



## FGWAS: Functional genome wide association analysis<sup>☆</sup>



Chao Huang<sup>a,b</sup>, Paul Thompson<sup>c</sup>, Yalin Wang<sup>d</sup>, Yang Yu<sup>e</sup>, Jingwen Zhang<sup>a,b</sup>, Dehan Kong<sup>f</sup>,  
Rivka R. Colen<sup>g</sup>, Rebecca C. Knickmeyer<sup>h</sup>, Hongtu Zhu<sup>a,b,\*</sup>, The Alzheimer's Disease  
Neuroimaging Initiative<sup>1</sup>

<sup>a</sup> Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

<sup>b</sup> Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

<sup>c</sup> Imaging Genetics Center, Stevens Institute for Neuroimaging and Informatics, University of Southern California, Marina del Rey, CA, USA

<sup>d</sup> School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

<sup>e</sup> Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

<sup>f</sup> Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

<sup>g</sup> Department of Diagnostic Radiology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

<sup>h</sup> Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

### ARTICLE INFO

#### Keywords:

Computational complexity  
Functional genome wide association analysis  
Multivariate varying coefficient model  
Wild bootstrap

### ABSTRACT

Functional phenotypes (e.g., subcortical surface representation), which commonly arise in imaging genetic studies, have been used to detect putative genes for complexly inherited neuropsychiatric and neurodegenerative disorders. However, existing statistical methods largely ignore the functional features (e.g., functional smoothness and correlation). The aim of this paper is to develop a functional genome-wide association analysis (FGWAS) framework to efficiently carry out whole-genome analyses of functional phenotypes. FGWAS consists of three components: a multivariate varying coefficient model, a global sure independence screening procedure, and a test procedure. Compared with the standard multivariate regression model, the multivariate varying coefficient model explicitly models the functional features of functional phenotypes through the integration of smooth coefficient functions and functional principal component analysis. Statistically, compared with existing methods for genome-wide association studies (GWAS), FGWAS can substantially boost the detection power for discovering important genetic variants influencing brain structure and function. Simulation studies show that FGWAS outperforms existing GWAS methods for searching sparse signals in an extremely large search space, while controlling for the family-wise error rate. We have successfully applied FGWAS to large-scale analysis of data from the Alzheimer's Disease Neuroimaging Initiative for 708 subjects, 30,000 vertices on the left and right hippocampal surfaces, and 501,584 SNPs.

### 1. Introduction

Functional responses that frequently arise in neuroimaging studies have been widely used to characterize brain structure and function (Miller and Qiu, 2009; Smith et al., 2006; Styner et al., 2005; Fischl, 2012; Goodlett et al., 2009; Yushkevich et al., 2008). For instance, in diffusion tensor imaging, various diffusion properties (e.g., fractional

anisotropy) have been extracted along major fiber bundles to reveal white matter tract maturation and integrity (Smith et al., 2006; Goodlett et al., 2009; Yushkevich et al., 2008). Shape analysis has been widely used to characterize features of brain cortical and subcortical structures, including cortical complexity, curvature, spectral content, and other indices (Miller and Qiu, 2009; Styner et al., 2005; Fischl, 2012). Thus, they have been widely used to better understand normal brain

<sup>☆</sup> This work was partially supported by U.S. NIH grants MH086633 and MH092335, NSF grants SES-1357666 and DMS-1407655, and a grant from the Cancer Prevention Research Institute of Texas. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The readers are welcome to request reprints from Dr. Hongtu Zhu. Email: [hzh5@mdanderson.org](mailto:hzh5@mdanderson.org); Phone: 346-814-0191.

\* Corresponding author. Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.  
E-mail address: [htzhu@email.unc.edu](mailto:htzhu@email.unc.edu) (H. Zhu).

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

development and the neurological bases of neuropsychiatric and neurodegenerative diseases. Therefore, these functional responses may be effective phenotypes that facilitate the identification of causal genes and the mechanistic understanding of pathophysiological processes of neurological disorders (Zhao and Castellanos, 2016). Our primary research interest is to identify novel genetic effects on the local changes of various functional responses.

Statistically, we are interested in developing a fast and efficient statistical method to correlate functional phenotypes measured at tens of thousands of grid points ( $N_V \sim 10^3 - 10^6$ ) with tens of millions of known genetic variants ( $N_G \sim 10^7$ ), or so-called big data squared. Conventional analysis of such imaging–genetic data is based on methods for voxel-wise genome-wide association analysis studies (VGWAS). Such VGWAS methods primarily consist of three major steps: Gaussian smoothing of the functional responses across subjects, a total of  $N_G N_V$  ( $\sim 10^{10} - 10^{12}$ ) massive univariate analyses, and correction for multiple comparisons in an expanded image  $\times$  gene search space with  $N_G N_V$  elements (Hibar et al., 2011; Shen et al., 2010; Huang et al., 2015; Medland et al., 2014; Zhang et al., 2014; Thompson et al., 2014; Liu and Calhoun, 2014). These methods are not only computationally extensive, but also involve major methodological limitations when searching for novel genetic markers associated with the local changes of functional phenotypes. Specifically, running VGWAS can pose significant computational challenges, including limited computer memory, finite CPU speed, and limited CPU nodes. For instance, for VGWAS, it is computationally intensive to compute all  $N_V N_G$  ( $\sim 10^{10} - 10^{13}$ ) test statistics for all  $M$  ( $\sim 10^3 - 10^4$ ) bootstrapping replicates and to store and manage all  $N_G N_V M$  ( $\sim 10^{13} - 10^{17}$ ) test statistic images in a limited computer hard drive. Moreover, due to massive model fitting, the statistical power is usually very low after adjusting for multiple comparisons, while the spatial correlation and smoothness features in functional phenotypes are not considered, leading to difficult interpretation of the results.

We propose two important strategies to address several fundamental bottlenecks of constructing brain–genetic association maps for functional responses in large-scale imaging genetic studies. First, instead of repeatedly fitting a univariate model to each voxel and each genetic marker, we treat all image measures as a single functional response

measured at all  $N_V$  grid points and focus on testing the coefficient function of interest, which is intrinsically low rank. We develop some functional data analysis (FDA) methods to explicitly account for the three key features of the functional phenotypes: spatial smoothness, spatial correlation, and the low-dimensional representation of functional data. The key advantage of using FDA is to reduce the dimension of the functional responses from  $N_V$  to an intrinsically low dimension, denoted as  $N_{V0}$ , which is much smaller than  $N_V$ . Second, we develop a new global sure independence screening (GSIS) procedure to eliminate most of the “noisy” genetic variants and a divide-and-conquer algorithm to efficiently perform multiple comparisons. The divide-and-conquer algorithm is critically important for performing FGWAS when  $N_G$  is extremely large, such as for whole-genome sequencing.

The aim of this paper is to develop a FGWAS pipeline with several formal FDA tools as a novel extension of VGWAS for functional responses. A schematic overview of FGWAS is given in Fig. 1. Although FDA methods have been widely studied in the literature, most focus on one-dimensional curves. See Ramsay and Silverman (2005); Wang et al. (2016); Morris (2015) and references therein for a comprehensive review of FDA. Although there are a few methods for the association mapping of longitudinal phenotypes (Nicolae, 2016; Wu and Lin, 2006; Reimherr and Nicolae, 2014), little has been done on the association mapping of functional phenotypes of two or higher dimensions. Compared with existing methods in the literature, five major methodological contributions of this paper are as follows.

- We use a multivariate varying coefficient model (MVCM) as a special function-on-scalar regression model to fit the functional phenotypes with a large number of genetic variants (Zhu et al., 2011; Di et al., 2009; Zipunnikov et al., 2011; Zhu et al., 2014; Guo, 2002; Lin et al., 2014), while explicitly accounting for their three key functional features as discussed above. The use of the MVCM can project  $N_V$  imaging measures into the  $N_{V0}$ -dimensional space, leading to computational and efficiency gains on the order of  $O(N_V/N_{V0})$ .
- Under MVCM, we use a local Wald-type test statistic to detect novel genetic variants that influence brain structure and function. Moreover, such a test statistic outperforms the test statistics used in other

## Functional Genome Wide Association analysis (FGWAS)

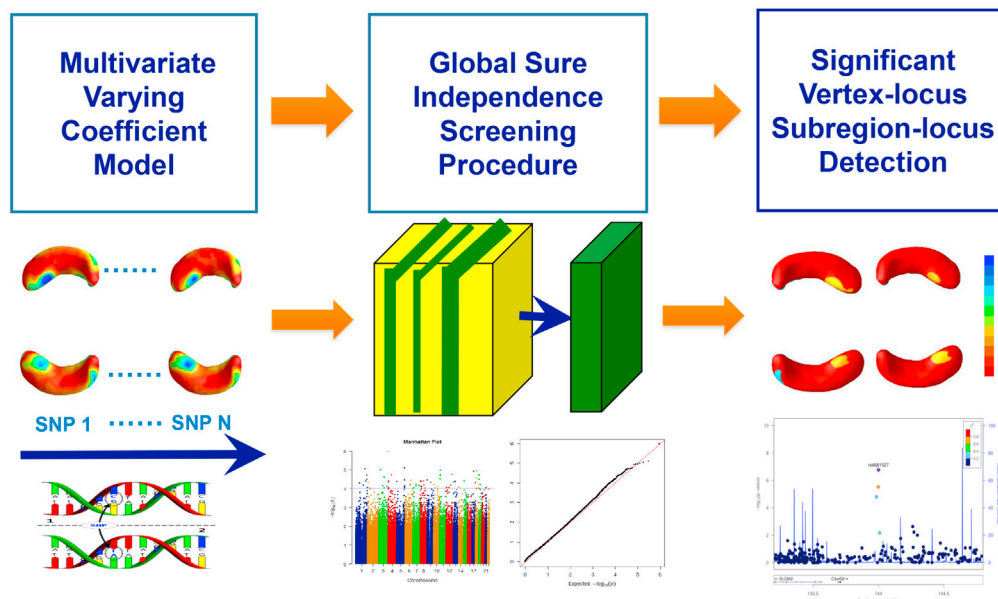


Fig. 1. Schematic overview of FGWAS.

association mapping methods in terms of statistical power (Nicolae, 2016; Wu and Lin, 2006; Reimherr and Nicolae, 2014; Huang et al., 2015).

- We introduce a GSIS procedure based on global test statistics to test the hypotheses of interest associated with functional phenotypes. The GSIS not only selects  $N_{G0}$  “important” genetic variants, but also offers the sure independence screening property (Fan and Lv, 2008) with a vanishing false selection rate. The use of GSIS can reduce the size of genetic variants from  $N_G$  to  $N_{G0}$ , leading to a computational gain on the order of  $O(N_G/N_{G0})$ .
- We develop a divide-and-conquer algorithm coupled with parallel computing to efficiently perform multiple comparisons in order to detect subregion-locus pairs, while controlling for their family-wise error (FWE) rate. Compared with VGWAS, FGWAS achieves an extensive computational gain in terms of both memory and speed.
- The package for FGWAS, along with its documentation, are freely accessible from the website “<http://odin.mdacc.tmc.edu/bigs2/>” and “<https://github.com/BIG-S2>”. To make it user-friendly, we developed a graphical user interface to package the code, which is also freely downloadable from the same website. Our FGWAS package can handle three types of functional phenotypes: curves, surfaces, and volumes in  $\mathbb{R}^3$ . To facilitate its application to real data, we use three computing languages, Rcpp, Matlab, and Python, to develop the corresponding versions.

## 2. Method

Suppose that we observe functional responses, clinical covariates and genetic markers for  $n$  unrelated subjects. Without loss of generality, we focus on a compact set, denoted as  $\mathcal{D} \subset \mathbb{R}^3$ , (e.g., cortical and subcortical regions), which is general enough to cover curves, surfaces, and volumes in  $\mathbb{R}^3$ . Let  $\mathbf{d}$  be a grid point in  $\mathcal{D}$ . It is assumed that there are  $N_V$  common grid points  $\mathbf{d}_1, \dots, \mathbf{d}_{N_V}$  across all subjects. Let  $\mathbf{g}$  be a locus in the set of  $N_G$  genetic markers, denoted as  $\mathcal{G} = \{g_1, \dots, g_{N_G}\}$ . Specifically, for the  $i$ -th subject, we observe a  $J \times 1$  vector of imaging measurements at  $\mathbf{d}$ , denoted as  $\mathbf{y}_i(\mathbf{d}) = (y_{i1}(\mathbf{d}), \dots, y_{iJ}(\mathbf{d}))^T$ , a  $p_c \times 1$  vector of covariates (e.g., age and gender), denoted as  $\mathbf{x}_i$  with its first component being 1, and a  $p_g \times 1$  vector associated with the genetic marker at the locus  $\mathbf{g}$  and/or  $\mathbf{x}_i$ , denoted as  $\mathbf{z}_i(\mathbf{g}) = (z_{i,1}(\mathbf{g}), \dots, z_{i,p_g}(\mathbf{g}))^T$ .

### 2.1. FGWAS

We developed the FGWAS pipeline to efficiently carry out the association mapping of functional phenotypes. A schematic overview of FGWAS is given in Fig. 1. Our FGWAS consists of three major components:

- (I) a multivariate varying coefficient model (MVCM);
- (II) a global sure independence screening procedure;
- (III) a test procedure based on the global and local test statistics.

We elaborate on these components below.

