NeuroImage 63 (2012) 1681–1694

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression ☆

Matt Silver [a], Eva Janousova [a,b], Xue Hua [c], Paul M. Thompson [c], Giovanni Montana [a,*] and The Alzheimer's Disease Neuroimaging Initiative [1]

[a] Statistics Section, Department of Mathematics, Imperial College London, UK
[b] Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic
[c] Laboratory of Neuro Imaging, Department of Neurology, UCLA School of Medicine, Los Angeles, CA, USA

ARTICLE INFO

Article history:
Accepted 3 August 2012
Available online 15 August 2012

Keywords:
Alzheimer's disease
Imaging genetics
Atrophy
Gene pathways
Sparse regression

ABSTRACT

We present a new method for the detection of gene pathways associated with a multivariate quantitative trait, and use it to identify causal pathways associated with an imaging endophenotype characteristic of longitudinal structural change in the brains of patients with Alzheimer's disease (AD). Our method, known as pathways sparse reduced-rank regression (PsRRR), uses group lasso penalised regression to jointly model the effects of genome-wide single nucleotide polymorphisms (SNPs), grouped into functional pathways using prior knowledge of gene–gene interactions. Pathways are ranked in order of importance using a resampling strategy that exploits finite sample variability. Our application study uses whole genome scans and MR images from 99 probable AD patients and 164 healthy elderly controls in the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. 66,182 SNPs are mapped to 185 gene pathways from the KEGG pathway database. Voxel-wise imaging signatures characteristic of AD are obtained by analysing 3D patterns of structural change at 6, 12 and 24 months relative to baseline. High-ranking, AD endophenotype-associated pathways in our study include those describing insulin signalling, vascular smooth muscle contraction and focal adhesion. All of these have been previously implicated in AD biology. In a secondary analysis, we investigate SNPs and genes that may be driving pathway selection. High ranking genes include a number previously linked in gene expression studies to β-amyloid plaque formation in the AD brain (*PIK3R3*, *PIK3CG*, *PRKCA* and *PRKCB*), and to AD related changes in hippocampal gene expression (*ADCY2*, *ACTN1*, *ACACA*, and *GNAI1*). Other high ranking previously validated AD endophenotype-related genes include *CR1*, *TOMM40* and *APOE*.

Introduction

A growing list of genetic variants has now been associated with greater susceptibility to develop early and late-onset Alzheimer's disease (AD), with the *APOE*ϵ4 allele consistently identified as having the greatest effect (for an up to date list see www.alzgene.org). Recently, case–control susceptibility studies have been augmented by studies using neuroimaging phenotypes. The rationale here is that the use of heritable imaging signatures (endophenotypes) of disease may increase the power to detect causal variants, since gene effects are expected to be more penetrant at this level (Meyer-Lindenberg and Weinberger, 2006). This 'imaging-genetic' approach has been used to identify genes associated with a range of AD-associated imaging phenotypes including measures of hippocampal volume (Stein et al., 2012), cortical thickness (Burggren et al., 2008) and longitudinal, structural change (Vounou et al., 2011).

AD is a moderate to highly heritable condition, yet as with many common heritable diseases, association studies have to date identified gene variants explaining only a relatively modest amount of known AD heritability (Braskie et al., 2011). One approach to uncovering this 'missing heritability' is motivated by the observation that in many cases disease states are likely to be driven by multiple genetic variants of small to moderate effect, mediated through their interaction in molecular networks or pathways, rather than by the effects of a few, highly penetrant mutations (Schadt, 2009). Where this assumption holds, the hope is that by considering the joint effects of multiple variants acting in concert, pathways genome-wide association studies (PGWAS) will reveal aspects of a disease's genetic architecture that would otherwise be missed when considering variants individually (Fridley and Biernacka, 2011; Wang et al., 2010). Another potential benefit of the

☆ The software is available for download at the author's web page: http://www2.imperial.ac.uk/~gmontana.

* Corresponding author.
E-mail address: g.montana@imperial.ac.uk (G. Montana).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

1053-8119
http://dx.doi.org/10.1016/j.neuroimage.2012.08.002

PGWAS approach is that it can help to elucidate the mechanisms of disease by providing a biological interpretation of association results (Cantor et al., 2010). In the case of AD for example, an understanding of the underlying mechanisms by which gene mutations impact disease aetiology may play an important role in the translation of basic AD biology into therapy and patient care (Sleegers et al., 2010).

In this paper, we present the first PGWAS method that is able to accommodate a multivariate quantitative phenotype, and apply our method to a pathway analysis of the ADNI cohort, comparing genome-wide single nucleotide polymorphism (SNP) data with voxel-wise tensor-based morphometry (TBM) maps describing longitudinal structural changes that are characteristic of AD. In this study we map SNPs to pathways from the KEGG pathway database, a curated collection of functional gene pathways representing current knowledge of molecular interaction and reaction networks (http://www.genome.jp/kegg/pathway.html). Our method is however able to accommodate alternative sources of information for the grouping of SNPs and genes, for example using gene ontology (GO) terms, or information from protein interaction networks (Jensen and Bork, 2010; Wu et al., 2010).

