Mining Outcome-relevant Brain Imaging Genetic Associations via Three-way Sparse Canonical Correlation Analysis in Alzheimer’s Disease

Citation
Hao, X., C. Li, L. Du, X. Yao, J. Yan, S. L. Risacher, A. J. Saykin, et al. 2017. “Mining Outcome-relevant Brain Imaging Genetic Associations via Three-way Sparse Canonical Correlation Analysis in Alzheimer’s Disease.” Scientific Reports 7 (1): 44272. doi:10.1038/srep44272. http://dx.doi.org/10.1038/srep44272.

Published Version
doi:10.1038/srep44272

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:32071994

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story
The Harvard community has made this article openly available. Please share how this access benefits you. Submit a story.

Accessibility
Mining Outcome-relevant Brain Imaging Genetic Associations via Three-way Sparse Canonical Correlation Analysis in Alzheimer’s Disease

Xiaoke Hao, Chanxiu Li, Lei Du, Xiaohui Yao, Jingwen Yan, Shannon L. Risacher, Andrew J. Saykin, Li Shen, Daoqiang Zhang & Alzheimer’s Disease Neuroimaging Initiative

Neuroimaging genetics is an emerging field that aims to identify the associations between genetic variants (e.g., single nucleotide polymorphisms (SNPs)) and quantitative traits (QTs) such as brain imaging phenotypes. In recent studies, in order to detect complex multi-SNP-multi-QT associations, bi-multivariate techniques such as various structured sparse canonical correlation analysis (SCCA) algorithms have been proposed and used in imaging genetics studies. However, associations between genetic markers and imaging QTs identified by existing bi-multivariate methods may not be all disease specific. To bridge this gap, we propose an analytical framework, based on three-way sparse canonical correlation analysis (T-SCCA), to explore the intrinsic associations among genetic markers, imaging QTs, and clinical scores of interest. We perform an empirical study using the Alzheimer’s Disease Neuroimaging Initiative (ADNI) cohort to discover the relationships among SNPs from AD risk gene APOE, imaging QTs extracted from structural magnetic resonance imaging scans, and cognitive and diagnostic outcomes. The proposed T-SCCA model not only outperforms the traditional SCCA method in terms of identifying strong associations, but also discovers robust outcome-relevant imaging genetic patterns, demonstrating its promise for improving disease-related mechanistic understanding.

Alzheimer’s disease (AD) is the most common form of dementia characterized by progressive impairment of memory and other cognitive functions in people over 65 years of age. It is an important research topic to develop methods for early diagnosis of AD. At present, many studies have focused on searching for biomarkers from brain imaging data as well as molecular and cellular data to investigate the pathological changes. To further aid the development of effective diagnostic and therapeutic approaches, it has received increasing attention to study AD at the system biology level. For example, revealing the biological pathway from the microcosmic genetic factors to macroscopic brain anatomy has the potential to understand pathogenicity mechanisms underlying the disordered cognition and behavior.

Enabled by recent advances in high-throughput genotyping and multimodal neuroimaging technologies, imaging genetics is becoming an emerging research field for discovering the associations between genetic markers such as single nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from structural or functional neuroimaging data. Thus, it holds great promise for us to understand the complex neurogenetic and neurobiological mechanism of complex brain disorders.

In prior imaging genetics studies, univariate and multivariate regression methods have been typically used to capture the effective associations between SNPs and neuroimaging data. More recently, bi-multivariate...
analysis techniques such as various structured sparse canonical correlation analysis (SCCA) models have attracted increasing attention in brain imaging genetics to detect complex multi-SNP-multi-QT associations. The essence of all the structured SCCA approaches, incorporating valuable prior knowledge, is to find the best linear transformations for imaging and genetic features respectively so that the strongest correlation between the imaging and genetic components can be achieved. These methods have the potential to discover effective imaging genetic associations, while the identified genotypic and phenotypic markers may not be disease specific. To overcome this limitation, in this paper, we focus on exploring three-way associations among genetic markers, imaging QTs, and cognitive and diagnostic outcomes. Our goal is to reveal the relationships among these multidimensional genetics, imaging and outcome data, and contribute to a better understanding of pathogenicity mechanisms in AD.

Of note, some diagnosis information guided methods have been proposed in the field of imaging genetics. The sRRR model proposed in used a two-step procedure for detecting genetic factors associated with disease relevant imaging phenotypes by using penalized linear discriminant analysis. More recently, a Bayesian framework was used to select the relevant features along the pathway from gene to imaging and then to symptom. It is worth noting that both models treated diagnosis information as binary status (e.g., AD and normal control (NC)) for imaging genetic association studies. Actually, in the spectrum between NC and AD, there exist other progressive stages. For example, in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, there were participant groups labeled as Significant Memory Concern (SMC), Early Mild Cognitive Impairment (EMCI) and Late Mild Cognitive Impairment (LMCI). In addition, the cognitive scores such as Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), ADNI Memory (ADNI-MEM) and ADNI Executive Functioning (ADNI-EF), which are neuropsychological assessment measures from different aspects, are often used as quantitative descriptions of symptom severity instead of binary diagnosis. Accordingly, identification of imaging genetic associations relevant to these diagnostic and cognitive outcomes may yield important information for a better understanding of disease-specific mechanisms.