### 2.2. FGWAS (I): multivariate varying coefficient model

For the genetic marker at locus  $\mathbf{g}$ , the MVCM is defined as

$$y_{ij}(\mathbf{d}) = \mathbf{x}_i^T \boldsymbol{\beta}_j^{(c)}(\mathbf{d}) + \mathbf{z}_i(\mathbf{g})^T \boldsymbol{\beta}_j^{(g)}(\mathbf{d}) + \eta_{ij}^{(g)}(\mathbf{d}) + \varepsilon_{ij}(\mathbf{d}), \quad (1)$$

where  $\boldsymbol{\beta}_j^{(c)}(\mathbf{d})$  is a  $p_c \times 1$  vector of non-genetic fixed effects,  $\boldsymbol{\beta}_j^{(g)}(\mathbf{d})$  is a  $p_g \times 1$  vector of fixed genetic effects,  $\boldsymbol{\eta}_i^{(g)}(\mathbf{d}) = (\eta_{i1}^{(g)}(\mathbf{d}), \dots, \eta_{iJ}^{(g)}(\mathbf{d}))^T$  characterizes both subject-specific and location-specific variability, and  $\varepsilon_i(\mathbf{d}) = (\varepsilon_{i1}(\mathbf{d}), \dots, \varepsilon_{iJ}(\mathbf{d}))^T$  are measurement errors. It is also assumed that  $\eta_i^{(g)}(\mathbf{d})$  and  $\varepsilon_i(\mathbf{d})$  are mutually independent and identical copies of

$SP(0, \Sigma_{\eta}^{(g)})$  and  $SP(0, \Sigma_{\varepsilon})$ , respectively, where  $SP(\boldsymbol{\mu}, \Sigma)$  denotes a stochastic process vector with mean function  $\boldsymbol{\mu}(\mathbf{d})$  and covariance function  $\Sigma(\mathbf{d}, \mathbf{d}')$ . Moreover,  $\Sigma_{\varepsilon}(\mathbf{d}, \mathbf{d}')$  takes the form of  $\Omega_{\varepsilon}(\mathbf{d})1(\mathbf{d} = \mathbf{d}')$ , where  $\Omega_{\varepsilon}(\mathbf{d})$  is a nonnegative function of  $\mathbf{d}$  and  $1(\cdot)$  is an indicator function of an event.

Compared with the standard linear regression model, MVCM explicitly accounts for spatial smoothness, spatial correlation, and the low-dimensional representation of functional responses (Zhu et al., 2011, 2014; Zipunnikov et al., 2011; Guo, 2002). The functional responses in neuroimaging studies can usually be regarded as a noisy version of a smooth function of  $\mathbf{d} \in \mathcal{D}$ . For spatial smoothness, it is reasonable to assume that  $\boldsymbol{\beta}_j^{(c)}(\cdot)$  and  $\boldsymbol{\beta}_j^{(g)}(\cdot)$  in MVCM may inherit the smooth feature from functional phenotypes and can be represented as a linear combination of a small number of basis functions, such as B-spline. For spatial correlation, it is assumed that  $\eta_{ij}^{(g)}(\cdot)$ 's are smooth functions and allow for the Karhunen-Loeve expansion as follows:

$$\eta_{ij}^{(g)}(\mathbf{d}) = \sum_{l=1}^{\infty} \xi_{i,j,l} \psi_{jl}(\mathbf{d}) \quad (2)$$

where  $\psi_{jl}(\cdot)$ 's are eigenfunctions of  $\sum_{ij}^{(g)}(\cdot, \cdot)$  such that  $\sum_{ij}^{(g)}(\mathbf{d}, \mathbf{d}') = \sum_{l=1}^{\infty} \lambda_{jl} \psi_{jl}(\mathbf{d}) \psi_{jl}(\mathbf{d}')$  with  $\sum_{l=1}^{\infty} \lambda_{jl} < \infty$  captures the spatial correlation of  $\eta_{ij}^{(g)}(\mathbf{d})$ . Moreover,  $\xi_{i,j,l}$  is the  $(j, l)$ -th functional principal component score of the  $i$ -th subject such that  $E(\xi_{i,j,l}) = 0$  and  $\text{Var}(\xi_{i,j,l}) = \lambda_{jl}$ . Thus, we can accurately approximate  $\eta_{ij}^{(g)}(\mathbf{d})$  by a small number of eigenfunctions such that  $\eta_{ij}^{(g)}(\mathbf{d}) \approx \sum_{l=1}^L \xi_{i,j,l} \psi_{jl}(\mathbf{d})$ , where  $L$  is a positive integer. In FGWAS, following the standard approach in functional data analysis, we use the fraction of variance explained by the first few top PC components to determine  $L$ . In both simulations and ADNI data analysis, the first three eigenfunctions are included in the model, since they explain more than 90% of variance. Finally, we can obtain a low-dimensional representation of  $y_{ij}(\mathbf{d})$  as follows:

$$y_{ij}(\mathbf{d}) \approx \mathbf{x}_i^T \boldsymbol{\beta}_j^{(c)}(\mathbf{d}) + \mathbf{z}_i(\mathbf{g})^T \boldsymbol{\beta}_j^{(g)}(\mathbf{d}) + \sum_{l=1}^L \xi_{i,j,l} \psi_{jl}(\mathbf{d}). \quad (3)$$

For each genetic marker, representation (3) ensures that the intrinsic dimension of  $y_{ij}(\cdot)$  is much lower than  $N_V$ . Moreover, since it is expected that  $\boldsymbol{\beta}_j^{(g)}(\mathbf{d}) = 0$  holds for most loci, the true search space should be much smaller than  $N_V \times N_G$ .

Under model (1), we start with a hypothesis testing problem on  $\boldsymbol{\beta}_j^{(g)}(\mathbf{d}), j = 1, \dots, J$ ,

$$H_0 : \boldsymbol{\beta}_j^{(g)}(\mathbf{d}) = 0 \text{ v.s. } H_1 : \boldsymbol{\beta}_j^{(g)}(\mathbf{d}) \neq 0 \text{ for each } (g, \mathbf{d}), \quad (4)$$

where  $\boldsymbol{\beta}_j^{(g)}(\mathbf{d}) = \text{vec}([\boldsymbol{\beta}_1^{(g)}(\mathbf{d}), \dots, \boldsymbol{\beta}_J^{(g)}(\mathbf{d})])$ , and  $\text{vec}(\cdot)$  is the vectorization of a matrix.

As an example, in our analysis of data from the Alzheimers Disease Neuroimaging Initiative (ADNI), we are interested in detecting novel genetic markers that influence the radial distance and determinant of the Jacobian matrix of both the left and right hippocampal surfaces. We consider MVCM (1) on the hippocampal surfaces with  $(y_{i1}(\mathbf{d}), y_{i2}(\mathbf{d}))^T = (\text{radial distance, determinant})^T$ ,  $\mathbf{x}_i = (\text{intercept, gender, age, apolipoprotein E (APOE) gene } \varepsilon 4, \text{ the top 5 principal component scores of all single-nucleotide polymorphisms [SNPs]})^T$ , and  $\mathbf{z}_i(\mathbf{g}) = (\text{SNP value})$ .

We introduce a local Wald-type test statistic  $T_n(g, \mathbf{d})$  as follows:

$$T_n(g, \mathbf{d}) = \mathbf{r}^{(g)}(\mathbf{d})^T \left[ \left\{ \widehat{\Sigma}_{\eta}^{(g)}(\mathbf{d}, \mathbf{d}) \right\}^{-1} \otimes \left\{ \mathbf{Z}_X^T(\mathbf{g}) \mathbf{Z}_X(\mathbf{g}) \right\}^{-1} \right] \mathbf{r}^{(g)}(\mathbf{d}), \quad (5)$$

where

$$\mathbf{r}^{(g)}(\mathbf{d}) = \widehat{\boldsymbol{\beta}}^{(g)}(\mathbf{d}) - \text{Bias}(\widehat{\boldsymbol{\beta}}^{(g)}(\mathbf{d})), \quad \mathbf{Z}_X(\mathbf{g}) = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T) \mathbf{Z}(\mathbf{g})^T, \quad \mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T,$$

$\mathbf{Z}(g) = (\mathbf{z}_1(g), \dots, \mathbf{z}_n(g))$ , and  $\otimes$  denotes the Kronecker product. The  $\hat{\beta}^{(g)}(\mathbf{d})$  and  $\hat{\Sigma}_\eta^{(g)}$  are estimates of the corresponding parameters, while  $\text{Bias}(\hat{\beta}^{(g)}(\mathbf{d}))$  is the bias term of  $\hat{\beta}^{(g)}(\mathbf{d})$ . Moreover, under the null hypothesis  $H_0$ , the limiting distribution of  $T_n(g, \mathbf{d})$  can be approximated by a weighted  $\chi^2$  distribution (Zhang and Chen, 2007).

To estimate all unknown parameters in model (1), we employ a weighted least squares (WLS) method based on the multivariate local polynomial kernel smoothing technique (Fan and Gijbels, 1996; Zhang and Chen, 2007). Let  $K(\cdot)$  be a kernel function,  $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ , and  $H$  be a bandwidth matrix with a simple diagonal form. We also denote that  $K_{H,m}(\mathbf{d}) = |H|^{-1} K(H^{-1}(\mathbf{d}_m - \mathbf{d}))$  and  $\mathbf{w}_H(\mathbf{d}_m - \mathbf{d}) = (1, (\mathbf{d}_m - \mathbf{d})^T H^{-1})^T$ . For each  $j$  and a fixed bandwidth matrix  $H_\beta$ , the WLS estimator of  $\beta_j^{(g)}(\mathbf{d})$  is given by

$$\hat{\beta}_j^{(g)}(\mathbf{d}) = \{\mathbf{Z}_X^T(g) \mathbf{Z}_X(g)\}^{-1} \mathbf{Z}_X^T(g) \sum_{m=1}^{N_V} a_m(H_\beta, \mathbf{d}) \mathbf{y}_{\cdot,j}(\mathbf{d}_m), \quad (6)$$

where

$a_m(H_\beta, \mathbf{d}) = \mathbf{e}^T [\sum_{m=1}^{N_V} K_{H_\beta,m}(\mathbf{d}) \{\mathbf{w}_{H_\beta}(\mathbf{d}_m - \mathbf{d})\}^{\otimes 2}]^{-1} K_{H_\beta,m}(\mathbf{d}) \mathbf{w}_{H_\beta}(\mathbf{d}_m - \mathbf{d})$ ,  $\mathbf{e} = (1, 0, 0, 0)^T$ , and  $\mathbf{y}_{\cdot,j}(\mathbf{d}) = (y_{1j}(\mathbf{d}), \dots, y_{nj}(\mathbf{d}))^T$ . Based on (6), for a fixed bandwidth matrix  $H_\eta$ , the WLS estimate of  $\eta_{ij}^{(g)}(\mathbf{d})$  is given by

$$\hat{\eta}_{ij}^{(g)}(\mathbf{d}) = \sum_{m=1}^{N_V} a_m(H_\eta, \mathbf{d}) \left\{ y_{ij}(\mathbf{d}_m) - \mathbf{x}_i^T \hat{\beta}_j^{(c)}(\mathbf{d}_m) - \mathbf{z}_i(g)^T \hat{\beta}_j^{(g)}(\mathbf{d}_m) \right\}. \quad (7)$$

Finally, we can estimate  $\sum_\eta^{(g)}(\mathbf{d}, \mathbf{d}')$  by using the sample covariance function of  $\hat{\eta}_i^{(g)}(\mathbf{d})$ , denoted as  $\hat{\Sigma}_\eta^{(g)}(\mathbf{d}, \mathbf{d}')$ .

To select the optimal bandwidth in  $\hat{\beta}_j^{(g)}(\mathbf{d})$  (or  $\hat{\eta}_{ij}^{(g)}(\mathbf{d})$ ), we use the generalized cross-validation score method (Zhang and Chen, 2007; Zhu et al., 2012). We standardize all covariates to have mean zero and standard deviation one; thus, we may choose a common bandwidth for all covariates. Moreover, following the arguments of Fan and Zhang (1999), a small bandwidth leads to a small value of  $\text{Bias}(\hat{\beta}^{(g)}(\mathbf{d}))$ , which can be dropped from the test statistics hereafter.

Four big-data challenges stem from the calculations of  $T_n(g, \mathbf{d})$ .

- (B1) Calculating  $\hat{\Sigma}_\eta^{(g)}(\mathbf{d}, \mathbf{d})$  across all loci and grid points ( $O(N_V N_G)$ ) is computationally intensive.
- (B2) Bandwidth selection in  $T_n(g, \mathbf{d})$  across all loci ( $O(N_G)$ ) can also be computationally intensive.
- (B3) Substantial computer resources are required to store all  $N_V \times N_G$  test statistics  $T_n(g, \mathbf{d})$ .
- (B4) Determining how to speed up the calculation of  $T_n(g, \mathbf{d})$ .

To solve the computational bottlenecks in (B1)–(B4), we propose the following solutions:

- (S1) Calculate  $\hat{\Sigma}_\eta^{(g)}(\mathbf{d}, \mathbf{d})$  under the null hypothesis  $H_0$  for all loci.
- (S2) Divide all loci into multiple groups based on their minor allele frequencies (MAFs), and select a common optimal bandwidth for each group.
- (S3) Develop a GSIS procedure to eliminate many ‘noisy’ loci based on a global Wald-type statistic.
- (S4) Set up a parallel computing strategy so that processing large-scale genetic data can be technically feasible with limited computer resources.

