follows, the 639 healthy and AD subjects form the *discovery* set, while the MCI converters and non-converters form the independent *validation* set.

Data processing

The imaging phenotype comprised the baseline brain cortical thickness maps estimated with FreeSurfer 5.3 [43] and the bilateral radial thickness maps for hippocampi, amygdalae, thalami, caudate, putamen, globus pallidus and nucleus accumbens. In detail, radial thickness of each subcortical surface model was based on the distance to a medial curve. We fit the medial curve using curve evolution individually for each shape [44]. Surfaces are then registered parametrically to achieve point-to-point correspondence by matching curvature and medial curve-based

features [45,46]. The procedure resembles the cortical surface registration on the sphere performed in FreeSurfer. Finally, the full imaging component comprises 327,684 cortical and 27,120 subcortical features per subject.

SNP genotype data (Illumina Human610-Quad BeadChip for ADNI-1, and Illumina Human Omni Express for ADNI-GO/2) was downloaded from the ADNI website and preprocessed with PLINK [47]. Standard quality control (QC) parameters were used to filter SNPs: minor allele frequency (MAF) < 0.01, genotype call rate <95% and Hardy-Weinberg equilibrium (HWE) p-value < $1x10^{-6}$. Finally, genotyped SNPs passing QC were used to impute SNPs in the HapMap III reference panel. Imputed SNPs underwent a separate QC regarding minor allele frequency (MAF > 0.01) and imputation quality (imputation R-squared > 0.3) in order to exclude poorly imputed SNPs. For the analysis the individuals' minor allele counts for each of the resulting 1,167,126 SNPs in the 22 autosomes were used.

Statistical Discovery

The joint relationship between the genetic and imaging modalities was investigated through partial least squares (PLS) modeling [48,49,50,51,52,53]. Among the several PLS versions proposed in the literature we focus on the symmetric formulation of PLS computed through the singular value decomposition (SVD) of the cross covariance matrix (Figure S1) [51,52,54]. Within this setting, the aim of PLS is to estimate the latent components that maximize the global covariance between the two input modalities. Each input feature receives a weight in the latent component that represents its relative importance for describing the global joint multimodal

relationship. Analyzing these weights helps identifying SNPs that are linked to the patterns of cortical thinning in the brain.

In this study we applied a robust approach for the stable estimation and interpretation of PLS weights in genome-wide genotyping data, aimed at promoting *sparsity* (i.e., selecting only few features for simplified interpretation) and *regularity* (by aggregating SNPs within the same genetic neighborhood). This is achieved through a stability selection procedure in which the reproducibility and robustness of the PLS parameters is assessed through a split-half cross-validation based scheme on 1,000,000 repetitions of the models on randomly sampled subgroups (Figure 1 and supplementary Methods).

By considering a pre-defined partition of each chromosome into contiguous loci of size 10kb, the procedure leads to the estimation of a confidence measure taking values ranged between 0.0 and 1.0 indicating the probability of each genetic loci to contain highly reproducible PLS weights, and therefore serving as a measure of importance of the genomic location (Figure 2).

A similar procedure was employed to assess the importance of the phenotype component (Figure 1). However, no regional binning was employed (Figure 3).

The procedure was applied to assess the parameter reproducibility of the first five PLS modes; subsequent analyses were performed only on components with relevant genetic and brain regions (i.e., reproducible PLS weights with selection frequency >95%). PLS components and probability measures will be made available at the author's website.

Gene identification

We analyzed the 10kb bins (genetic loci) with the selection frequency exceeding *0.95,* i.e., bins selected in 95% or more of the 1,000,000 replications. Within these bins we then identified the influential SNPs: a SNP was declared influential if it was associated with the weights of greatest magnitude in the PLS components estimated on the full data sample, i.e., SNPs with absolute weights exceeding the 99th quantile of all weights in the component. These weights are the ones contributing to the high selection frequency in the split-half procedure, and are representative of the significant variation modeled in the data. Focusing on the features associated with these weights allows us to restrict the functional prioritization on a SNPs subset of reduced dimensionality, by focusing only on the most representative elements.
In order to link SNPs to corresponding genes we used the Ensembl Variant Effect Predictor (VEP) for GRCh37 (date accessed: 17th October 2016) [55] with the GENCODE gene annotation. SNPs tagged as 'regulatory' were manually investigated and annotated with the nearby genes.