With these observations, we consider the outcome-relevant imaging genetic association study as a multi-view multivariate correlation problem, which can be solved by CCA and partial least squares (PLS), as well as their sparse versions (including SCCA and SPLS). Thus, following the existing imaging genetic studies via bi-multivariate SCCA, we propose a three-way sparse canonical correlation analysis (T-SCCA) framework to explore the intrinsic associations among SNP loci, neuroimaging features and clinical score outcomes. Specifically, in this study, the outcomes of interest include cognitive scores (CS) and diagnosis status (DS). We evaluate the effectiveness of the proposed method by identifying three-way associations among 85 candidate SNPs from the top AD risk gene APOE, imaging QTs extracted from structural magnetic resonance imaging (MRI) scans, and relevant cognitive and diagnostic outcomes, using the Alzheimer's Disease Neuroimaging Initiative (ADNI) data as a test bed. The experimental results demonstrate that the proposed T-SCCA model not only outperforms the standard two-way SCCA method in terms of identifying strong associations, but also discovers robust outcome-relevant imaging genetic patterns, demonstrating its promise for improving disease-related mechanistic understanding.

**Imaging genetic associations**

**Imaging genetic associations via bi-multivariate analysis.** We first describe relevant notation. We use lowercase letters to denote vectors, and uppercase letters to denote matrices. For a given matrix $X = [x_{ij}]$, we denote its $i$-th row and $j$-th column as $x_i$ and $x_j$, respectively. Let $X = [x_1, ..., x_n] \in \mathbb{R}^{n \times p}$ be the SNP genotype data, $Y = [y_1, ..., y_n] \in \mathbb{R}^{n \times r}$ be the imaging QT data (i.e., voxel-based morphometry measures in this work), where $n$ is the number of participants, and $p$ and $q$ are the number of SNPs and QTs, respectively.

For detecting complex multi-SNP-multi-QT associations, sparse canonical correlation analysis (SCCA) seeks linear transformations of variables $X$ and $Y$ to achieve the maximal correlation between $Xw_1$ and $Yw_2$. We introduce penalty terms simultaneously, which can be formulated as:

$$\max_{w_1, w_2} \mathbf{w}_1^T X^T Y \mathbf{w}_2$$

s.t. $w_1^T X^T X w_1 \leq 1$, $w_2^T Y^T Y w_2 \leq 1$, $\|w_1\|_1 \leq c_1$, $\|w_2\|_1 \leq c_2$  \hspace{1cm} (1)

where $w_1$ and $w_2$ are canonical loadings or weights, reflecting the contribution of each feature in the identified canonical correlation. Note that $w_1^T X^T X w_1 \leq 1$, $w_2^T Y^T Y w_2 \leq 1$ are used to embrace the covariance structure of the data in the model, $\|w_1\|_1 \leq c_1$, $\|w_2\|_1 \leq c_2$ are constraints for controlling the sparsity so that only a small number of relevant features will be selected automatically from the SNP and imaging data.

**Cognitive score (CS)-relevant imaging genetic associations via three-way SCCA.** In this study, for revealing the biological mechanism specific to the disease, we aim to discover imaging genetic associations that are relevant to cognitive scores (CSs) or diagnosis status. The first attempt focuses on involving multi-assessment CSs (i.e., MMSE, CDR, ADNI-MEM and ADNI-EF). Let $Z = [z_1, ..., z_r] \in \mathbb{R}^{n \times r}$ be the CSs, where $n$ is the number of participants, $r$ is the number of CSs. Since different CSs can provide complementary perspectives on neuropsychological assessments, we aim to seek a set of linear transforms to estimate the contribution of each individual cognitive score in imaging genetic associations.

The formulation of CS-relevant imaging genetic associations can be extended from eq. (1) as follows:
is a newly added constraint for selecting a small number of CS that are related to both imaging and genetics measures.

$$\text{Maximize} \quad w_1^T X^T Y w_2 + w_1^T X^T Z w_3 + w_2^T Y^T Z w_3$$

subject to:

$$w_1^T X w_1 \leq 1, \quad w_2^T Y w_2 \leq 1, \quad w_3^T Z w_3 \leq 1,$$

$$\|w_1\| \leq c_1, \quad \|w_2\| \leq c_2, \quad \|w_3\| \leq c_3$$

(2)

where $w_i$ is the canonical loading, reflecting the contribution of each neuropsychological assessment in the identified canonical correlation. Similar to the existing constraints in eq. (1), $w_3^T Z^T Z w_3 \leq 1$ is a newly added constraint to embrace the covariance structure, and $\|w_3\| \leq c_3$ is another newly added constraint for selecting a small number of CS that are related to both imaging and genetics measures.

**Diagnosis status (DS)-relevant imaging genetic associations via three-way SCCA.** Besides the CSs, the diagnosis status (DS) considered as a qualitative measurement can also describe the progressive stages of AD. Accordingly, we propose another outcome-relevant imaging genetic association that involves DS. Let $Z \in \mathbb{R}^{n \times 1}$ be the DS labels for all participants. The DS-relevant formulation, similar to eq. (2), can be used for mining DS-relevant imaging genetic associations.

It is worth noting that the equation of DS-relevant imaging genetic association is a special case of eq. (2) if $r=1$. Similar to eq. (2), $w_i$ is the canonical loading reflecting the contribution of DS in the identified canonical correlation, so that the optimization of $w_1, w_2$ and $w_3$ can be also solved by the proposed algorithm.

**Results**

**Characteristics of the dataset.** The dataset comprises of 913 non-Hispanic Caucasian participants, including 211 normal control (NC), 82 significant memory concern (SMC), 273 early mild cognitive impairment (EMCI), 187 late mild cognitive impairment (LMCI) and 160 AD. In our experiments, we used the baseline structural MRI data with average voxel-based morphometry (VBM) measures in 116 regions of interest (ROIs), genotyping data with 85 candidate SNP loci, as well as cognitive scores (including MMSE, CDR, ADNI-MEM and ADNI-EF) and diagnosis status (including NC, SMC, EMCI, LMCI and AD). For more details about the demographic and data processing, please see the complete information in the Methods section.