The use of high-dimensional endophenotypes in imaging genetic studies has become increasingly commonplace, since it enables the voxel-wise mapping of genetic effects across the brain (Thompson et al., 2010). Previous work has demonstrated that a sparse reduced-rank regression (sRRR) approach that exploits the multivariate nature of the phenotype can be more powerful than a mass-univariate linear modelling approach in which each phenotype is regressed against each SNP (Vounou et al., 2010). Furthermore, multivariate, high-dimensional phenotypes have also been shown to offer an increased signal to noise ratio over low dimensional or univariate phenotypes, provided that uninformative voxels that are not characteristic of the disease under study are removed (Vounou et al., 2011). In this study we use a high-dimensional phenotype describing structural change relative to baseline over three time points in subjects with AD, and in healthy controls. From this we extract an imaging endophenotype that is highly characteristic of AD in our sample by using a stringent statistical threshold to exclude voxels that do not discriminate between AD and CN. Our main objective here is not to build a robust statistical classifier for AD, but instead to produce a quantitative phenotype having maximal sample variability between AD and CN for the subsequent gene mapping stage of our analysis.

Many existing PGWAS methods, such as GenGen (Wang et al., 2009) and ALLIGATOR (Holmans et al., 2009) rely on univariate statistics of association, whereby each SNP in the study is first independently tested for association with a univariate quantitative or dichotomous (case–control) phenotype. SNPs are assigned to pathways by mapping them to adjacent genes within a specified distance, and individual SNP or gene statistics are then combined across each pathway to give a measure of pathway significance, corrected for multiple testing. Methods must also account for the potentially biasing effects of gene and pathway size and linkage disequilibrium (LD), and this is generally done through permutation. A potential disadvantage of these methods is that each SNP is considered separately at the first step, with no account taken of SNP–SNP dependencies. In contrast, a multilocus or multivariate model that considers all SNPs simultaneously may characterise SNP effects more accurately by aiding the identification of weak signals while diminishing the importance of false ones (Hoggart et al., 2008).

In earlier work we developed a multivariate PGWAS method for identifying pathways associated with a single quantitative trait (Silver and Montana, 2012). We used a sparse regression model – the group lasso – with SNPs grouped into pathways. We demonstrated in simulation studies using real SNP and pathway data, that our method showed high sensitivity and specificity for the detection of important pathways, when compared with an alternative pathway method based on univariate SNP statistics. Our method showed the greatest relative gains in performance where marginal SNP effect sizes are small. Here we extend our previous model to accommodate the case of a multivariate neuroimaging phenotype. We do this by incorporating a group sparsity constraint on genotype coefficients in a multivariate sparse reduced-rank regression model, previously developed for the identification of single causal variants (Vounou et al., 2010). Our proposed 'pathways sparse reduced-rank regression' (PsRRR) algorithm incorporates phenotypes and genotypes in a single model, and accounts for potential biasing factors such as dependencies between voxels and SNPs using an adaptive, weight-tuning procedure.

To the best of our knowledge, few other multilocus methods for the identification of biological pathways currently exist. The GRASS method (Chen et al., 2010) and the method proposed by Zhao et al. (2011) use sparse regression techniques to measure pathway significance. These methods are currently implemented for case–control data only, and are unable to accommodate a multivariate phenotype. Each method makes different assumptions about the distribution of important SNPs and genes affecting the phenotype. GRASS assumes sparsity at the SNP level within each pathway gene, while Zhao's method assumes sparsity at the gene level. In contrast, our PsRRR method assumes sparsity only at the pathway level (although we subsequently perform SNP and gene selection as a second step in selected pathways). As such, each method is expected to perform differently, depending on the 'true' distribution of causal SNPs and genes. GRASS and Zhao's methods also use a pre-processing dimensionality reduction step on SNPs within each gene using PCA. While this has been shown to be advantageous in certain circumstances (Wang and Abbott, 2008), we elect to retain original SNP genotypes in our model, since this facilitates sparse SNP selection. A further distinguishing feature of our method is that we include all pathways together in a single regression model. By doing this we hope to gain a better measure of the relative importance of different pathways, by ensuring that they compete against each other.

The article is presented as follows. We begin in the Imaging data section with a description of the voxel-wise TBM maps used in the study, and in the Phenotype extraction section we outline how we use these maps to generate an imaging signature characteristic of structural change in AD, that is able to discriminate between AD patients and controls. In the Genotype data section we describe the genotype data used in the study, together with quality control procedures, and in the SNP to pathway mapping section we explain how this genotype data is mapped to gene pathways. The theoretical underpinnings of the PsRRR method are described in the Pathways sparse reduced-rank regression section. We explain our method for ranking AD-associated pathways, SNPs and genes using a resampling procedure in the Pathway, gene and SNP ranking section, and discuss our strategies for addressing the significant computational challenge of fitting a regression-based model with such high dimensional datasets in the Computational issues section. Pathway, SNP and gene ranking results are presented in the Results section, and we conclude with a Discussion.

Materials and methods

Imaging and genotype data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by 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 organisations, as a 5-year public–private partnership. 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 AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.