The key idea of (S1) is to calculate  $\hat{\Sigma}_\eta^{(g)}(\mathbf{d}, \mathbf{d})$  under the null hypothesis in (4), since it is expected that the null hypothesis  $H_0$  holds for most loci. Similar to the estimation procedure in (7), the estimate of  $\sum_\eta^{(g)}(\mathbf{d})$  under  $H_0$  is given by

$$\hat{\Sigma}_\eta^{(g)}(\mathbf{d}, \mathbf{d}) = n^{-1} \sum_{i=1}^n \{\hat{\eta}_i(\mathbf{d}) - \bar{\eta}(\mathbf{d})\}^{\otimes 2}, \quad (8)$$

where  $\hat{\eta}_i(\mathbf{d}) = \sum_{m=1}^{N_V} a_m(H_\eta, \mathbf{d}) \{\mathbf{y}_i^T(\mathbf{d}_m) - \mathbf{x}_i^T \hat{\beta}^{(c)}(\mathbf{d}_m)\}^T$  and  $\bar{\eta}(\mathbf{d}) = n^{-1} \sum_{i=1}^n \hat{\eta}_i(\mathbf{d})$ . Since  $\hat{\Sigma}_\eta^{(g)}(\mathbf{d}, \mathbf{d})$  is invariant across all loci, we only need to calculate  $\hat{\Sigma}_\eta^{(g)}(\mathbf{d}, \mathbf{d})$  at each vertex  $\mathbf{d}$  and denote it as  $\hat{\Sigma}_\eta(\mathbf{d})$  from here on. Moreover, since  $\hat{\Sigma}_\eta(\mathbf{d})$  in (8) is only related to the non-genetic covariates, the optimal bandwidth matrix  $H_\eta$  is calculated once at most for all loci. Thus, the total complexity of computing all  $\{\hat{\Sigma}_\eta(\mathbf{d})\}$  is at least  $\min(N_V, N_G)$  times faster than calculating  $\{\hat{\Sigma}_\eta^{(g)}(\mathbf{d}, \mathbf{d})\}$ .

The key idea of (S2) is to select the common optimal bandwidth matrix  $H_\beta$  in  $\hat{\beta}^{(g)}(\mathbf{d})$  according to the MAFs. Specifically, we divide all the genetic markers into 6 groups according to their MAFs, including  $\text{MAF} \in (0.075, 0.15]$ ,  $\text{MAF} \in (0.15, 0.25]$ ,  $\text{MAF} \in (0.25, 0.35]$ ,  $\text{MAF} \in (0.35, 0.45]$ , and  $\text{MAF} \in (0.45, 0.50]$ . For each MAF group, we randomly select  $k_H$  SNPs (e.g.,  $k_H = 10$ ), and calculate the optimal bandwidth in  $\hat{\beta}^{(g)}(\mathbf{d})$  when each genetic marker is included in model (1). Consequently, the optimal bandwidth in each group is determined as the average of all the  $k_H$  bandwidths. Moreover, the number of MAF groups can be larger than 6. We elaborate on (S3) and (S4) in the next subsection.

### 2.3. FGWAS (II): A global sure independence screening procedure

The key idea of the GSIS procedure in (S3) is to detect potentially causal genetic markers by using a dimension reduction method and an approximation method (Huang et al., 2015). Specifically, since only a small number of causal genetic markers are expected to contribute to the imaging phenotypic measures, we consider a global Wald-type statistic to eliminate many loci with weak or even no genetic effects. Let  $w(\mathbf{d})$  be a prior distribution of  $\mathbf{d}$  in  $\mathcal{D}$ . The global Wald-type statistic at locus  $g$ , denoted as  $T(g)$ , is an integral of  $T_n(g, \mathbf{d})w(\mathbf{d})$  with respect to  $\mathbf{d} \in \mathcal{D}$ ; that is,  $T(g) = \int_{\mathcal{D}} T_n(g, \mathbf{d})w(\mathbf{d})dL(\mathbf{d})$ , where  $L(\mathbf{d})$  is the Lebesgue measure.

Selecting different  $w(\mathbf{d})$  allows us to introduce the prior information of specific regions of interest (ROIs). If there is no such prior information, then a uniform prior can be used. In this case, except for a constant scalar,  $T(g)$  can be approximated by

$$T_n(g) = \frac{1}{N_V} \text{tr} \left( \left[ \sum_{m=1}^{N_V} \mathbf{Y}_w(\mathbf{d}_m) \{\hat{\Sigma}_\eta(\mathbf{d}_m)\}^{-1} \mathbf{Y}_w^T(\mathbf{d}_m) \right] \otimes \mathbf{Q}_{z|x}^{-1} \text{vec}(\mathbf{Z}_X^T(g))^{\otimes 2} \right), \quad (9)$$

where  $\mathbf{Y}_w(\mathbf{d}) = \sum_{m=1}^{N_V} a_m(H_\beta, \mathbf{d}) [\mathbf{y}_{1\cdot}(\mathbf{d}_m), \dots, \mathbf{y}_{n\cdot}(\mathbf{d}_m)]^T$ , and  $\mathbf{Q}_{z|x}(g) = \mathbf{Z}_X^T(g) \mathbf{Z}_X(g)$ . At a specific locus  $g$ , if the area of the true genetic effect region, denoted by  $\mathcal{D}^*(g)$ , is relatively large and its corresponding measurements are moderate, then the value of  $T_n(g)$  should be relatively large. Thus, if the value of  $T_n(g)$  is large, then the locus  $g$  is more likely to be a causal locus.

Our GSIS procedure consists of the following steps:

- Step (II.1). Calculate  $\mathbf{X}^T \mathbf{X}$  and  $\mathbf{P}_X$  with the computational complexity of  $O(np^2)$ .
- Step (II.2). Calculate  $\sum_{m=1}^{N_V} \mathbf{Y}_w(\mathbf{d}_m) \{\hat{\Sigma}_\eta(\mathbf{d}_m)\}^{-1} \mathbf{Y}_w^T(\mathbf{d}_m)$  with the computational complexity of  $O(N_V^2 n^2)$ .
- Step (II.3). For the locus  $g$ , calculate  $T_n(g)$  with the computational complexity of  $O((p_c + p_g)^2 n^2)$ .
- Step (II.4). Repeat Step (II.3) for all loci and calculate the  $p$ -value of  $T_n(g)$  using an approximation method (Zhang and Chen, 2007; Zhu



et al., 2012; Huang et al., 2015). Specifically,  $T_n(g)$  can be approximated by a  $\chi^2$ -type random variable  $\alpha_1\chi^2(\alpha_2) + \alpha_3$ , where  $\alpha_1$ ,  $\alpha_2$ , and  $\alpha_3$  are respectively given by

$$\alpha_1 = \frac{\kappa_3(T)}{4\kappa_2(T)}, \alpha_2 = \frac{8\kappa_3^2(T)}{\kappa_3^2(T)}, \text{ and } \alpha_3 = \kappa_1(T) - \frac{2\kappa_2^2(T)}{\kappa_3(T)}, \quad (10)$$

where  $\kappa_k(T)$ ,  $k = 1, 2, 3$ , are respectively the first three sample cumulants of  $T_n(g)$ . Finally, the  $p$ -value of  $T_n(g)$  can be approximated by using  $P(\chi^2(\alpha_2) \geq [T_n(g) - \alpha_3]/\alpha_1)$ .

- Step (II.5). Sort the  $-\log_{10}(p)$ -values of all  $T_n(g)$ s and select the top  $N_0$  loci (e.g.,  $N_0 = 1,000$ ), denoted by  $\mathcal{S}_0^* = \{g_1^*, \dots, g_{N_0}^*\}$ . From here on, we call  $\mathcal{S}_0^*$  a *candidate significant locus set*. Usually, we choose a relatively large  $N_0$  so as to guarantee that all true positive loci are contained in  $\mathcal{S}_0^*$  with high probability.

The computational complexity of GSIS is primarily associated with the number of loci,  $N_G$ . If  $N_G$  is of a super large scale (e.g.,  $O(10^8)$ ), the GSIS procedure can be very time consuming or even fail on a single computer core with limited computer memory. To address this issue, we propose a divide-and-conquer algorithm along with a parallel computing strategy, since the calculation of  $T_n(g)$  can be done independently. First, we divide the whole genetic data set into  $K_G$  groups (e.g., each chromosome as a group). Then, we perform Steps (II.1)-(II.3) independently for each group of genetic markers. Finally, we combine the  $T_n(g)$ 's across all groups and approximate their corresponding  $p$ -values based on the method used in Step (II.4). Subsequently, we determine the candidate significant locus set  $\mathcal{S}_0^*$ . More details on the parallel computing strategy are provided in the next subsection.

#### 2.4. FGWAS (III): A test procedure

The objectives of the test procedure are

- (O.1) to detect the genetic markers that are significantly associated with the functional phenotype as a whole; and
- (O.2) to detect the subregion(s) (or compact set(s)) of the functional phenotype that are significantly associated with some genetic marker(s).

Note that it is important to detect significant voxel-locus pairs for VGWAS, but such detection is less meaningful for the functional responses, which are intrinsically smooth functions. Moreover, the existing GWAS methods for imaging phenotypes focus on (O.1), whereas we are particularly interested in (O.2).

To achieve (O.1), we calculate a maximum statistic of all  $T_n(g)$ 's across all loci as follows:

$$T_{n,\mathcal{G}} = \max_{g \in \mathcal{G}} T_n(g). \quad (11)$$

The maximum statistic  $T_{n,\mathcal{G}}$  plays a crucial role in controlling the FWE rate.

To achieve (O.2), we resort to cluster size inference, which plays an important role in assessing the significance of each subregion that consists of interconnected grid points for which the  $p$ -values are greater than a given threshold, say  $\alpha_l = 0.005$  or  $0.001$  (Smith and Nichols, 2009; Ge et al., 2012). For functional phenotypes, we prefer to replace the cluster by the subregion from here on. For the locus  $g$ , let  $p(g, \mathbf{d})$  be the  $p$ -value of  $T_n(g, \mathbf{d})$  at the grid point  $\mathbf{d}$  and let  $A(g, \alpha_l)$  be the largest subregion at a given threshold  $\alpha_l$  based on the map of  $\{p(g, \mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$ . To detect significant subregion-locus pairs, we consider a maximum subregion statistic, i.e.,

$$A(\mathcal{S}, \alpha_l) = \max_{g \in \mathcal{S}} A(g, \alpha_l), \quad (12)$$

which, in practice, can be approximated by a local maximum subregion statistic  $A(\mathcal{S}_0^*, \alpha_l)$  in terms of both size and distribution (Huang et al., 2015).

We use the wild bootstrap method to approximate the null distribution of  $T_{n,\mathcal{G}}$  and that of  $A(\mathcal{S}_0^*, \alpha_l)$  under the assumption that the null hypothesis  $H_0$  in (4) holds for all  $(g, \mathbf{d}) \in \mathcal{G} \times \mathcal{D}$ . We propose an efficient wild bootstrap procedure to simultaneously detect significant loci and subregion-locus pairs as follows:

- Step (III.1). Calculate  $T_n(g^*, \mathbf{d})$  for each pair  $(g^*, \mathbf{d}) \in \mathcal{S}_0^* \times \mathcal{D}$  as

$$\text{tr} \left( \left[ \mathbf{Y}_w(\mathbf{d}_m) \{ \hat{\Sigma}_\eta(\mathbf{d}_m) \}^{-1} \mathbf{Y}_w^T(\mathbf{d}_m) \right] \otimes \mathbf{Q}_{\text{GLS}}^{-1}(g^*) \text{vec}(\mathbf{Z}_X^T(g^*))^{\otimes 2} \right) \quad (13)$$

- Step (III.2). Calculate the uncorrected  $p$ -values of  $T_n(g^*, \mathbf{d})$  across all  $(g^*, \mathbf{d}) \in \mathcal{S}_0^* \times \mathcal{D}$  based on the  $F$  distribution.
- Step (III.3). Calculate  $A(g^*, \alpha_l)$  based on the  $p$ -values of  $\{T_n(g^*, \mathbf{d})\}$  obtained in Step (III.2).
- Step (III.4). Apply the wild bootstrap method to the set  $\mathcal{S}_0^*$ .

- Step (III.4.1). Fit model (1) under the null hypothesis  $H_0$ , which yields  $\hat{\beta}^{(c)*}(\mathbf{d})$ ,  $\hat{\eta}_i^*(\mathbf{d})$  and  $\hat{\epsilon}_i^*(\mathbf{d})$  for all  $i$  and  $\mathbf{d}$ .

- Step (III.4.2). Generate a random sample  $\nu_i^b$  and  $\nu_i^b(\mathbf{d}_m)$  from a  $N(0, 1)$  generator for  $i = 1, \dots, n$  and  $m = 1, \dots, N_v$ .  $B$  bootstrap samples are constructed as

$$\mathbf{y}_i^{(b)}(\mathbf{d}_m) = \mathbf{x}_i^T \hat{\beta}^{(c)*}(\mathbf{d}_m) + \nu_i^b \hat{\eta}_i^*(\mathbf{d}_m) + \nu_i^b(\mathbf{d}_m) \hat{\epsilon}_i^*(\mathbf{d}_m), b = 1, \dots, B$$

for all  $i$  and  $\mathbf{d}_m \in \mathcal{D}$ .

- Step (III.4.3). For all  $g \in \mathcal{G}$ , calculate the global Wald-type statistic  $T_n^{(b)}(g)$  based on the bootstrap samples.
- Step (III.4.4). Sort all  $\{T_n^{(b)}(g)\}$  according to their magnitudes and select the top  $N_0$  loci to form  $\mathcal{S}_0^{*(b)}$ .
- Step (III.4.5). Calculate  $T_{n,\mathcal{G}}^{(b)} = \max_{g \in \mathcal{G}} \{T_n^{(b)}(g)\}$ .
- Step (III.4.6). Calculate  $A^{(b)}(\mathcal{S}_0^{*(b)}, \alpha_l)$  based on  $\{T_n^{(b)}(g, \mathbf{d}), (g, \mathbf{d}) \in \mathcal{S}_0^{*(b)} \times \mathcal{D}\}$ . For computational efficiency, we suggest directly comparing  $T_n^{(b)}(g, \mathbf{d})$  with the  $100(1 - \alpha_l)$ th percentile of the  $F$  distribution in order to determine significant subregions at each locus  $g$ .

- Step (III.5). Calculate the FWE corrected  $p$ -values of  $T_n(g)$  based on the empirical distribution of  $\{T_{n,\mathcal{G}}^{(b)}\}_{b=1, \dots, B}$ . Since  $N_G$  is much larger than the sample size, choose a significance level, say  $\alpha = 0.5$ .