**Experimental settings.** In this imaging genetic association study, 5-fold cross-validation strategy is adopted to evaluate the effectiveness of our proposed method. For parameters of regularization, we determine their values by another nested 5-fold cross-validation on the training set. It is used to fine tune the parameters in the objective function in the range of $\{10^{-3}, 3 \times 10^{-3}, 10^{-2}, 3 \times 10^{-2}, 10^{-1}, 3 \times 10^{-1}, 1, 3, 10, 30, 100\}$. The parameters yielding the best performance in the inner cross-validation are finally used in the resulting model.

In the current experiments, we compare BM-SCCA (denoting conventional bi-multivariate SCCA), CS-SCCA (denoting cognitive scores-guided method via three-way SCCA), DS-SCCA (denoting diagnosis-guided method via three-way SCCA). Both CS-SCCA and DS-SCCA belong to the T-SCCA category.

**Improved association between risk SNPs and phenotypic imaging markers.** We compare our proposed T-SCCA methods (including CS-SCCA and DS-SCCA) with the conventional method (BM-SCCA). The performance on each dataset is assessed using the correlation coefficient (CC) between SNP and imaging data, which is widely used for association analysis measurements. The average results of CC across the 5-fold training and testing data are calculated respectively. As shown in Table 1 and Table 2, CS-SCCA and DS-SCCA yield the CC values of 0.2633 (0.3667) and 0.2711 (0.3723) on test (training) set, respectively, which are better than those of BM-SCCA. These results indicate that the search space of BM-SCCA could be too large such that the algorithm could converge to local optima without prior knowledge, while the regularizations of the restrictions

| Method     | F1     | F2     | F3     | F4     | F5     | Mean + Std |
|------------|--------|--------|--------|--------|--------|------------|
| BM-SCCA   | 0.2619 | 0.2810 | 0.1846 | 0.2679 | 0.2755 | 0.2542 ± 0.0396 |
| CS-SCCA   | 0.3436 | 0.3819 | 0.3743 | 0.3536 | 0.3798 | 0.3667 ± 0.0171 |
| DS-SCCA   | 0.3519 | 0.3843 | 0.3848 | 0.3584 | 0.3822 | 0.3723 ± 0.0159 |

Table 2. 5-fold cross-validation results on ADNI: The model learned from the training data is used to estimate the correlation coefficients on the testing set.
on outcome-relevant information might be able to guide imaging genetic associations out of over-fitting. This demonstrates the disease information could help improve the performances of correlations between genotypes and imaging phenotypes.

In addition, we perform a permutation test using 1000 permutations with retraining each of BM-SCCA, SC-SCCA and DS-SCCA models to assess the statistical significance of the identified correlations on the test set. The p-value corresponds to the fraction of times that the correlation coefficient is greater or equal to the result from original data. The resulting p-values ($p < 0.001$) are statistically significant in all three cases, including BM-SCCA, CS-SCCA and DS-SCCA, respectively.

Besides improving correlation performances, one major goal of this study is to identify genotypic and phenotypic markers that are not only highly correlated to each other, but also relevant to cognitive or diagnostic outcomes specific to AD. Figure 1 shows the heat map of average estimated canonical loadings on 85 APOE SNPs associated with 116 brain ROIs across 5-fold cross-validation respect to different methods.

Figure 1. Heat map of average estimated canonical loadings on 85 APOE SNPs associated with 116 brain ROIs across 5-fold cross-validation respect to different methods.
Discussion

As expected, the well-known locus rs429358 is identified to be associated with gray matter loss in multiple AD-relevant ROIs, which is in accordance with the previous studies. The C allele increases the risk of AD in APOE e4, which is encoded by rs429358 (www.snpedia.com/index.php/APOE) (www.alzgene.org). BM-SCCA seems to yield other loci such as rs405697 (this genetic variant in APOE region has shown only to be associated with human longevity in the literature) and rs157594.

For the phenotype identifications, besides the well-known AD-related ROI left hippocampus, signals in the cerebellum and vermis areas are detected by BM-SCCA. However, the morphometric changes of the cerebellum have not been widely validated as AD biomarkers. On the other hand, a few additional ROIs such as right amygdala, right calcine, right cuneus, left and right frontal-sup-medial gyrus, right parahippocampal gurus have been detected as top 10 features associated with the risk genotype biomarker rs429358 by the proposed T-SCCA (including CS-SCCA and DS-SCCA). It's worth noting that CS-SCCA captures similar patterns on brain imaging as DS-SCCA does for the most part. More interestingly, the same ROIs selected simultaneously by two types of T-SCCA have similar weight values, which demonstrates the robust and consistent biomarker findings. Note that these weights have different signs, which are caused by the negative directionality of the cognitive score values in relation to the diagnosis values. The top 10 selected MRI-VBM imaging features, as well as their averaged estimated canonical loadings generated by T-SCCA (combination of SC-SCCA and DS-SCCA) across 5 cross-validation trials, are visualized in Fig. 2 by mapping them onto the human brain. The colors of the selected brain regions indicate the canonical loadings of the corresponding markers. The identified regions have potential clinical correlates in typical clinically well-described AD impairments. To our knowledge, the hippocampus is one of the first regions of the brain to suffer damage including memory loss and disorientation. In addition, amygdala atrophy is related to aberrant motor behavior, with potential relationships to anxiety and irritability. Some existing results suggest that the magnitude of amygdala atrophy is comparable to that of the hippocampus in the earliest clinical stages of AD. The analytical result reassures that our method identifies a well-known correlation between genotypes and phenotypes that is severely and consistently affected by pathology in AD. Besides confirming the prior findings, our method also yields the associations between APOE rs429538 and other eminent AD markers such as both left and right frontal-sup-medial gyrus. There also appear to be specific relationships among genotypes, phenotypes and neuropsychiatric symptoms that deserve further investigation.