Functional prioritization

All SNPs successfully annotated with a gene were subjected to functional prioritization through expression quantitative trait loci (eQTL) analysis based on the Genotype-Tissue-Expression project (GTEx) data. The sample size in GTEx for relevant brain tissues in AD was rather small (e.g., N=81 for hippocampus). Therefore, we added five more tissues with large samples sizes that were more distantly relevant to AD: nerve tibial (N=256) was added as a proxy for nervous tissue; whole blood (N=338) and artery tibial (N=285) were included to cover blood-based changes and effects on blood vessels [56] adipose subcutaneous (N=298) was

selected due to links between AD and obesity, type-2 diabetes and metabolic disease [57,58]. Finally, transformed fibroblasts (N=272) were included as a general-purpose cell line. P-values were corrected for multiple testing using the Bonferroni method.

Model validation in independent MCI subjects

The genes that were found to be under expression control by the identified SNPs were validated for their capacity to predict clinical conversion in MCI subjects. To this end, for each identified gene we applied the PLS weights estimated on the discovery set on the validation set, with the genetic component restricted to SNPs +/- 20kb of the gene borders. The identified latent projections (i.e., a weighted sum of SNPs) results in one score per subject per gene. For each gene the association of the projection score with conversion status was assessed by statistically comparing the scores distribution between healthy individuals and AD patients, and between MCI converters and non-converters (Kruskal-Wallis non parametric test for two sample comparison, Bonferroni correction for multiple comparisons).

Acknowledgment

SO receives funding from the EPSRC (EP/H046410/1, EP/J020990/1, EP/K005278), the MRC (MR/J01107X/1), the EU-FP7 project VPH-DARE@IT (FP7-ICT-2011-9-601055), the NIHR Biomedical Research Unit (Dementia) at UCL and the National Institute for Health Research University College London Hospitals Biomedical Research Centre (NIHR BRC UCLH/UCL High Impact Initiative-BW.mn.BRC10269). EPSRC grants EP/J020990/01 and EP/M020533/1 support DA and SO's work on this topic. ML, DA, JS, and SO also received support from the *European Union's Horizon 2020 research and innovation programme* under grant

agreement No 666992 (EuroPOND) for this work. JMS acknowledges the support of the National Institute for Health Research University College London Hospitals Biomedical Research Centre, Wolfson Foundation, EPSRC (EP/J020990/1), MRC (MR/L023784/1), ARUK (ARUK-Network 2012-6-ICE; ARUK-PG2017-1946; ARUK-PG2017-1946), Brain Research Trust (UCC14191) and European Union's Horizon 2020 research and innovation programme (Grant 666992). AA holds an MRC eMedLab Medical Bioinformatics Career Development Fellowship. SW and CA are supported by the NIHR Queen Square Biomedical Research Unit in Dementia and Alzheimer's Research UK, and by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. Contribution to this work by BG and PT was funded by the National Institutes of Health "Big Data to Knowledge" (BD2K) award (NIH U54 EB020403, PI: Thompson).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). 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.; 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 organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for NeuroImaging at the University of Southern California.

Conflict of Interest

The authors do not report any conflict of interest.