- Step (III.6). For each locus  $g \in \mathcal{S}_0^*$ , calculate all possible subregions and their associated FWE corrected  $p$ -values based on the empirical distribution of  $\{A^{(b)}(\mathcal{S}_0^{*(b)}, \alpha_l)\}_{b=1, \dots, B}$ .

Similar to the GSIS procedure, the computational issue still exists in the test procedure for large  $N_G$ . We also use the divide-and-conquer algorithm here. Specifically, after generating bootstrap samples in Steps (III.4.1)-(III.4.2), we divide the whole genetic data set into  $K_G$  disjoint groups such that  $\mathcal{G} = \cup_{k=1}^{K_G} \mathcal{G}_k$  and  $\mathcal{G}_k \cap \mathcal{G}_{k'} = \emptyset$  for  $k \neq k'$ . Then, Steps (III.4.3)-(III.4.6) are independently performed for bootstrap samples on each group of genetic markers. For each group, we can obtain the relevant  $\{T_{n,\mathcal{G}_k}^{(b)}, A^{(b)}(\mathcal{S}_{0,k}^{*(b)}, \alpha_l)\}_{b=1}^B$  for  $k = 1, \dots, K_G$ . Then the maximum statistics  $T_{n,\mathcal{G}}^{(b)}$  and  $A(\mathcal{S}_0^{*(b)}, \alpha_l)$  across all groups are calculated as follows:

$$T_{n,\mathcal{G}}^{(b)} = \max_{1 \leq k \leq K_G} \{T_{n,\mathcal{G}_k}^{(b)}\}, b = 1, \dots, B, \quad (14)$$

$$A(\mathcal{S}_0^{*(b)}, \alpha_l) = \max_{1 \leq k \leq K_G} \{A^{(b)}(\mathcal{S}_{0,k}^{*(b)}, \alpha_l)\}, b = 1, \dots, B, \quad (15)$$

which lead to the empirical distributions of  $\{T_{n,\mathcal{S}}^{(b)}\}_{b=1}^B$  and  $\{A(\mathcal{S}_0^{*(b)}, \alpha_t)\}_{b=1}^B$  under  $H_0$ . Consequently, the corresponding corrected  $p$ -values are derived in Steps (III.5)–(III.6). In addition, this parallel computing strategy can be conducted on the bootstrap sampling level, which means calculations on different bootstrap samples in Step (III.4) can be carried out at the same time on different cores.

The divide-and-conquer algorithm coupled with parallel computing can achieve computational gain in terms of both memory and speed, while having the same statistical power as the standard method for significant locus detection. Specifically, in GSIS, since the computation of global test statistic across all loci is independent of each other, the divide-and-conquer algorithm in Steps (II.1)–(II.3) does not change the test statistics. The same comment is also valid for  $T_{n,\mathcal{S}}^{(b)}$  and  $A(\mathcal{S}_0^{*(b)}, \alpha_t)$  in the test procedure.

### 3. Simulation studies

In this section, we use Monte Carlo simulations to evaluate the finite-sample performance of FGWAS. The hypothesis testing problem we focus on is to test the null hypothesis of no association for all the functional phenotypes at each locus. All computations for these numerical examples were done in Matlab on a Dell C6100 server. The computation for FGWAS is efficient even for large-scale imaging genetic data, as shown in the real data analysis.

We simulated imaging surface data at  $N_V = 15,000$  vertices on the right hippocampus obtained from the publicly accessible ADNI data. More information on the ADNI data used in the current study is given in the next section. We only considered the additive genetic effect of SNPs on the right hippocampal surface data. The  $y_{ij}(\mathbf{d})$ s were generated from model (1) given by

$$y_{ij}(\mathbf{d}) = \mathbf{x}_i^T \beta_j^{(c)}(\mathbf{d}) + \sum_{g=1}^{N_G} z_i(g) \beta_j^{(g)}(\mathbf{d}) + \eta_{ij}(\mathbf{d}) + \varepsilon_{ij}(\mathbf{d}) \quad (16)$$

for  $i = 1, \dots, n$  and  $j = 1, 2$ , where  $\varepsilon_i(\mathbf{d}) \sim N(0, \Omega = \text{diag}(\sigma_1^2, \sigma_2^2))$ ,  $z_i(g)$  were simulated genetic data as described below, and  $\mathbf{x}_i = (1, x_{i1}, \dots, x_{i8})^T$  were generated from  $U(0, 1)$  for continuous variables or from the Bernoulli distribution with a success probability of 0.5 for discrete variables.

The true values of  $\beta_j^{(c)}(\mathbf{d})$  and  $\eta_{ij}(\mathbf{d})$  were set to the estimates  $\hat{\beta}_j^{(c)}(\mathbf{d})$  and  $\hat{\eta}_{ij}(\mathbf{d})$  by fitting model (1) without genetic covariates to the real ADNI data set introduced in the next section. The elements in  $\beta_j^{(g)}(\mathbf{d})$  for  $j = 1$  and 2 corresponding to the pre-specified pairs of causal SNPs and affected ROIs were set to affect magnitude  $\{\beta_j^*, j = 1, 2\}$  and zero otherwise. In addition, the affected ROI associated with the causal SNPs was pre-fixed as a circular region with radius  $r$  (Fig. 2).

We simulated genetic data  $z_i(g)$  as follows. We used linkage disequilibrium (LD) blocks defined by the default method (Gabriel, 2002) of Haploview (Barrett et al., 2005) and PLINK (Purcell et al., 2007) to form SNP sets. To calculate the LD blocks,  $n$  subjects were simulated by

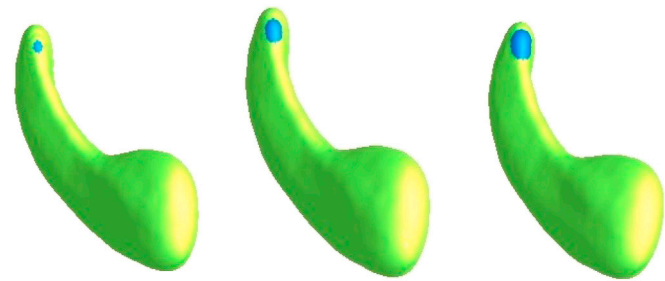


Fig. 2. Simulation settings: Green and blue regions in each panel respectively represent the right hippocampal surface and the affected ROI associated with the causal SNPs. From left to right, the radii of the affected ROIs are respectively set to 3, 6, and 9.

randomly combining haplotypes of HapMap CEU subjects. We used PLINK to determine the LD blocks based on these subjects. We randomly selected 2000 blocks, and combined haplotypes of HapMap CEU subjects in each block to form genotype variables for these subjects. We randomly selected 10 SNPs in each block; thus, we had  $N_G = 20,000$  SNPs for each subject. Moreover, we chose the first  $q$  SNPs as the causal SNPs. We set the sample size ( $n$ ) and the number of causal SNPs ( $q$ ) to be 1,000 and 100, respectively. Finally, we used 100 Monte Carlo realizations.

First, we evaluated the finite sample performance of the proposed GSIS for different settings of  $(N_0, \beta_1^*, \beta_2^*, \Omega)$ . Specifically, we set  $\beta_1^* = 0.01, \beta_2^* = a\beta_1^*$ , in which  $a$  was chosen from the set  $\{0.5, 1, 1.5, 2\}$ , and  $N_0$  between 100 and 2000. Moreover, we set the radius of the ROI as  $r = 6$ . We considered two different settings of  $\Omega$ , including (i)  $\Omega = \text{diag}(\sigma_1 = \sigma_2 = 0.5)$ ; and (ii)  $\Omega = \text{diag}(\sigma_1 = 0.8, \sigma_2 = 1)$ . We defined the causal SNP rate as the ratio of the number of causal SNPs in  $\mathcal{S}_0^*$  over the total number of causal SNPs. Table 1 includes the causal SNP rates under different settings. As expected, the causal SNP rate increases as  $N_0$  and  $a$  increase. However, the causal SNP rate is low for small  $N_0$ , especially when  $\sigma_1$  and  $\sigma_2$  are quite large. When  $N_0$  is set to be larger than 900, almost all causal SNPs are included in the set  $\mathcal{S}_0^*$  for most settings of  $(\beta_1^*, \beta_2^*, \Omega)$ . See Table 1 for more details.

Second, we evaluated the finite sample performance of FGWAS when model is misspecified. Assume that the  $y_{ij}(\mathbf{d})$ s were generated from the model given by

$$y_{ij}(\mathbf{d}) = \mathbf{x}_i^T \beta_j^{(c)}(\mathbf{d}) + \sum_{g=1}^{100} z_i(g) \beta_j^{(g)}(\mathbf{d}) + \beta^l(\mathbf{d}) x_{i2} \sum_{g=101}^{100+q^l} z_i(g) + \eta_{ij}(\mathbf{d}) + \varepsilon_{ij}(\mathbf{d}) \quad (17)$$

for  $i = 1, \dots, n$  and  $j = 1, 2$ . It can be seen that, the first  $100 + q^l$  SNPs are treated as causal SNPs here. The key difference between (17) and (16) is that there exists an extra interaction term (non-genetic effect  $\times$  genetic effect) in (17). Here  $\mathbf{x}_i, \beta_j^{(c)}(\mathbf{d}), \eta_{ij}(\mathbf{d})$ , and the genetic data  $z_i(g)$  were set in the same way as those in (16). The elements of genetic effect coefficient  $\beta_1^* = \beta_2^* = \beta^* = 0.02$ . The genetic data  $\{z_i(g_k)\}_{k=1}^{q^l}$  in the interaction term are observations from  $q^l$  casual SNPs. The interaction effect is assumed to be homogeneous with non-zero magnitude across the whole ROI and zero otherwise. Specifically, we set  $\beta^l = a^l \beta^*$ , in which  $a^l$  was chosen from the set  $\{0.1, 0.2, 0.5\}$ ,  $q^l$  was chosen from the set  $\{10, 20, 50\}$ , and  $N_0$  between 100 and 2000. Moreover, we set the radius of ROI as  $r = 6$ . The causal SNP rates under different settings were presented in Table 2. According to the results, it can be found that, when the

Table 1

Simulation results: causal SNP rates correspond to different settings of  $(N_0, \beta^*, \Omega)$  in the affected ROI, with radius  $r = 6$ . The causal SNP rate is defined as the ratio of the number of causal SNPs in  $\mathcal{S}_0^*$  over the total number of causal SNPs.

FGWAS: $(\beta_1^* = 0.01, \sigma_1 = 0.5, \sigma_2 = 0.5)$								
$\beta_2^*/\beta_1^*$	$N_0$							
	100	300	600	900	1,200	1,500	1800	2000
0.5	0.16	0.32	0.45	0.52	0.60	0.69	0.75	0.76
1	0.15	0.43	0.72	0.99	1.00	1.00	1.00	1.00
1.5	0.18	0.44	0.76	0.98	1.00	1.00	1.00	1.00
2	0.19	0.44	0.78	0.99	1.00	1.00	1.00	1.00
FGWAS: $(\beta_1^* = 0.01, \sigma_1 = 0.5, \sigma_2 = 1)$								
$\beta_2^*/\beta_1^*$	$N_0$							
	100	300	600	900	1,200	1,500	1800	2000
0.5	0.04	0.05	0.08	0.12	0.17	0.20	0.21	0.24
1	0.16	0.32	0.45	0.52	0.59	0.69	0.74	0.76
15	0.15	0.42	0.67	0.94	0.98	0.99	1.00	1.00
2	0.18	0.44	0.72	0.98	1.00	1.00	1.00	1.00

**Table 2**

Simulation results (model misspecification): causal SNP rates correspond to different settings of  $(N_0, \beta^*, q^l)$  in the affected ROI with radius  $r = 6$ .

FGWAS: $(\beta^* = 0.02, q^l = 20, \sigma = 0.5)$								
$\beta^l / \beta^*$	$N_0$							
	100	300	600	900	1,200	1,500	1800	2000
0.1	0.08	0.15	0.27	0.40	0.63	0.77	0.91	0.99
0.2	0.10	0.16	0.30	0.44	0.69	0.86	0.98	1.00
0.5	0.14	0.22	0.45	0.68	0.85	0.99	1.00	1.00

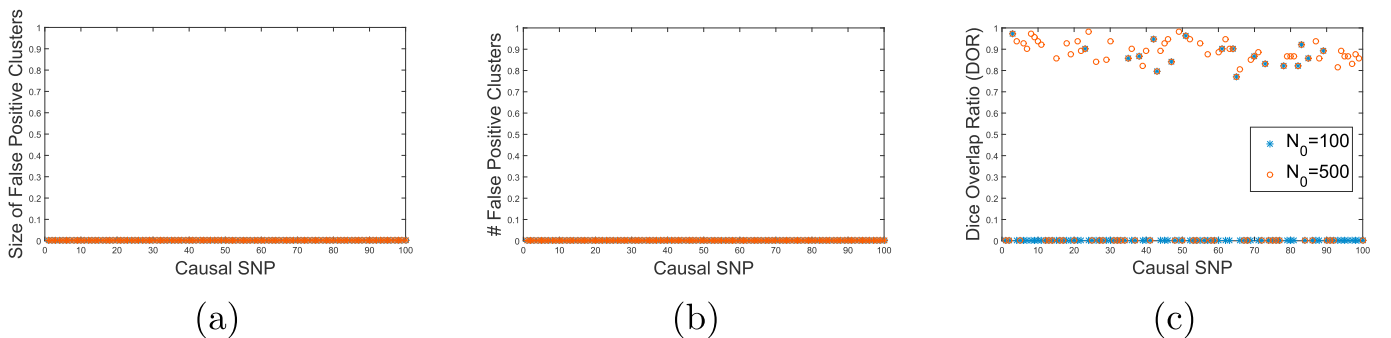
FGWAS: $(\beta^* = 0.02, \beta^l = 0.002, \sigma = 0.5)$								
$q^l$	$N_0$							
	100	300	600	900	1,200	1,500	1800	2000
10	0.07	0.13	0.23	0.34	0.57	0.74	0.88	0.94
20	0.08	0.15	0.27	0.40	0.63	0.77	0.91	0.99
50	0.11	0.17	0.33	0.46	0.69	0.85	0.99	1.00

model is misspecified, FGWAS can still achieve a high causal SNP rate when  $N_0$  is large. Nevertheless, the causal SNP rate increases when either the interaction effect or the number of non-causal SNPs included in the interaction term becomes larger.