As mentioned earlier, the quantitative CSs (MMSE, CDR, ADNI-MEM and ADNI-EF) used to index cognitive decline for disease severity are able to show the graded differences in participants, and we induce the sparsity constraint for selecting the related CSs in this imaging genetic associations study. The CS-SCCA yields the loading values of 0.0095, −0.0103, 0.0238, and −0.0026 on MMSE, CDR, ADNI-MEM and ADNI-EF, respectively.
It demonstrates ADNI-MEM is the top-ranked score that contributes to the multiple associations among gene, neuroimaging and cognition. Since all CSs are neuropsychological assessment measures from different aspects, the CS-SCCA provides a simple evaluating approach to investigate the relative contribution of each clinical score. In summary, we have performed a neuroimaging genetics study for Alzheimer's disease (AD) to explore the relationship between genetic variations in the APOE gene and brain ROIs measured by voxel based morphometry (VBM). Various existing sparse canonical correlation analysis (SCCA) methods are only designed for bi-multivariate analysis, and often yield suboptimal results without considering the cognitive or diagnostic outcomes. With these observations, we have investigated a three-way sparse canonical correlation analysis (T-SCCA) framework to discover the multiple associations among SNP loci, neuroimaging features and phenotypic outcomes (including cognitive scores (CSs) and diagnosis status (DS)). The experimental results performed on 913 subjects from ADNI show that our proposed T-SCCA model can substantially achieve higher correlations between genotypic and phenotypic features. Specifically, besides the improved correlation performances, CS-SCCA captures similar patterns on canonical loadings as DS-SCCA. This supports the benefit of our general T-SCCA model that can also identify some significant and robust biomarkers in imaging genetic associations, revealing disease-specific patterns on the complex mechanisms. However, there could be different mechanisms leading to three-way associations between genetics, imaging and diagnosis. For example, on top of possible pathway from genetics to imaging and then to disease, part of the genetic influences on disease might not be mediated through features captured by neuroimaging. In addition, genetics might independently affect disease susceptibility and imaging features, resulting in association between imaging and diagnosis, or there might be hidden or confounding variables that drive these associations. Thus, it warrants further investigation to reveal the underlying mechanisms related to the three-way associations discovered by our proposed T-SCCA methods.

In this initial study, where the number of samples exceeds the number of total features, the T-SCCA model can be successfully applied for association discovery coupled with feature selection. However, if the dataset has more features than samples, this ill-conditioned problem can be addressed via dimensionality reduction or regularization. In particular, when the datasets contain far more features (e.g., SNPs at the genome-wide magnitude), it will greatly increase the computational complexity and memory requirement. Note that the normal equation in the optimization contains matrix inversion operations (the time complexity is $O(n^3)$, where $n$ is the number of features). Therefore, it is an interesting future topic to develop a more efficient solution for our proposed T-SCCA and to identify potential markers from high-throughput genome-wide variants and neuroimaging quantitative traits in outcome-relevant imaging genetic studies.

In addition, in this study, we have explored the imaging genetic associations within a single population of non-Hispanic Caucasians. However, the effect of population structure is another important topic, and it may affect the identifications in multivariate associations due to the potential bias introduced by multiple populations in a study. In this case, the population structure adjustment should be considered in the study.

**Methods**

**Subjects.** Data used in the preparation of this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership by several organizations, including the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The study protocols were approved by the institutional review boards of all participating centers (Nanjing University of Aeronautics and Astronautics, Northwestern Polytechnical University, Indiana University and ADNI (A complete list of ADNI sites is available at http://www.adni-info.org/)) and written informed consent was obtained from all participants or authorized representatives. All the analytical methods were performed on the de-identified ADNI data, and were determined by Indiana University Human Subjects Office as IU IRB Review Not Required. In addition, these methods were carried out in accordance with the approved guidelines.

Participants were screened and enrolled according to criteria demonstrated in ADNI study protocol (http://adni-info.org/Scientists/ADNISTudyProcedures.html#). The general inclusion/exclusion criteria of the subjects from ADNI procedures manual (http://www.adni-info.org) are briefly described as follows:

1. NC participants have no subjective or informant-based complaint of memory decline and normal cognitive performance. The MMSE scores on NC should be between 24 and 30, CDR should be 0.
2. SMC participants have subjective memory concerns as assessed using the Cognitive Change Index (CCI; total score from first 12 items >16), no informant-based complaint of memory impairment or decline, and normal cognitive performance on the Wechsler Logical Memory Delayed Recall (LM-delayed) and the MMSE<sup>2</sup>.
3. EMCI participants have a memory concern reported by the subject, informant, clinician, abnormal memory function approximately 1 standard deviation below normative performance adjusted for education level on the LM-delayed, an MMSE total score greater than 24.
4. Besides a subjective memory concern as reported by subject, study partner or clinician, CDR on LMCI subjects is 0.5 and Memory Box (MB) score must be at least 0.5.
5. MMSE score on AD should be between 20 and 26, and CDR should be 0.5 or 1.0.