Footnotes

[1] http://gtexportal.org/home/
[2] http://www.braineac.org/

References

1 Rabinovici, G. D., & Jagust, W. J. Amyloid imaging in aging and dementia: testing the amyloid hypothesis in vivo. *Behavioural neurology*, 21, 1-2 (2009), 117-128.
2 Villemagne, V. L., & Okamura, N. In vivo tau imaging: obstacles and progress. *Alzheimer's & Dementia*, 10, 3 (2014), S254-S264.
3 Mosconi, L., Berti, V., Glodzik, L., Pupi, A., De Santi, S., & de Leon, M. J. Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging. *Journal of Alzheimer's Disease*, 20, 3 (2010), 843-854.
4 Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., & Thompson, P. M. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6, 2 (2010), 67-77.
5 Bigos, K. L., Hariri, A. R., & Weinberger, D. R. *Neuroimaging Genetics: Principles and Practices*. Oxford University Press, Oxford, 2016.
6 Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature genetics*, 45, 12 (2013), 1452-1458.
7 Potkin, S. G., Guffanti, G., Lakatos, A., Turner, J. A., Kruggel, F., Fallon, J. H., et al. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PloS one*, 4, 8 (2009), e6501.
8 Ramanan, V. K., Risacher, S. L., Nho, K., Kim, S., Swaminathan, S., Shen, L., et al, R. C. APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study. *Molecular psychiatry*, 19, 3 (2014), 351-357.
9 Kam-Thong, T., Azencott, C. A., Cayton, L., Pütz, B., Altmann, A., Karbalai, N., et al. GLIDE: GPU-based linear regression for detection of epistasis. *Human heredity*, 73, 4 (2012), 220-236.
10 Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., et al. Voxelwise genome-wide association study (vGWAS). *NeuroImage*, 53, 3 (2010), 1160-1174..
11 Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., et al. PRoNTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics*, 11, 3 (2013), 319-337.
12 Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., et al. Machine learning in genome-wide association studies. *Genetic epidemiology*, 33, S1 (2009), S51-S57.
13 Liu, J., & Calhoun, V. D. A review of multivariate analyses in imaging genetics. *Front. Neuroinform.*, 8, 29 (2014).
14 Le Floch, É., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., et al. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage* , 63, 1 (2012), 11-24.
15 Vounou, M., Nichols, T. E., Montana, G. Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. *NeuroImage*, 53, 3 (2010), 1147-1159.
16 Silver, M., Janousova, E., Hua, X., Thompson, P. M., Montana, G. Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage*, 63, 3 (2012), 1681-1694.
17 Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I., & Calhoun, V. Combining fMRI and SNP data to investigate connections between

brain function and genetics using parallel ICA. *Human brain mapping*, 30, 1 (2009), 241-255.
18 Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and biobanking*, 13, 5 (2015), 311-319.
19 Trabzuni, D., Ryten, M., Walker, R., Smith, C., Imran, S., Ramasamy, A., et al. Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *Journal of neurochemistry,* , 119, 2 (2011), 275-282.
20 Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. Circos: an information aesthetic for comparative genomics. *Genome research*, 19, 9 (2009), 1639-1645.
21 Machiela, M. J., & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics*, 31, 21 (2015), 3555-3557.
22 Du, K., Herzig, S., Kulkarni, R. N., & Montminy, M. TRB3: a tribbles homolog that inhibits Akt/PKB activation by insulin in liver. (2003), 1574-1577.
23 Zareen, N., Biswas, S. C., & Greene, L. A. A feed-forward loop involving Trib3, Akt and FoxO mediates death of NGF-deprived neurons. *Cell Death & Differentiation*, 20, 12 (2013), 1719-1730.
24 Aimé, P., Sun, X., Zareen, N., Rao, A., Berman, Z., Volpicelli-Daley, L., et al. Trib3 Is elevated in Parkinson's disease and mediates death in Parkinson's disease models. *The Journal of Neuroscience*, 35, 30 (2015), 10731-10749.
25 Hua, F., Li, K., Yu, J. J., Lv, X. X., Yan, J., Zhang, X. W., et al. TRB3 links insulin/IGF to tumour promotion by interacting with p62 and impeding autophagic/proteasomal degradations. *Nature communications*, 13, 6 (2015), 7951.
26 Menzies, F. M., Fleming, A., & Rubinsztein, D. C. Compromised autophagy and neurodegenerative diseases. *Nature Reviews Neuroscience*, 16, 6 (2015), 345-357.
27 Zhou, Y., Li, L., Liu, Q., Xing, G., Kuai, X., Sun, J., et al. E3 ubiquitin ligase SIAH1 mediates ubiquitination and degradation of TRB3. *Cellular signalling*, 20, 5 (2008), 942-948.
28 Saleem, S., & Biswas, S. C. Tribbles Pseudokinase 3 Induces Both Apoptosis and Autophagy in Amyloid-β induced Neuronal Death. *Journal of Biological Chemistry*, jbc-M116 (2016).
29 Zhang, W., Wu, M., Kim, T., Jariwala, R. H., Garvey, W. J., Luo, N., et al. Skeletal Muscle TRIB3 Mediates Glucose Toxicity in Diabetes and High Fat Diet-Induced Insulin Resistance. *Diabetes*, 65, 8 (2016), 2380-2391.
30 Sims-Robinson, C., Kim, B., Rosko, A., & Feldman, E. L. How does diabetes accelerate Alzheimer disease pathology? *Nature Reviews Neurology*, 6, 10 (2010), 551-559.
31 Ribe, E. M., & Lovestone, S. Insulin signalling in Alzheimer′s disease and diabetes: from epidemiology to molecular links. *Journal of Internal Medicine*, 280, 5 (2016), 430-442.
32 Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segre, A. V., Steinthorsdottir, V., et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44, 9 (2012), 980-981.