Third, we evaluated the finite sample performance of FGWAS in detecting the causal SNP and subregion pairs. We set  $n = 1,000$ ,  $q = 100$ ,  $\Omega = 0.5I_2$ ,  $(\beta_1^*, \beta_2^*) = (0.01, 0.01)$ , and  $r = 6$ . Moreover, we used an uncorrected  $\alpha_l = 0.005$   $p$ -value threshold to identify subregions consisting of contiguous supra-threshold vertices. If the vertices in a thresholded cluster overlapped with some vertices in the affected ROI at a causal SNP, then we call these vertices “true positive vertices”. If a thresholded subregion did not overlap with any vertices of the affected ROI at any causal SNP, we call such a subregion a “false positive” subregion. We summarized the results by using the Dice overlap ratio (DOR), the number of false positive subregions, and the size in the number of vertices in false positive subregions. DOR is the ratio between the number of true positive pixels over the size of the affected ROI (Huang et al., 2015). Thus, a higher DOR means a higher probability of detecting the affected ROI. As shown in Fig. 3, no false positive subregion is detected. These results further demonstrate that the GSIS procedure can effectively detect and localize relatively strong genetic effects. Moreover, the average DOR of  $N_0 = 500$  is higher than that of  $N_0 = 100$ .

Fourth, we compared the proposed FGWAS method with other two methods, i.e., the standard functional GWAS (Reimherr and Nicolae, 2014) and the FVGWAS package (Huang et al., 2015). In order to make a fair comparison, we applied all three methods to the same simulated data sets. Since both standard functional GWAS and FVGWAS are feasible only for univariate imaging phenotypic measurements, we simplified model (16) by considering only one imaging measurement, i.e.,

$$y_i(\mathbf{d}) = \mathbf{x}_i^T \boldsymbol{\beta}^{(c)}(\mathbf{d}) + \sum_{g=1}^{N_G} z_i(g) \boldsymbol{\beta}^{(g)}(\mathbf{d}) + \eta_i(\mathbf{d}) + \varepsilon_i(\mathbf{d}), i = 1, \dots, n. \quad (18)$$



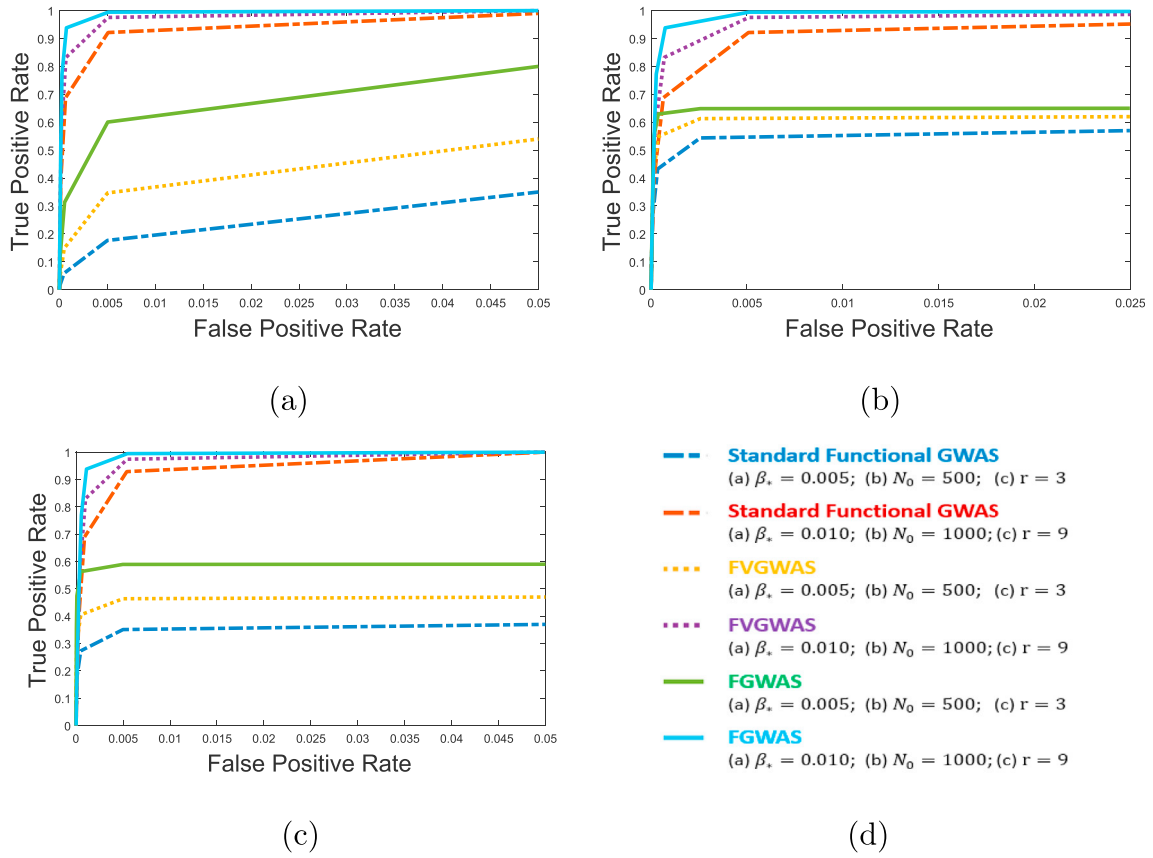
**Fig. 3.** Simulation results for the association between SNPs and subregions: (a) size in the number of vertices of false positive subregions in each causal SNP; (b) number of false positive subregions in each causal SNP; and (c) Dice overlap ratio (DOR) in each causal SNP. Parameters  $(\beta_1^*, \beta_2^*)$ ,  $\Omega$ , and  $r$  are set to  $(0.01, 0.01)$ ,  $0.5I_2$ , and 6, respectively.

Three factors are considered in the comparisons: (i) the genetic effect  $\beta_*(\mathbf{d})$  in the affected region, (ii) the number of candidate significant loci  $N_0$ , and (iii) the radius  $r$  of the affected ROI. In order to illustrate how each factor affects the finite sample performance of the three methods, we fixed two factors and chose different values for one factor in each setting. Fig. 4 presents the receiver operating characteristic (ROC) curves for all three methods, corresponding to three different cases. In case 1, the genetic effect is set to  $\beta_* = 0.005$  and  $\beta_* = 0.01$ , whereas other parameters  $\Omega$ ,  $n$ ,  $N_0$ , and  $r$  are set to 0.5, 1,000, 1,000, and 6, respectively. In case 2, the number of candidate loci is set to  $N_0 = 500$  and  $N_0 = 1,000$ , whereas other parameters  $\Omega$ ,  $n$ ,  $\beta_*$ , and  $r$  are set to 0.5, 1,000, 0.01, and 6, respectively. In case 3, the radius of the affected ROI is set to  $r = 3$  and  $r = 9$ , whereas other parameters  $\Omega$ ,  $n$ ,  $N_0$ , and  $\beta_*$  are set to 0.5, 1,000, 1,000, and 0.01, respectively. It can be found that, for each case, as the factor increases, the areas under the ROC curves for all the methods increase as well. Moreover, FGWAS outperforms both the standard functional GWAS and FVGWAS in all three cases, indicating that compared with standard functional GWAS and FVGWAS, FGWAS dramatically boosts the power for detecting various settings of genetic effects and affected ROIs.

#### 4. ADNI hippocampal surface data analysis

##### 4.1. ADNI data description

Data used in the preparation of this article were obtained from the ADNI database ([adni.loni.usc.edu](http://adni.loni.usc.edu)). The ADNI was launched in 2003 by the National Institute on Aging, National Institute of Biomedical Imaging and Bioengineering, Food and Drug Administration, private pharmaceutical companies and non-profit organizations as a \$60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography, 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). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as lessening the time and cost of clinical trials. The principal investigator of this initiative is Michael W. Weiner, MD, at the VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The goal was to recruit 800 subjects, but the initial study (ADNI-1) has been followed by ADNI-GO and ADNI-2. To date, these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the



**Fig. 4.** Simulation results for comparisons among FVGWAS, standard functional GWAS, and FGWAS in identifying significant voxel-locus pairs: (a) Case 1. ROC curves of all three methods with  $\beta_s = 0.005$  and  $\beta_e = 0.01$ , whereas other parameters  $\Omega$ ,  $n$ ,  $N_0$ , and  $r$  are set to 0.5, 1000, 1000, and 6, respectively. (b) Case 2. ROC curves of all three methods with  $N_0 = 500$  and  $N_0 = 1,000$ , whereas other parameters  $\Omega$ ,  $n$ ,  $\beta_s$ , and  $r$  are set to 0.5, 1000, 0.01, and 6, respectively. (c) Case 3. ROC curves of all three methods with  $r = 3$  and  $r = 9$ , whereas other parameters  $\Omega$ ,  $n$ ,  $N_0$ , and  $\beta_s$  are set to 0.5, 1000, 1000, and 0.01, respectively.

option to be followed in ADNI-2. For up-to-date information, see [www.adni-info.org](http://www.adni-info.org).

#### 4.2. Data processing

We applied FGWAS to the joint analysis of anatomical MRI and genetic data collected through ADNI-1. In this data analysis, we included 708 MRI scans from healthy controls and individuals with AD or MCI (186 AD, 388 MCI, and 224 healthy controls) from ADNI-1. The scans (from 462 men and 336 women, ages  $75.42 \pm 6.83$  years), which were performed on a variety of 1.5 T MRI scanners with protocols individualized for each scanner, include standard T1-weighted images obtained using volumetric 3-dimensional sagittal MPRAGE or equivalent protocols with varying resolutions. The typical protocol includes: repetition time = 2400 ms, inversion time = 1000 ms, flip angle =  $8^\circ$ , and field of view = 24 cm, with a  $256 \times 256 \times 170$  acquisition matrix in the  $x$ -,  $y$ -, and  $z$ -dimensions, which yields a voxel size of  $1.25 \times 1.26 \times 1.2$  mm<sup>3</sup>. We processed the MRI data by using standard steps, including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removing, intensity inhomogeneity correction, segmentation, and registration. Subsequently, we carried out automatic regional labeling by labeling the template and by transferring the labels following the deformable registration of subject images. After labeling 93 ROIs, we were able to compute volumes for each of these ROIs for each subject.

We adopted a hippocampal subregional analysis package based on surface fluid registration (Shi et al., 2013) that uses isothermal coordinates and fluid registration to generate one-to-one hippocampal surface registration for computing the surface statistics. Then, we computed the various surface statistics on the registered surface, such as multivariate tensor-based morphometry statistics, which retain the full

tensor information of the deformation Jacobian matrix, together with the radial distance, which retains information on the deformation along the surface normal direction. More details can be found in (Wang et al., 2011).

We considered the 818 subjects' genotype variables acquired by using the Human 610-Quad BeadChip (Illumina, Inc., San Diego, CA) in the ADNI database, which includes 620,901 SNPs. To reduce the population stratification effect, we used data from 749 Caucasians among all 818 subjects with complete imaging measurements at baseline. Quality control procedures included (i) call rate check per subject and per SNP marker, (ii) gender check, (iii) sibling pair identification, (iv) the Hardy-Weinberg equilibrium test, (v) marker removal by MAF, and (vi) population stratification. The second line preprocessing steps included removal of SNPs with (i) more than 5% missing values, (ii) MAF smaller than 5%, and (iii) Hardy-Weinberg equilibrium  $p$ -value  $< 10^{-6}$ . Remaining missing genotype variables were imputed as the modal value. After the quality control procedures, 708 subjects and 501,584 SNPs remained in the final data analysis.

#### 4.3. Data analysis

The hippocampus is believed to be involved in memory, spatial navigation and memory, and behavioral inhibition. In AD, the hippocampus is one of the first regions of the brain to be affected, leading to the confusion and loss of memory so commonly seen in the early stages of the disease. Recent work has revealed that the hippocampus is structurally and functionally asymmetric, and hippocampal asymmetry changes with AD progression, with the left hippocampus affected first by dementia, followed by atrophy in the right hippocampus after a time lag.

The objective of this data analysis was to examine the genetic effect of



each of 501,584 SNPs on either the left or right hippocampus and whether the genetic pathway of the left hippocampus overlaps with that of the right after partialing out the genetic effect of APOE  $\epsilon 4$ . To achieve this objective, we applied FGWAS with either the left or right hippocampal surface data as the functional phenotypes. Specifically, we chose the radial distance and determinant of Jacobian matrix as two different surface measurements. Moreover, in model (1), we included an intercept, gender, age, APOE  $\epsilon 4$ , and the top 5 principal component scores of all SNPs as covariates. We had 708 subjects, 30,000 vertices on the hippocampal surface (15,000 on each side), and 501,584 SNPs. The number of candidate loci was set as  $N_0 = 2,000$ . Then, the total computational time was 91,423s and 92,091s for the left and right hippocampi, respectively.