In the practical diagnosis of AD, multiple clinical variables are generally acquired, e.g., MMSE, CDR, ADNI-MEM (composite score for memory) and ADNI-EF (composite score for executive functioning), etc.
let them be zero: the models, each SNP value was coded in an additive fashion as 0, 1 or 2, indicating the number of minor alleles. For the input in gene boundary based on the ANNOVAR (http://annovar.openbioinformatics.org) annotation, which include a total number of 85 SNPs as candidates. For the input in 20 k base pairs of the analysis on all SNPs within 2 visit scan, and then created normalized gray matter density maps from the MRI data in the standard Montreal Neurological Institute (MNI) space as 2 × 2 × 2 mm³ voxels SPM software package. 116 ROI level measurements of mean gray matter densities were further extracted based on the MarsBaR AAL atlas. All measurements were pre-adjusted for age, gender, and education.

Table 3. Characteristics of the subjects. Note: NC = Normal Control, SMC = Significant Memory Concern, ECMI = Early Mild Cognitive Impairment, LCMI = Late Mild Cognitive Impairment, AD = Alzheimer’s disease.

| Subjects | NC | SMC | EMCI | LMCI | AD |
|----------|----|-----|------|------|----|
| Number   | 211| 82  | 273  | 187  | 160|
| Gender(M/F) | 109/102 | 35/49 | 153/120 | 108/79 | 95/65|
| Age      | 76.14±6.53 | 72.45±5.67 | 71.48±7.12 | 73.86±8.44 | 75.18±7.88|
| Education| 16.45±2.62 | 16.78±2.67 | 16.08±2.62 | 16.38±2.81 | 15.86±2.75|
| MMSE     | 29.01±1.23 | 29.00±1.22 | 28.38±1.54 | 27.71±1.73 | 24.00±2.62|
| CDR      | 0.01±0.07 | 0.00±0.00 | 0.49±0.08 | 0.49±0.07 | 0.72±0.27|
| ADNI-MEM | 1.02±0.58 | 1.12±0.57 | 0.60±0.60 | 0.07±0.67 | -0.76±0.61|
| ADNI-EF | 0.85±0.69 | 0.73±0.81 | 0.51±0.74 | 0.18±0.81 | -0.53±0.91|

SNP genotype data. Since genetic risk factors can help scientists focus on relevant biological pathways and form effective hypothesis for drug design, identifying risk genetic markers associated with brain imaging can help understand the underlying biological mechanisms. We downloaded the ADNI-GO/2 genotyping data, and performed quality control and population stratification using the approach described in the previous study. Specifically, MMSE is used to examine functions including registration, attention and calculation, recall, language, ability to follow simple commands and orientation. CDR is a numeric scale used to assess a patient’s cognitive and functional performance in six areas: memory, orientation, problem solving, community affairs, hobbies and personal care. There are two derived composite scores for MEM and EF from ADNI. The formation of ADNI-MEM is complicated by the use of different word lists in the Rey Auditory Verbal Learning Test (RAVLT) and the ADAS-Cog, and by Logical Memory I data missing by design. The formation of ADNI-EF includes Category Fluency-animals, Category Fluency-vegetables, Trails A and B, Digit span backwards, WAIS-R Digit Symbol Substitution, and 5 Clock Drawing items (circle, symbol, numbers, hands, time).

Imaging phenotype data. The MRI data used in this paper were also obtained from the ADNI database. We aligned the preprocessed imaging data (i.e., voxel based morphometry (VBM)) to each participant’s same visit scan, and then created normalized gray matter density maps from the MRI data in the standard Montreal Neurological Institute (MNI) space as 2 × 2 × 2 mm³ voxels SPM software package. 116 ROI level measurements of mean gray matter densities were further extracted based on the MarsBaR AAL atlas. All measurements were pre-adjusted for age, gender, and education.

Objective function and algorithm design. In this section, we design an algorithm to solve the optimization problem defined in eq. (2). For the general formulation, using the Lagrange multiplier and writing the penalization problem defined in eq. (2). For the general formulation, using the Lagrange multiplier and writing the penalties into the matrix form, the objective function for mining CS-relevant or DS-relevant imaging genetic associations via three-way sparse canonical correlation analysis (T-SCCA) is as follows:

$$L(w_1,w_2,w_3) = -w_1^T X^T Y w_2 - w_1^T X^T Z w_3 - w_2^T Y^T Z w_3 + \frac{1}{2} \|Xw_1\|^2 + \frac{1}{2} \|Yw_2\|^2$$

$$+ \frac{1}{2} \|Zw_3\|^2 + \beta_1 \|w_1\|_1 + \beta_2 \|w_2\|_1 + \beta_3 \|w_3\|_1$$

(3)

where $({\beta_1},{\beta_2},{\beta_3})$ and are the set of model parameters. Take the derivative regarding $w_1$, $w_2$ and $w_3$ separately and let them be zero:

$$w_1 = (X^T X + \beta_1 D_1)^{-1} X^T (Y w_2 + Z w_3)$$

(4)

$$w_2 = (Y^T Y + \beta_2 D_2)^{-1} Y^T (X w_1 + Z w_3)$$

(5)

$$w_3 = (Z^T Z + \beta_3 D_3)^{-1} Z^T (X w_1 + Y w_2)$$

(6)

where $D_i$ is a diagonal matrix with the $k_i$-th element as $1$ ($k_i \in [1, p]$), $D_2$ is a diagonal matrix with the $k_2$-th element as $1$ ($k_2 \in [1, q]$), and $D_3$ is a diagonal matrix with the $k_3$-th element as $1$ ($k_3 \in [1, r]$).
Since $D_1$ relies on $w_1$, $D_2$ relies on $w_2$, and $D_3$ relies on $w_3$, we introduce an iterative procedure to solve this objective. In each iteration, we first fix $w_0$ and $w_3$ to solve $w_1$, then fix $w_1$ and $w_3$ to solve $w_2$, and finally fix $w_1$ and $w_2$ to solve $w_3$. The procedure stops until it satisfies a predefined stopping criterion. Algorithm 1 shows the pseudo code of the T-SCCA algorithm for mining outcome-relevant imaging genetic associations.