33 Luciano, M., Hansell, N. K., Lahti, J., Davies, G., Medland, S. E., Räikkönen, K., et al. Whole genome association scan for genetic polymorphisms influencing information processing speed. *Biological psychology*, 86, 3 (2011), 193-202.
34 Giedraitis, V., Kilander, L., Degerman-Gunnarsson, M., Sundelöf, J., Axelsson, T., Syvänen, A. C., et al. Genetic analysis of Alzheimer's disease in the Uppsala Longitudinal Study of Adult Men. *Dementia and geriatric cognitive disorders*, 27, 1 (2009), 59-68.
35 Li, H., Wetten, S., Li, L., Jean, P. L. S., Upmanyu, R., Surh, L., et al. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Archives of neurology*, 65, 1 (2008), 45-53.
36 Oláh, J., Vincze, O., Virók, D., Simon, D., Bozsó, Z., Tőkési, N., et al. Interactions of Pathological Hallmark Proteins: Tubulin Polymerization Promoting Protein/p25, β-Amyloid, AND α-Synuclein. *Journal of Biological Chemistry*, 286, 39 (2011), 34088-34100.
37 Li, M. O., & Flavell, R. A. Contextual regulation of inflammation: a duet by transforming growth factor-β and interleukin-10. *Immunity*, 28, 4 (2008), 468-476.
38 Matarin, M., Salih, D. A., Yasvoina, M., Cummings, D. M., Guelfi, S., Liu, W., et al. A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. *Cell reports*, 10, 4 (2015), 633-644.
39 Guillot-Sestier, M. V., Doty, K. R., Gate, D., Rodriguez, J., Leung, B. P., Rezai-Zadeh, K., et al. Il10 deficiency rebalances innate immunity to mitigate Alzheimer-like pathology. *Neuron*, 85, 3 (2015), 534-548.
40 Kajkowski, E. M., Lo, C. F., Ning, X., Walker, S., Sofia, H. J., Wang, W., et al. β-Amyloid peptide-induced apoptosis regulated by a novel protein containing a G protein activation module. *Journal of Biological Chemistry*, 276, 22 (2011), 18748-18756.
41 Stage, E, Duran, T, Risacher, SL, Goukasian, N, Do, TM, West, JD, Wilhalme, H, Nho, K, Phillips, M, Elashoff, D, Saykin, AJ & Apostolova, LG. The effect of the top 20 Alzheimer disease risk genes on gray-matter density and FDG PET brain metabolism. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring*, 5 (2016), 53-66.
42 Mendelson, A. F., Zuluaga, M. A., Lorenzi, M., Hutton, B. F., Ourselin, S.. Selection bias in the reported performances of AD classification pipelines. *NeuroImage: Clinical*, 14 (2016), 400-416.
43 Fischl, B. FreeSurfer. In *NeuroImage* 62, 2 (2012), 774-781.
44 Gutman BA, Wang Y, Rajagopalan P, Toga AW, Thompson PM. Shape matching with medial curves and 1-D group-wise registration. In *9th IEEE International Symposium on Biomedical Imaging (ISBI)* (2012), IEEE, 716-719.
45 Gutman BA, Madsen SK, Toga AW, Thompson PM. A family of fast Spherical registration algorithms for cortical shapes. In *International Workshop on Multimodal Brain Image Analysis* (2013), Springer International Publishing, 246-257.
46 Roshchupkin GV, Gutman BA, Vernooij MW, Jahanshad N, Martin NG, et al. Heritability of the shape of subcortical brain structures in the general population. *Nature Communications*, 7 (2016).
47 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. PLINK: a tool set for whole-genome association and population-based linkage