We have the following findings. Fig. 5(a)–(d) shows the Manhattan and QQ plots of the GWAS results for the left and right hippocampal surfaces, respectively. Moreover, Fig. 6(a, b) shows the density of the global Wald-type statistic and its  $\chi^2$  approximation in the GSIS procedure, which are very close to each other, indicating the accuracy of the  $\chi^2$  approximation. Tables 3 and 4 present the top 50 SNPs associated with the left and right hippocampal surfaces. At the  $10^{-5}$  significance level, 11

SNPs were detected as being associated with the left hippocampal surface, while 17 SNPs were found to be associated with the right hippocampal surface. Fig. 7(a, b) presents the LocusZoom plot (Pruim et al., 2010), which shows the regional association results near the top 1 SNP (rs657132 on gene HRH4, chr 18) from the GSIS procedure on the left hippocampal surface, and the top 1 SNP (rs4681527 on gene C3orf58, chr 3) on the right hippocampal surface. In particular, histamine receptor H4 (HRH4) is a protein-coding gene, and disease associated with HRH4 includes cerebellar degeneration. Moreover, cholinergic receptor M4 (CHRM) is an important paralog of HRH4, and the loss of M4 receptors has been found in the hippocampus of AD patients (Mulugeta et al., 2003). Further information about all top 2000 SNPs on each side of the hippocampal surfaces are available online at "[http://odin.mdacc.tmc.edu/bigs2/Top\\_2000\\_SNPs.html](http://odin.mdacc.tmc.edu/bigs2/Top_2000_SNPs.html)". We also included the group effects into our MVCM and reran our FGWAS. The corresponding Manhattan and QQ plots are presented in Fig. 10 in Appendix. In order to compare the GWAS results obtained under the model including group effects with those under the model without group effects, the rank-rank analysis (Plaisier et al., 2010) was conducted to compare the rank consistency of

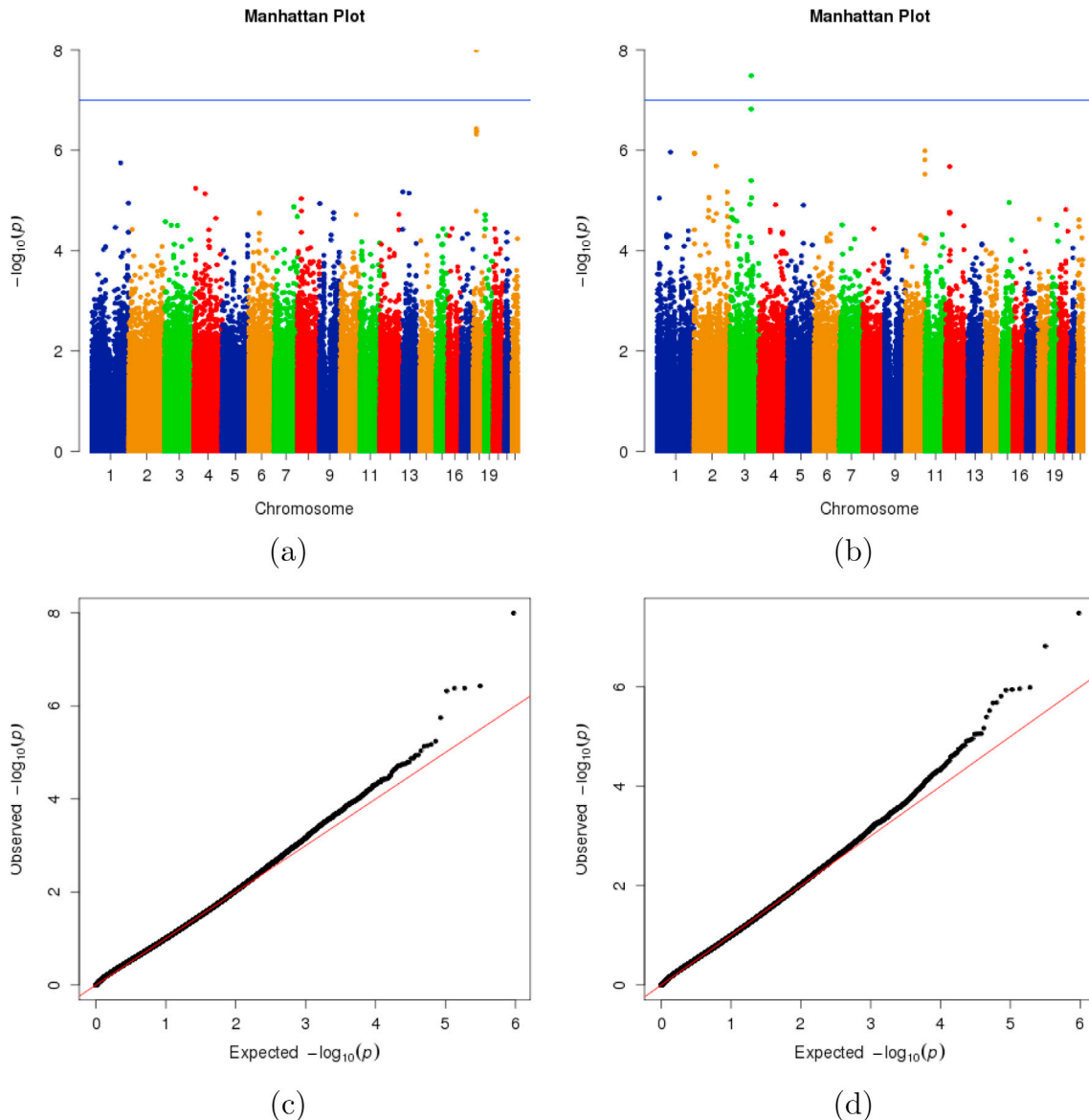
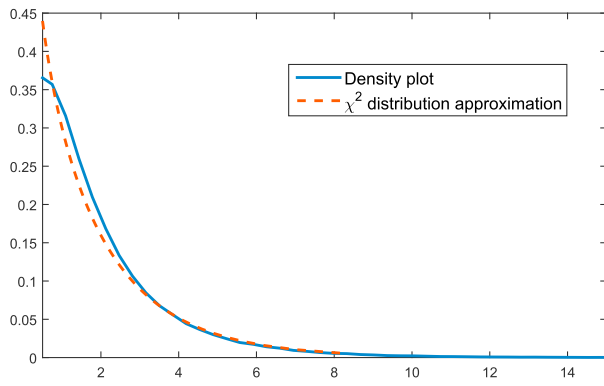
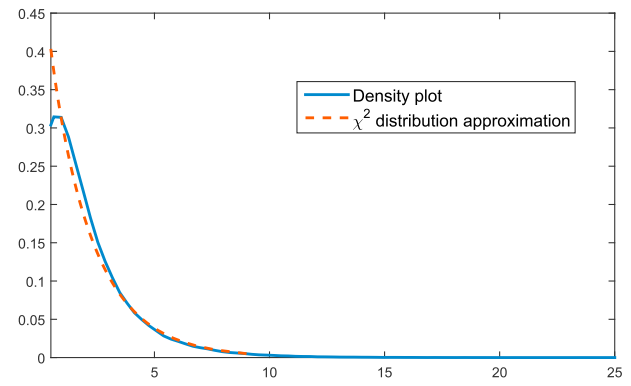


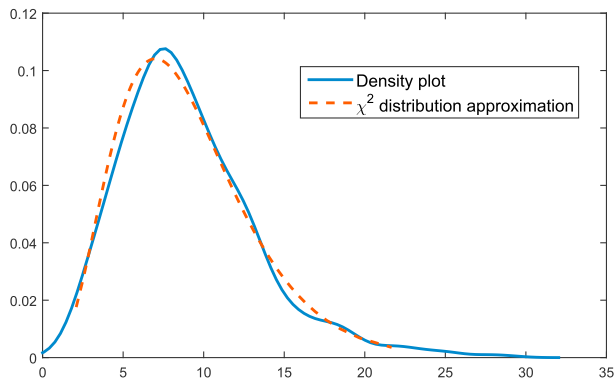
Fig. 5. ADNI hippocampal surface GWAS: (a) Manhattan plot (left hippocampal surface); (b) Manhattan plot (right hippocampal surface); (c) QQ plot (left hippocampal surface); (d) QQ plot (right hippocampal surface).



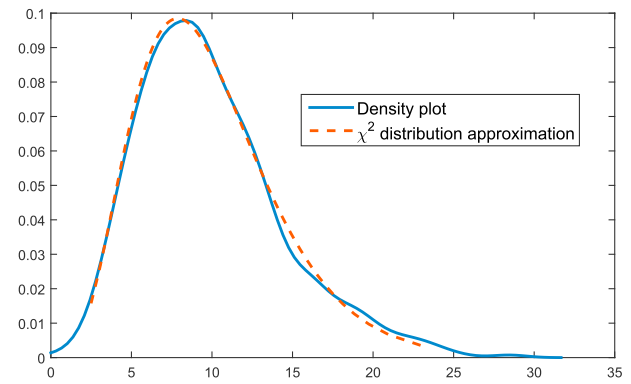
(a)



(b)



(c)



(d)

**Fig. 6.** ADNI hippocampal surface GWAS: Density plot of  $T_n(g)$  and its  $\chi^2$  distribution approximation from the GSIS procedure on the (a) left and (b) right hippocampal surfaces. The density plot of  $T_n(g^*, d)$  and its  $\chi^2$  distribution approximation from the bootstrapping procedure on (c) the left and (d) right hippocampal surfaces.

**Table 3**

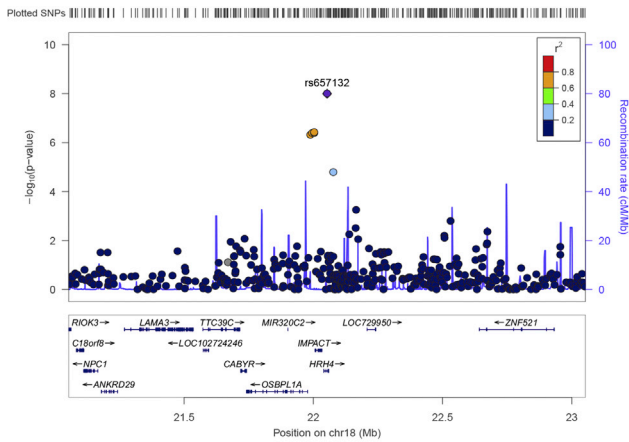
ADNI hippocampal surface GWAS: selected top 50 SNPs associated with the left hippocampal surface.

SNP	CHR	BP	<i>p</i> -value	Gene	SNP	CHR	BP	<i>p</i> -value	Gene
rs657132	18	22,053,274	1.01E-08	HRH4	rs10821312	9	96,944,581	2.31E-05	MIRLET7DHG
rs604345	18	22,003,302	3.71E-07	IMPACT	rs887500	19	2,862,785	2.47E-05	ZNF555
rs582110	18	21,995,436	4.17E-07	IMPACT	rs7631664	3	4,751,252	2.66E-05	ITPR1
rs546000	18	22,003,116	4.17E-07	IMPACT	rs730004	3	44,221,458	3.12E-05	TOPAZ1
rs489631	18	21,989,024	4.81E-07	IMPACT	rs6786876	3	86,462,805	3.16E-05	THAP12P2
rs16837577	1	194,870,594	1.79E-06	KCNT2	rs7528690	1	158,765,934	3.44E-05	OR2AQ1P
rs11730805	4	12,174,662	5.74E-06	HS3ST1	rs2037173	16	27,126,509	3.62E-05	C16orf82
rs9580112	13	19,317,385	6.74E-06	RP11-38M15.7	rs4807347	19	2,857,287	3.66E-05	ZNF555
rs3812872	13	61,986,918	7.16E-06	PCDH20	rs5011374	20	9,212,186	3.68E-05	PLCB4
rs6826085	4	76,870,229	7.38E-06	SDAD1	rs17360351	15	69,542,047	3.68E-05	GLCE
rs17197236	8	25,193,338	9.16E-06	DOCK5	rs9552579	13	19,377,679	3.75E-05	RP11-38M15.8
rs9783081	1	247,016,787	1.13E-05	AHCTF1	rs10495737	2	23,238,004	3.77E-05	AC016768.1
rs2890548	9	4,105,330	1.15E-05	GLIS3	rs1390931	4	101,762,475	3.82E-05	EMCN
rs2042067	7	132,651,302	1.34E-05	CHCHD3	rs10492041	12	126,592,053	3.83E-05	RP3-446N13.2
rs929714	7	132,630,046	1.34E-05	CHCHD3	rs2800219	1	247,055,427	4.29E-05	AHCTF1
rs2709627	8	25,174,710	1.62E-05	DOCK5	rs2800221	1	247,058,705	4.29E-05	AHCTF1
rs628674	18	22,077,035	1.63E-05	HRH4	rs2960173	8	25,186,301	4.29E-05	DOCK5
rs4744321	9	96,932,414	1.75E-05	RP11-2B6.3	rs2830058	21	27,484,968	4.36E-05	APP
rs9354911	6	70,971,130	1.78E-05	COL9A1	rs9446340	6	71,826,158	4.47E-05	U3
rs10806631	6	70,976,818	1.78E-05	COL9A1	rs2289672	17	42,932,244	4.63E-05	HIGD1B
rs7969873	12	128,129,387	1.89E-05	RP11-526P6.1	rs902498	5	171,560,741	4.78E-05	STK10
rs2418828	10	108,654,356	1.92E-05	SORCS1	rs1328531	9	80,389,921	4.87E-05	GNAQ
rs10418996	19	2,860,166	1.93E-05	ZNF555	rs9302364	15	93,699,827	4.88E-05	RGMA
rs12113716	7	155,975,007	2.09E-05	Y_RNA	rs12518980	5	175,168,827	5.03E-05	HRH2
rs17475359	4	148,566,463	2.26E-05	PRMT9	rs17670627	16	10,096,340	5.06E-05	GRIN2A

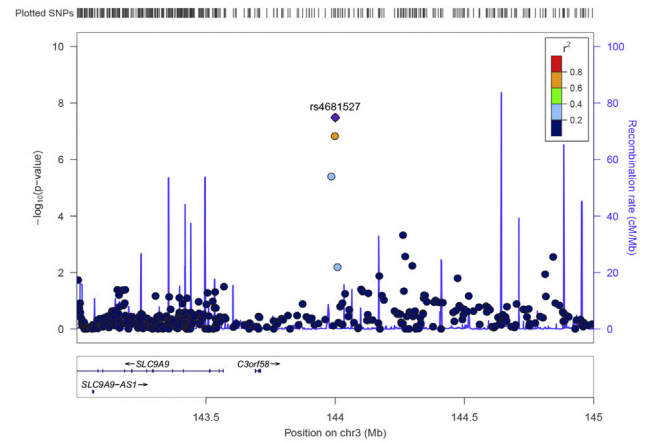
**Table 4**

ADNI hippocampal surface GWAS: selected top 50 SNPs associated with the right hippocampal surface.