### Algorithm 1: T-SCCA for Mining Outcome-relevant Imaging Genetic Associations

**Input:** candidate SNPs $X=[x_1, \ldots, x_s] \in \mathbb{R}^{s \times n}$
neuroimaging ROIs $Y=[y_1, \ldots, y_r] \in \mathbb{R}^{r \times n}$
cognitive scores (CS) $Z=[z_1, \ldots, z_k] \in \mathbb{R}^{k \times n}$ or diagnosis status (DS) ($r=1$ for $Z$

**Ensure:** canonical vectors $w_0, w_1, w_2, w_3$

**Initialization:** $w_0 \in \mathbb{R}^{n \times 1}, w_1 \in \mathbb{R}^{n \times 1}, w_2 \in \mathbb{R}^{n \times 1}$

**While** not converged regarding to $w_0, w_1, w_2, w_3$

- Calculate the diagonal matrix $D_1$, where the $k$-th element is $\frac{1}{2 \| y_k \|^2}$
- Update $w_1 = (X'X + \lambda D_1)^{-1}X'(Yw_2 + Zw_3)$
- Scale $w_1$ so that $\|Xw_1\| = 1$

- Calculate the diagonal matrix $D_2$, where the $k$-th element is $\frac{1}{2 \| y_k \|^2}$
- Update $w_2 = (Y'Y + \lambda D_2)^{-1}Y'(Xw_1 + Zw_2)$
- Scale $w_2$ so that $\|Yw_2\| = 1$

- Calculate the diagonal matrix $D_3$, where the $k$-th element is $\frac{1}{2 \| y_k \|^2}$
- Update $w_3 = (Z'Z + \lambda D_3)^{-1}Z'(Xw_1 + Yw_2)$
- Scale $w_3$ so that $\|Zw_3\| = 1$

**End while**

### References

1. Brookmeyer, R., Johnson, E., Ziegler-Graham, K. & Arrighi, H. M. Forecasting the global burden of Alzheimer’s disease. *Alzheimers Dement* 3, 186–191 (2007).
2. Pasinetti, G. M. & Hiller-Sturmfholz, S. Systems biology in the study of neurological disorders: focus on Alzheimer’s disease. *Alzheimers Dement* 5, 3–9 (2009).
3. Glahn, D. C., Thompson, P. M. & Blangero, J. Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function. *Hum Brain Mapp* 28, 488–501 (2007).
4. Gottfman, I. I. & Gould, T. D. The endophenotype concept in psychiatry: Etymology and strategic intentions. *Am J Psychiat* 160, 636–645 (2003).
5. Tian Ge, G. S., Jianfeng Feng. Imaging genetics—towards discovery neuroscience. *Quant. Biol. 1*, 227–245 (2013).
6. Shen, L. et al. Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *Neuroimage* 53, 1051–1063 (2010).
7. Stein, J. L. et al. Voxelwise genome-wide association study (VGWAS). *Neuroimage* 53, 1160–1174 (2010).
8. Hiba, D. P., Kohannim, O., Stein, J. L., Chiang, M. C. & Thompson, P. M. Multilocus genetic analysis of brain images. *Front Genet* 2, 73 (2011).
9. Kohannim, O. et al. Predicting Temporal Lobe Volume on MRI from Genotypes Using L1(1)-L2(2) Regularized Regression. *Proc IEEE Int Symp Biomed Imaging*. 1160–1163 (2012).
10. Wang, H. et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics* 28, 229–237 (2012).
11. Chi, E. C. et al. Imaging Genetics Via Sparse Canonical Correlation Analysis. *Proc IEEE Int Symp Biomed Imaging*. 740–743 (2013).
12. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10, 515–534 (2009).
13. Witten, D. M. & Tibshirani, R. J. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Stat Appl Genet Mol Bio* 8 (2009).
14. Du, L. et al. Structured Sparse Canonical Correlation Analysis for Brain Imaging Genetics: An Improved GraphNet Method. *Bioinformatics* (2016).
15. Lin, D., Calhoun, V. D. & Wang, Y. P. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Med Image Anal* 18, 891–902 (2014).
16. Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D. & Li, H. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* 14, 244–258 (2013).
17. Chen, X. & Liu, H. An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. *Statistics in Biosciences* 4, 3–26 (2012).
18. Yan, J. et al. Transcriptome-guided amyloid imaging genetic analysis via a novel structured sparse learning algorithm. *Bioinformatics* 30, i564–571 (2014).
19. Vounou, M., Nichols, T. E. & Montana, G. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage* 53, 1147–1159 (2010).
20. Vounou, M. et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer’s disease. *Neuroimage* 60, 700–716 (2012).
21. Batmanghelich, N. K., Dalca, A. V., Sabuncu, M. R. & Polina, G. Joint modeling of imaging and genetics. *Inf Process Med Imaging* 23, 766–777 (2013).
22. Krishnan, A., Williams, L. J., McIntosh, A. R. & Abd, H. Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. *Neuroimage* 56, 455–475 (2011).
23. Le Cao, K. A., Rossouw, D., Robert-Granie, C. & Besse, P. A Sparse PLS for Variable Selection when Integrating Omics Data. *Stat Appl Genet Mol Bio* 7 (2008).
Mining Outcome-relevant Brain Imaging Genetic Associations via Three-Way Sparse Canonical Correlation Analysis in Alzheimer's Disease.