analyses. The American Journal of Human Genetics. *Am J Hum Genet*, 81, 3 (2007), 559-575.
48 Wold, H. *Estimation of principal components and related models by iterative least squares. Multivariate analysis*. Academic Press, NewYork, 1966.
49 Martens, Harald, and Tormod Naes. *Multivariate calibration*. Springer, Dordrecht, 1984.
50 McIntosh, A. R., & Lobaugh, N. J. Partial least squares analysis of neuroimaging data: applications and advances. *NeuroImage*, 23 (2004), S250-S263.
51 McIntosh, A. R., Bookstein, F. L., Haxby, J. V., & Grady, C. L. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, 3, 3 (1996), 143-157.
52 Worsley, K. J. An overview and some new developments in the statistical analysis of PET and fMRI data. *Human Brain Mapping*, 5, 4 (1997), 254-258.
53 Lorenzi, M., Gutman, B., Thompson, P. M., Alexander, D. C., Ourselin, S., & Altmann, A. Secure multivariate large-scale multi-centric analysis through on-line learning: an imaging genetics case study. In *12th International Symposium on Medical Information Processing and Analysis* (2017), Society of Photo-Optical Instrumentation Engineers.
54 Friston, K. J., Frith, C. D., Liddle, P. F., & Frackowiak, R. S. J. Functional connectivity: the principal-component analysis of large (PET) data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13, 1 (1993), 5-14.
55 Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., et al. The Ensembl gene annotation system. *Database (Oxford).* , baw093 (2016).
56 Kimbrough, I. F., Robel, S., Roberson, E. D., & Sontheimer, H. Vascular amyloidosis impairs the gliovascular unit in a mouse model of Alzheimer's disease. *Brain*, 138, 12 (2015), 3716-3733.
57 Luchsinger, J. A., Gustafson, D. R. Oxidative Stress, Reactive Metabolites, Inflammation, and RAGE – Building a Bridge from Alzheimer's Disease to Diabetes and Vice Versa. *Journal of Alzheimer's Disease*, Journal of Alzheimer's Disease, 4 (2009), 693-704.
58 Ferreira, S. T., Clarke, J. R., Bomfim, T. R., & De Felice, F. G. Inflammation, defective insulin signaling, and neuronal dysfunction in Alzheimer's disease. *Alzheimer's & dementia*, 10, 1 (2014), S76-S83.

Figure legends

Figure 1. *Cross-validation scheme for the assessment of the genetic loci of maximal genotype-phenotype correlation identified by the PLS model.* The whole procedure is repeated 1,000,000 times, and the resulting array is further analyzed. **a)** PLS is applied in a split-half setting. For each of the two non-overlapping randomly sampled groups, the PLS components of joint phenotype and genotype variation are independently estimated. **b)** Left. Each chromosome is partitioned in bins of 10k base-pairs size, which are labeled 1 if they contain a SNP associated to the largest PLS weights (top 10% of absolute values), or 0 otherwise. To obtain stable estimates of the loci of maximal weights, the resulting binary arrays independently estimated in the two groups are merged (bin-wise AND operation). The same procedure is applied on the mesh-based PLS weights associated to the phenotype component. **c)** Steps a) and b) are repeated across 1,000,000 folds, and the results are subsequently averaged to obtain the confidence maps associated to genetic and phenotype components (figures 2 and 3).

Figure 2. *PLS genotype component:* the outer circular plots show the probability of a given genetic locus to be associated with the phenotype components shown in Figure 3. The plots show the probability of a given genetic bin of size 10kb of being relevant in the PLS model, i.e., to contain a SNP that is ranked in the top 10% of the absolute weights of the genotype component. Spatially contiguous loci generally show similar importance values, which is caused by LD of these regions. The genes close to the important loci ($p>0.95$) are listed in the innermost circle depending on their genomic position; genes with eQTLs are highlighted by red font. The inner circular plots show the PLS weights associated to each genetic locus (red: positive, blue: negative). Loci with large absolute weight value are usually characterized by high relevance. The red radial lines are located in correspondence of known AD genes: *ABCA7, APOE, APP,*

BIN1, CASS4, CD2AP, CD33, CELF1, CLU, CR1, DSG2, EPHA1, FERMT2, HLA-DRB5, INPP5D, MAPT, MEF2C, MS4, NME8, PICALM, PSEN1, PSEN2, PTK2B, SLC24A4, SORL1, ZCWPW1. High-resolution circular plots for each component are provided in Supplementary Figures S2, S3 and S4.

Figure 3. *PLS Phenotype component:* the figures in the top row show the topographical distribution of the PLS weights associated to the cortical and subcortical brain areas. The absolute value of the weights is proportional to the importance of the underlying brain areas. The relevance of the brain areas is quantified in the bottom row. The colours (red to white) indicate the probability of a brain area to be associated with the genotype component shown in Figure 2, and quantify the probability of each cortical mesh points of being relevant in the PLS model, i.e., to be ranked among the top 10% of the absolute weights of the phenotype component.