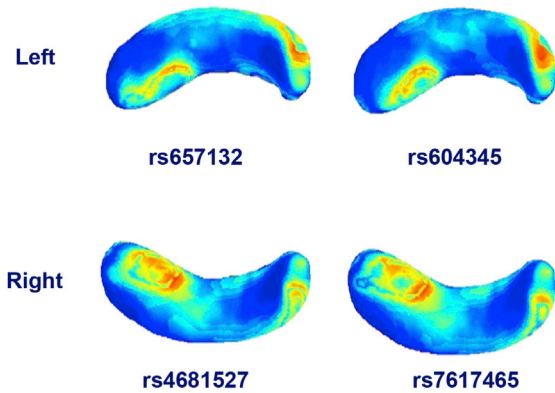
SNP	CHR	BP	p-value	Gene	SNP	CHR	BP	p-value	Gene
rs4681527	3	144,000,439	3.27E-08	RNA5SP144	rs10843350	12	29,385,946	1.81E-05	FAR2
rs7617465	3	143,998,527	1.52E-07	RNA5SP144	rs16828194	2	151,311,055	1.82E-05	RND3
rs12264728	10	132,139,574	1.02E-06	RP11-339B9.1	rs4674860	2	224,921,814	2.08E-05	SERPINE2
rs10801705	1	89,500,382	1.09E-06	GBP3	rs6798713	3	12,305,587	2.18E-05	GSTM5P1
rs749788	2	2,846,181	1.13E-06	AC011995.1	rs2871213	2	98,807,030	2.19E-05	VWA3B
rs823246	2	2,849,167	1.17E-06	AC011995.1	rs1719990	18	5,560,422	2.36E-05	EPB41L3
rs652911	10	132,139,875	1.54E-06	RP11-339B9.1	rs9847186	3	25,081,857	2.39E-05	AC133680.1
rs3108514	2	151,279,247	2.07E-06	RND3	rs6004683	22	25,898,162	2.42E-05	CRYBB2P1
rs7312068	12	29,435,225	2.11E-06	FAR2	rs1574605	2	130,653,885	2.52E-05	AC079776.1
rs366346	10	132,140,841	2.99E-06	RP11-339B9.1	rs7433347	3	46,716,523	2.59E-05	ALS2CL
rs1354316	3	143,985,479	4.04E-06	RNA5SP144	rs17141117	7	19,349,040	3.08E-05	AC007091.1
rs282268	2	224,920,176	6.77E-06	SERPINE2	rs2075650	19	45,395,619	3.08E-05	APOE
rs4599142	2	101,254,300	8.75E-06	NANOGNBP1	rs937341	12	125,093,430	3.22E-05	RP11-83B20.4
rs6542972	2	101,257,185	8.77E-06	FAT3	rs2013369	22	25,905,668	3.34E-05	CRYBB2P1
rs1512890	3	145,951,330	8.83E-06	PLSCR4	rs1367873	2	236,212,377	3.40E-05	AC114814.3
rs3817959	1	14,408,015	8.96E-06	KAZN	rs6988179	8	73,035,769	3.65E-05	MSC-AS1
rs748608	15	74,516,427	1.10E-05	CCDC33	rs12988342	2	2,849,451	3.83E-05	AC011995.1
rs4673085	2	224,925,417	1.15E-05	SERPINE2	rs7656798	4	75,387,973	3.87E-05	AC142293.3
rs11927997	3	135,140,012	1.19E-05	RP11-65709.1	rs1566602	1	211,191,060	4.04E-05	KCNH1
rs12499874	4	111,129,756	1.22E-05	ELOVL6	rs11906462	20	61,158,952	4.09E-05	C20orf166
rs13187102	5	107,995,317	1.24E-05	LINC01023	rs4349644	4	157,418,154	4.28E-05	RP11-171N4.4
rs16995208	20	49,036,909	1.51E-05	RN7SL636P	rs4694691	4	75,383,332	4.30E-05	AC142293.3
rs12635120	3	12,304,614	1.51E-05	AC091492.3	rs802682	6	111,121,188	4.60E-05	CDK19
rs6734569	2	99,156,296	1.58E-05	INPP4A	rs13150416	4	157,403,221	4.71E-05	RP11-171N4.4
rs9300188	12	29,388,162	1.71E-05	FAR2	rs10194024	2	224,890,840	4.74E-05	SERPINE2



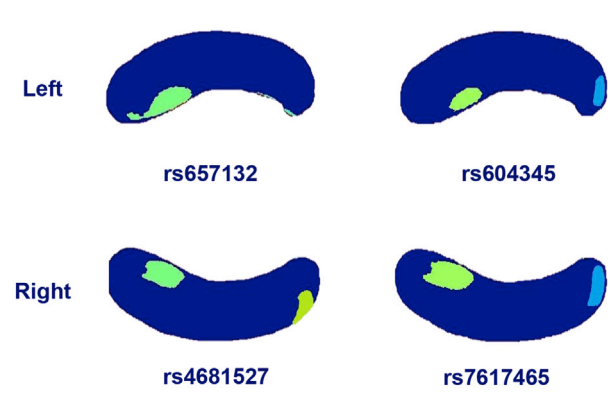
(a)



(b)



(c)



(d)

**Fig. 7.** ADNI hippocampal surface GWAS: LocusZoom plot showing the regional association results near (a) the top 1 SNP (rs657132) from the GSIS procedure on the left hippocampal surface and (b) the top 1 SNP (rs4681527) from the GSIS procedure on right hippocampal surface; (c) corrected  $-\log_{10}(p)$ -values across all vertices corresponding to top 2 SNPs on both the left and right hippocampal surfaces (rs657132 and rs604345 for the left hippocampal surface, rs4681527 and rs3108514 for the right hippocampal surface); (d)  $-\log_{10}(p)$ -values for significant subregions to top 2 SNPs on both the left and right hippocampal surfaces (rs657132 and rs604345 for the left hippocampal surface, rs4681527 and rs3108514 for the right hippocampal surface).

top 2000 genes detected in each model. It shows that the results under the two models are highly consistent with each other.

In Step (III), we first calculated the corrected  $p$ -values of  $T_n(g^*, d)$  across all vertices and candidate loci in  $\mathcal{S}_0^*$  to further detect significant vertex-locus pairs. In order to obtain the empirical distribution of  $T_n(g^*, d)$  under  $H_0$ , the wild bootstrap method was adopted. We set  $N_0 = 2,000$  and generated  $B = 500$  bootstrapped samples. We considered a parallel computing strategy and divided the genetic data into  $K_G = 10$  pieces. Fig. 6(c, d) shows the density plots of  $T_n(g^*, d)$  for  $N_0 = 2,000$ , corresponding to the left and right hippocampal surfaces, respectively, which are close to their  $\chi^2$  approximations. Subsequently, we calculated the corrected  $p$ -values of  $T_n(g^*, d)$ . Fig. 7(c) shows the corrected  $-\log_{10}(p)$ -values corresponding to the top 2 SNPs on the left and right hippocampal surfaces, where the color bar is presented as well.

In order to detect significant subregion-locus pairs, we set  $\alpha_l = 0.0001$  and calculated all possible subregions and their associated  $p$ -values against the top  $N_0 = 2,000$  SNPs. Fig. 7(d) shows the  $-\log_{10}(p)$ -values for significant subregions that correspond to the top 2 SNPs on both hippocampal surfaces, where the color bar is also presented. In particular, for the top 1 SNP, two subregions are found for the right hippocampal surface and one for the left hippocampal surface; whereas for the second top SNP, two subregions are found for each side of the hippocampal surfaces. Moreover, most significant subregions are likely to be symmetric across the left and right hippocampal surfaces. To specify the exact locations of significant subregions on the hippocampal surfaces, we recalled the cytoarchitectonic subregions mapped on blank MR-based models at 3 T of the hippocampal formation (Duvernoy, 2005; Frisoni et al., 2008), which are presented in Fig. 8(a). It shows that all the significant subregions are found in the CA1 subfield. Specifically, the most significant subregion (blue region indicated in Fig. 7(b)) is found on the lateral and medial aspects of the tail (CA1 subfield), and other subregions are found on the dorsolateral aspect of the head (CA1 subfield). It is interesting to note that volumes of similar hippocampal subregions were found to be affected in AD (Frisoni et al., 2008), indicating that the results obtained from FGWAS are in agreement with those of previous work.

We applied the rank-rank scatter plot in order to investigate the

connection between the genetic pathway for the left hippocampus and that for the right hippocampus. We first selected the top 2000 genes in the GWAS result on each side of the hippocampal surface and combined them, i.e., 3,562 genes in total. Note that the rank information of each gene was calculated based on the largest  $-\log_{10}(p)$ -value of SNPs associated with this gene. According to the rank information, the rank-rank scatter plot is presented in Fig. 8(b). It can be found that the two genetic pathways have weak connection. Specifically, only a few top genes have similar rank information on the left and right hippocampal surfaces, indicating that hippocampal asymmetry exists. In fact, the hippocampus was found to be structurally and functionally asymmetric in both healthy adults and AD patients (Shi et al., 2009; Maruszak and Thuret, 2013). Furthermore, different gene-environment interaction effects are found on different hippocampal subfields (Rabl et al., 2014). Therefore, the hippocampal asymmetry in our studies may be sound. Apart from the existence of hippocampal asymmetry, we wanted to examine whether any common genetic effect is associated with AD on both the left and right hippocampal surfaces. There is one gene, RBFox1, with similar rank information on both sides of the hippocampus (rank: 134 on the left hippocampal surface and 137 on the right hippocampal surface). Specifically, the amyloid precursor protein (APP) was found to be altered by transient RBFox1 expression in HEK293 and HeLa cells. Moreover, proteolytic processing of APP leads to the formation of  $\beta$ -amyloid (A $\beta$ ) peptides, which accumulate in the brains of those affected by AD (Ghiso and Frangione, 2002). Therefore, RBFox1, which presents a common genetic effect on both sides of the hippocampus, may play an important role in the progression of AD.

Finally, we calculated the polygenic risk score (PRS) at each vertex for multiple thresholds of  $p$ -values (i.e., 0.001, 0.01, 0.05, and 0.1). Then, we used a vertex-wise linear regression model and the coefficient of determination,  $R^2$ , to assess the proportion of variation in each imaging measurement that is explained by the PRS. The estimated  $R^2$  values across all vertices for two different types of imaging measurements are reported in Fig. 9. As the threshold increases, the estimated  $R^2$  values across all vertices increase. When the threshold is 0.05 or 0.1, the estimated  $R^2$  values are all above 0.5. Interestingly, comparing Fig. 9 with

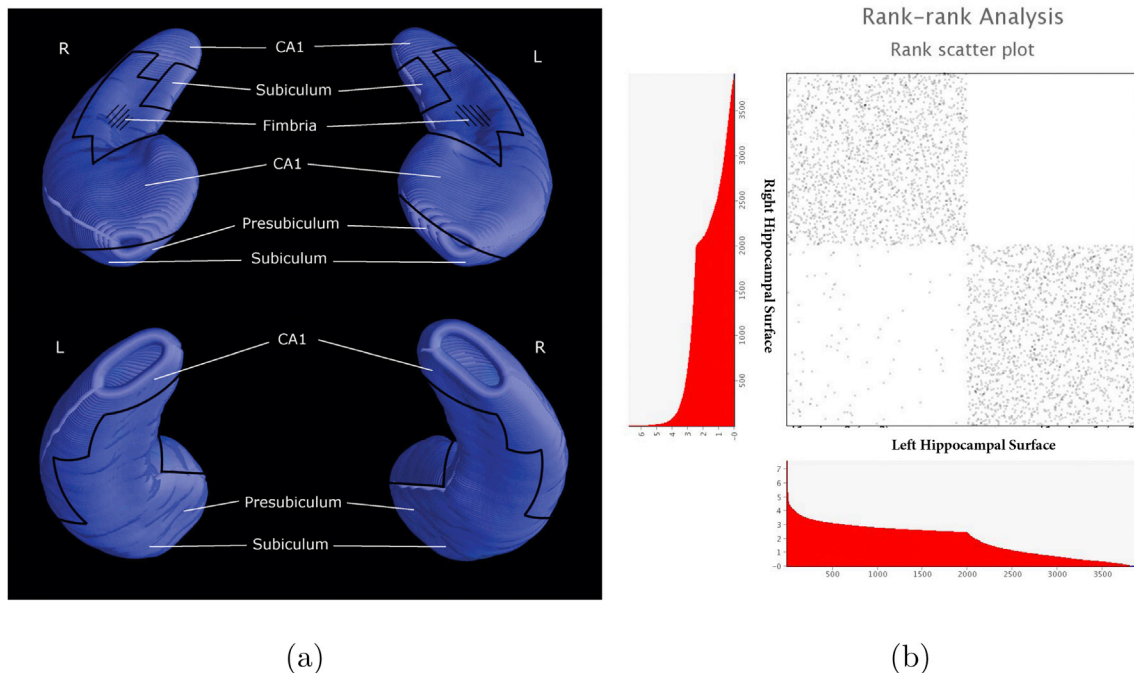
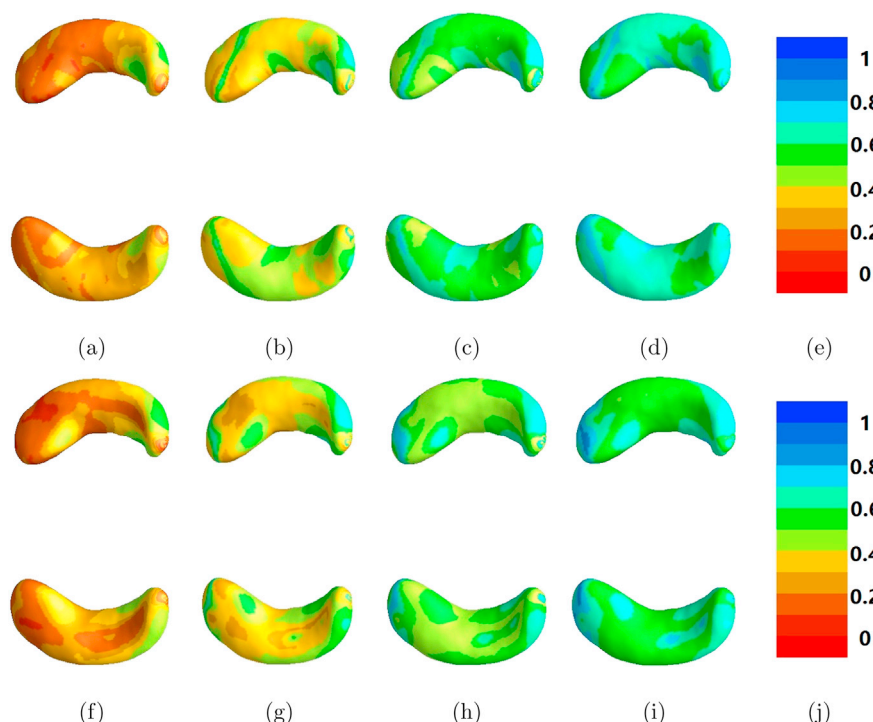