Xiao Hao, Dan Zhang, Ling Shi, and Alan J. Sinclair

How to cite this article: Hao, X. et al. Mining Outcome-relevant Brain Imaging Genetic Associations via Three-Way Sparse Canonical Correlation Analysis in Alzheimer's Disease. Sci. Rep. 7, 44272; doi: 10.1038/srep44272 (2017).

Acknowledgements

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.; Mesoblast Ltd.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is conducted by the Alzheimer's Disease Neuroimaging Initiative (ADNI). A complete listing of ADNI investigators can be found at t://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. This research is supported by the National Natural Science Foundation of China (Nos 61422204, 61473149, 61501230), the NUAA Fundamental Research Funds (No. NE2013105), the Jiangsu Qinglan Project of China. At Indiana University, this work was supported by NIH R01 LM011360, U01 AG024904, RC2 AG036535, R01 AG19771, P30 AG10133, UL1 TR001108, R01 AG042437, and R01 AG046171; NSF IIS-1117335; DOD W81XWH-14-2-0151, W81XWH-13-1-0259, and W81XWH-12-2-0012; NCA
classical MRI single-subject brain. Neuroimaging 15, 273–289 (2002).

Author Contributions

X.H., D.Z. and L.S. conceived and designed the research. X.H., C.L., L.D., and J.Y. carried out the study analysis. X.Y., S.L.R. and A.J.S. collected the data from ADNI database. X.H., D.Z., L.S. and A.J.S. analyzed the results and wrote the manuscript. 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.

Additional Information

Competing Interests: The authors declare no competing financial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Consortia
Alzheimer’s Disease Neuroimaging Initiative

Michael W. Weiner5, Paul Aisen6, Ronald Petersen7,8, Clifford R. Jack Jr.8, Sara S. Mason8, Colleen S. Albers8, David Knopman8, Kris Johnson9, William Jagust9, John Q. Trojanowski10, Arthur W. Toga11, Laurel Beckett12, Robert C. Green13, Martin R. Farlow13, Ann Marie Hake13, Brandy R. Matthews13, Jared R. Brosch13, Scott Herring13, Cynthia Hunt13, Lesley M. Shaw14, Beau Ances14, John C. Morris14, Maria Carroll14, Mary L. Creech14, Erin Franklin15, Mark A. Mintun15, Stacy Schneider15, Angela Oliver16, Jeffrey Kaye15, Joseph Quinn15, Lisa Silbert15, Betty Lind15, Raina Carter15, Sara Dolen15, Lon S. Schneider15, Sonia Pawluczky15, Mauricio Beccera15, Liberty Teodoro15, Bryan M. Spann15, James Brewer16, Helen Vanderswag16, Adam Fleisher16,17, Pierre Tariot17, Anna Burke17, Nadira Trincic17, Stephanie Reeder17, Judith L. Heidebrink18, Joanne L. Lord18, Rachelle S. Doody19, Javier Villanueva-Meyer19, Munir Chowdhury19, Susan Rountree19, Mimi Dang19, Yaakov Stern20, Lawrence S. Honig20, Karen L. Bell20, Daniel Marson20, Randall Griffith21, David Clark21, David Geldmacher21, John Brockington21, Erik Roberson21, Marissa Natelson Love22, Hillel Grossman22, Effie Mitsis22, Raj C. Shah23, Leyla deToledo-Morrell23, Ranjan Duara24, Daniel Varon24, Maria T. Greig24, Peggy Roberts24, Marilyn Albert24, Chiadi Onyike25, Daniel D’Agostino25, Stephanie Kielb25, James E. Galvin26, Brittany Cerbone26, Christina A. Michel26, Dana M. Pogorelec26, Henry Rusinek26, Mony J. de Leon26, Lidia Godzik26, Susan De Santi26, P. Murari Doraiswamy27, Jeffrey R. Petrella27, Salvador Borges-Neto27, Terence Z. Wong27, Edward Coleman27, Charles D. Smith27, Greg Jicha27, Peter Hardy27, Partha Sinha27, Elizabeth Oates28, Gary Conrad28, Anton P. Posteionsson29, Bonnie S. Goldstein29, Kim Martin29, Kelly M. Makino29, M. Saleem Ismaiel29, Connie Brand29, Ruth A. Mulnard29, Gary Thai29, Catherine Mc-Adams-Ortiz29, Kyle Womack31, Dana Matthews31, Mary Quiceno31, Allan I. Levy32, James J. Lah32, Janet S. Cellar32, Jeffrey M. Burns33, Russell H. Swerdlow33, William M. Brooks33, Liana Apostolova34, Kathleen Tiusgs34, Ellen Woo34, Daniel H. Silverman34, Po H. Lu35, George Bartzokis35, Neill R. Graff-Radford35, Francine Parftt35, Tracy Kendall35, Heather Johnson35, Christopher H. van Dyck36, Richard E. Carson36, Martha G. MacAvoy36, Pradeep Varma36, Howard Chertkow37, Howard Bergman37, Chris Hosein37, Sandra Black38, Bojana Stefanovic38, Curtis Caldwell38, Ging-Yuek Robin Hsiung39, Howard Feldman39, Benita Mudge39, Michele Assaly39, Elizabeth Finger40, Stephen Pasternack40, Irina Rachisky40, Dick Trost40, Andrew Kertesz40, Charles Bernick41, Donna Munic41, Marek- Marsel Mesulam41, Kristine Lipowski42, Sandra Weintraub42, Borna Bonakdarpour42, Diana Kerwin42, Chuang-Kuo Wu42, Nancy Johnson42, Carl Sadowsky43, Teresa Villena43, Raymond Scott Turner44, Kathleen Johnson44, Brigid Reynolds44, Reisa A. Sperling45, Keith A. Johnson45, Gad Marshall45, Jerome Yesavage46, Joy L. Taylor46, Barton Lane46, Allynson Rosen46, Jared Tinklenberg46, Marwan N. Sabbagh47, Christine M. Belden47, Sandra A. Jacobson47, Sherye A. Sirrel47, Neil Kowali48, Ronald Killiany48, Andrew E. Budson48, Alexander Norbash48, Patricia Lynn Johnson48, Thomas O. Obisesan49, Saba Wolday49, Joanne Allard49, Alan Lerner50, Paula Ogrocki50, Curtis Tatsuoka50, Parianne Fatica50, Evan Fletcher50, Pauline Maillard51, John Olichney51, Charles DeCarli51, Owen Carmichael51, Smita Kittur52, Michael Borris52, T.-Y. Lee52, Rob Bartha53, Sterling Johnson53, Sanjay Asthana54, Cynthia M. Carlsson54, Steven G. Potkin55, Adrian Preda55, Dana Nguyen55, Vernice Bates56, Horacio Capote56, Michelle Rainka56, Douglas W. Scharre57, Maria Katakis57, Anahita Adeli57, Earl A. Zimmerman58, Dzintra Celmins58, Alice D. Brown58, Godfrey Pearl59, Karen Blank59, Karen Anderson59, Laura A. Flashman60, Marc Seltzer60, Mary L. Hynes60, Robert B. Santulli60, Kaycee M. Sink61, Leslie Gordiner61, Jeff D. Williamson61, Pradeep Garg61, Franklin Watkins61, Brian R. Ott62, Henry Querfurth62, Geoffrey Tremont62, Stephen Salloway63, Paul Malloy63, Stephen Correia63, Howard J. Rosen64, Bruce L. Miller64, David Perry64, Jacobo Mintzer65, Kenneth Spencer65, David Bachman65, Nunzio Pomara66, Raymond Hernando66, Antero Sarraell66, Norman Relkin67, Gloria Chaing67, Michael Lin67, Lisa Radvin67, Amanda Smith68, Balebail Ashok Raj68 & Kristin Faragher68

5Magnetic Resonance Unit at the VA Medical Center and Radiology, Medicine, Psychiatry and Neurology, University of California, San Francisco, USA. 6San Diego School of Medicine, University of California, California, USA. 7Mayo Clinic, Minnesota, USA. 8Mayo Clinic, Rochester, USA. 9University of California, Berkeley, USA. 10University of Pennsylvania, Pennsylvania, USA. 11University of Southern California, California, USA. 12University of California, Davis, California, USA. 13MPH Brigham and Women’s Hospital/Harvard Medical School, Massachusetts, USA.
Washington University St. Louis, Missouri, USA. Oregon Health and Science University, Oregon, USA. University of California—San Diego, California, USA. Banner Alzheimer’s Institute, USA. University of Michigan, Michigan, USA. Baylor College of Medicine, Houston, State of Texas, USA. Columbia University Medical Center, South Carolina, USA. University of Alabama – Birmingham, Alabama, USA. Mount Sinai School of Medicine, New York, USA. Rush University Medical Center, Rush University, Illinois, USA. Wien Center, Florida, USA. Johns Hopkins University, Maryland, USA. New York University, NY, USA. Duke University Medical Center, North Carolina, USA. University of Kentucky, Kentucky, USA. University of Rochester Medical Center, NY, USA. University of California, Irvine, California, USA. University of Texas Southwestern Medical School, Texas, USA. Emory University, Georgia, USA. University of Kansas, Medical Center, Kansas, USA. University of California, Los Angeles, California, USA. Mayo Clinic, Jacksonville, USA. Yale University School of Medicine, Connecticut, USA. McGill University, Montreal-Jewish General Hospital, Canada. Sunnybrook Health Sciences, Ontario, USA. U.B.C. Clinic for AD & Related Disorders, Canada. Cognitive Neurology - St. Joseph’s, Ontario, USA. Cleveland Clinic Lou Ruvo Center for Brain Health, Ohio, USA. Northwestern University, USA. Premiere Research Inst (Palm Beach Neurology), USA. Georgetown University Medical Center, Washington D.C, USA. Brigham and Women’s Hospital, Massachusetts, USA. Stanford University, California, USA. Banner Sun Health Research Institute, USA. Boston University, Massachusetts, USA. Howard University, Washington D.C, USA. Case Western Reserve University, Ohio, USA. University of California, Davis – Sacramento, California, USA. Neurological Care of CNY, USA. Parkwood Hospital, Pennsylvania, USA. University of Wisconsin, Wisconsin, USA. University of California, Irvine – BIC, USA. Dent Neurologic Institute, NY, USA. Ohio State University, Ohio, USA. Albany Medical College, NY, USA. Hartford Hospital, Olin Neuropsychiatry Research Center, Connecticut, USA. Dartmouth-Hitchcock Medical Center, New Hampshire, USA. Wake Forest University Health Sciences, North Carolina, USA. Rhode Island Hospital, state of Rhode Island, USA. Butler Hospital, Providence, Rhode Island, USA. University of California, San Francisco, USA. Medical University South Carolina, USA. Nathan Kline Institute, Orangeburg, New York, USA. Cornell University, Ithaca, New York, USA. USF Health Byrd Alzheimer’s Institute, University of South Florida, USA.