Fig. 8. ADNI hippocampal surface GWAS: (a) Cytoarchitectonic subregions mapped on blank MR-based models at 3 T of the hippocampal formation (Duvernoy, 2005; Frisoni et al., 2006, 2008); (b) Rank-rank scatter plots. The rank information of both the top 2000 genes on the left hippocampal surface and those on the right hippocampal surface, which were calculated based on the largest  $-\log_{10}(p)$ -values of SNPs associated with each gene.





**Fig. 9.** ADNI hippocampal surface GWAS: Top: estimated  $R^2$  for the imaging measurement (radial distance) across all vertices in multiple thresholds of p-values ((a) 0.001, (b) 0.01, (c) 0.05, and (d) 0.1); Bottom: estimated  $R^2$  for the imaging measurement (determinant of Jacobian matrix) across all vertices in multiple thresholds of p-values ((f) 0.001, (g) 0.01, (h) 0.05, and (i) 0.1).

Fig. 7(d) reveals that the estimated  $R^2$  values in the significant sub-regions, which were detected in Step (III), are larger than those in other regions and more likely increase as the threshold increases.

## 5. Conclusion and discussion

We have developed a FGWAS pipeline for efficiently carrying out genome-wide association analysis of surface-based imaging genetic data. Our proposed FGWAS consists of an MVCM, a GSIS procedure, and a detection procedure based on wild bootstrapping methods. Three key advantages of FGWAS have been discovered: (i) the spatial correlation structure of imaging data and variability of multiple phenotypic measurements considered in the multivariate varying coefficient model; (ii) much lower computational complexity compared to standard functional GWAS (Reimherr and Nicolae, 2014), and (iii) a parallel computing strategy that makes FGWAS feasible for super large-scale genetic data. Simulation studies have been conducted to evaluate the finite sample performance of FGWAS. We successfully applied FGWAS to hippocampal surface data and genetic data from the ADNI. Our FGWAS is a valuable statistical toolbox for fast, large-scale imaging genetic analysis.

There are two substantial issues to be addressed in our future research. First, since our FGWAS is still a single SNP analysis framework (Huang et al., 2015), the power of FGWAS may be undermined by unobserved causal SNPs, correlation among adjacent SNPs, and SNP-SNP interactions (Tzeng et al., 2011; Wu et al., 2011). It has been shown that alternative approaches for testing the association between a single SNP set and individual phenotypes are promising for improving the power of GWAS (Ge et al., 2012; Thompson et al., 2013). Therefore, it is of great importance to generalize our FGWAS for mapping the association between a SNP set and functional neuroimaging.

Second, in this paper, our FGWAS can only be used to detect the genetic markers that influence neuroimaging phenotypes. However, as in GWAS of AD, a question of interest is to test the null hypothesis of no association between functional phenotypes and the genotypes or genetic

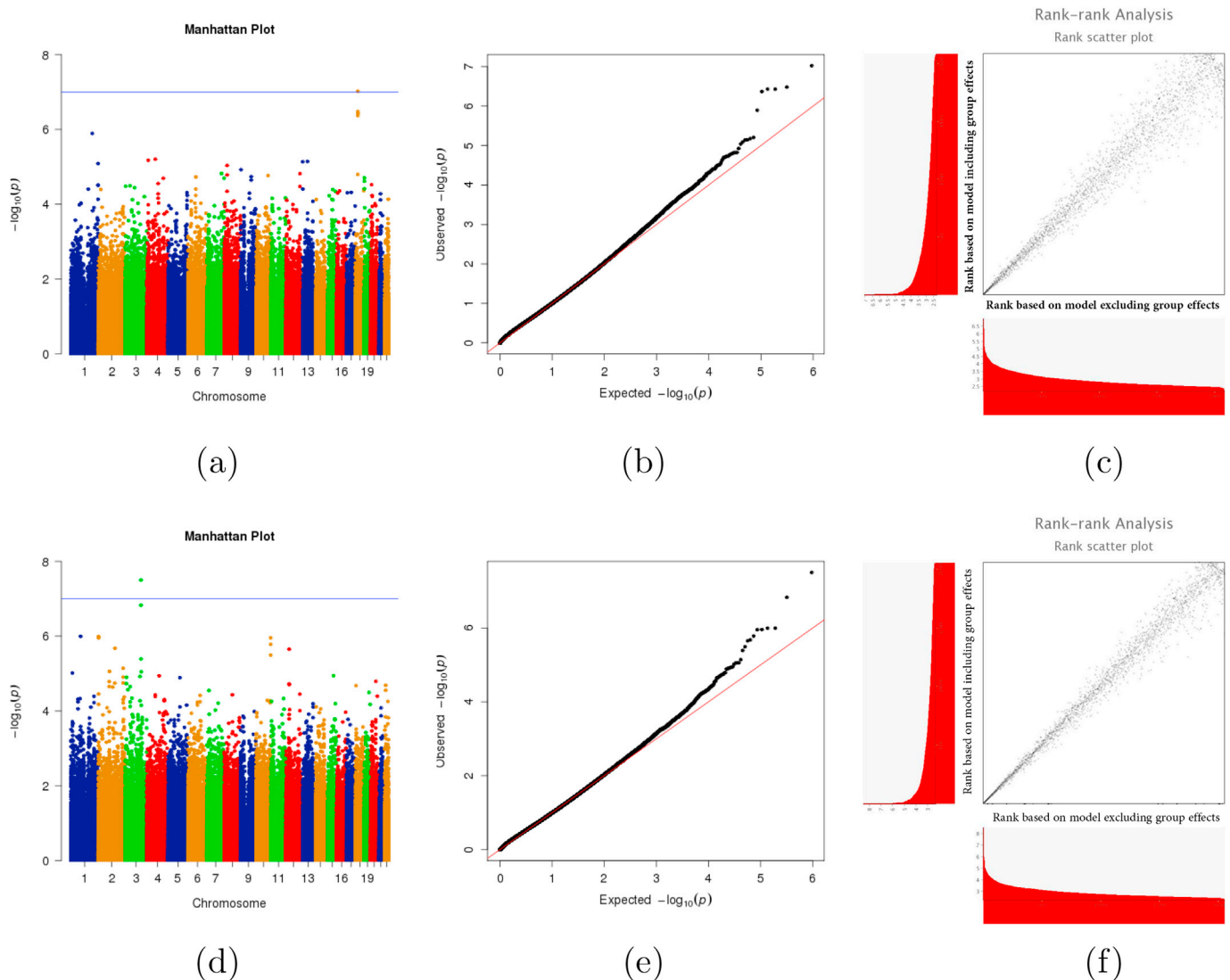
interactions (gene-environment), for example, genome-wide interaction analysis of relating SNPs to education level (Frost et al., 2016), case control conditions or memory scores (Yan et al., 2015). Meanwhile, detecting these interactions within genome-wide data can be challenging due to the loss in statistical power and computational efficiency. Therefore, generalizing our FGWAS for testing genetic interaction effects will be another aim in our future work.

## Acknowledgement

The collection and sharing of data that we analyzed for this project were 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: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; 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.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. 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](http://www.fnih.org)). The grantee organization 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 Neuro Imaging at the University of Southern California.

## Appendix

**Fig. 10:** ADNI hippocampal surface GWAS (model with group effects): (a) Manhattan plot (left hippocampal surface); (b) QQ plot (left hippocampal surface); (c) rank-rank scatter plot (left hippocampal surface); (d) Manhattan plot (right hippocampal surface); (e) QQ plot (right hippocampal surface); (f) rank-rank scatter plot (right hippocampal surface).



## References

- Barrett, J.C., Fry, B., Maller, J., Daly, M.J., 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21 (2), 263–265.
- Di, C.Z., Crainiceanu, C.M., Caffo, B.S., Punjabi, N.M., 2009. Multilevel functional principal component analysis. *Ann. Appl. Stat.* 3, 458–488.
- Duvernoy, H.M., 2005. The Human hippocampus: Functional Anatomy, Vascularization and Serial Sections with MRI. Springer Science & Business Media.
- Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman and Hall, London.
- Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. B* 70, 849–911.
- Fan, J., Zhang, W., 1999. Statistical estimation in varying coefficient models. *Ann. Stat.* 27 (5), 1491–1518.
- Fischl, B., 2012. Freesurfer. *Neuroimage* 62 (2), 774–781.
- Frisoni, G.B., Sabattoli, F., Lee, A.D., Dutton, R.A., Toga, A.W., Thompson, P.M., 2006. In vivo neuropathology of the hippocampal formation in ad: a radial mapping mr-based study. *Neuroimage* 32 (1), 104–110.
- Frisoni, G.B., Ganzola, R., Canu, E., Rüb, U., Pizzini, F.B., Alessandrini, F., Zoccatelli, G., Beltramello, A., Caltagirone, C., Thompson, P.M., 2008. Mapping local hippocampal changes in alzheimer's disease and normal ageing with mri at 3 tesla. *Brain* 131 (12), 3266–3276.
- Frost, H.R., Shen, L., Saykin, A.J., Williams, S.M., Moore, J.H., 2016. Identifying significant gene-environment interactions using a combination of screening testing and hierarchical false discovery rate control. *Genet. Epidemiol.* 40 (7), 544–557.
- Gabriel, S.B., 2002. The structure of haplotype blocks in the human genome. *Science* 296 (5576), 2225–2229.
- Ge, T., Feng, J., Hibar, D.P., Thompson, P.M., Nichols, T.E., 2012. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *NeuroImage* 63, 858–873.
- Ghisso, J., Frangione, B., 2002. Amyloidosis and alzheimer's disease. *Adv. Drug Deliv. Rev.* 54 (12), 1539–1551.
- Goodlett, C.B., Fletcher, P.T., Gilmore, J.H., Gerig, G., 2009. Group analysis of dti fiber tract statistics with application to neurodevelopment. *Neuroimage* 45, S133–S142.
- Guo, W., 2002. Functional mixed effects models. *Biometrics* 58 (1), 121–128.
- Hibar, D.P., Stein, J.L., Kohannim, O., Jahanshad, N., Saykin, A.J., Shen, L., Kim, S., Pankratz, N., Foroud, T., Huentelman, M.J., Potkin, S.G., Jack, C.R., Weiner, M.W., Toga, A.W., Thompson, P.M., Voxelwise, A.D.N.I., 2011. gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage* 56, 1875–1891.
- Huang, M., Nichols, T., Huang, C., Yang, Y., Lu, Z., Knickmeyer, R.C., Feng, Q., Zhu, H.T., 2015. Fvgwas: fast voxelwise genome wide association analysis of large-scale imaging genetic data. *NeuroImage* 118, 613–627.

- Lin, J., Zhu, H., Mihye, A., Sun, W., Ibrahim, J.G., 2014. Functional-mixed effects models for candidate genetic mapping in imaging genetic studies. *Genet. Epidemiol.* 38 (8), 680–691.
- Liu, J.Y., Calhoun, V.D., 2014. A review of multivariate analyses in imaging genetics. *Front. Neuroinf.* 8.
- Maruszak, A., Thuret, S., 2013. Why looking at the whole hippocampus is not enough—a critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for alzheimer's disease diagnosis. *Front. Cell. Neurosci.* 8 (95), 1–11.
- Medland, S.E., Jahanshad, N., Neale, B.M., Thompson, P.M., 2014. Whole-genome analyses of whole-brain data: working within an expanded search space. *Nat. Neurosci.* 17 (6), 791–800.
- Miller, M.I., Qiu, A., 2009. The emerging discipline of computational functional anatomy. *NeuroImage* 45, S16–S39.
- Morris, J.S., 2015. Functional regression. *Annu. Rev. Stat. Appl.* 2, 321–359.
- Mulugeta, E., Karlsson, E., Islam, A., Kalaria, R., Mangat, H., Winblad, B., Adem, A., 2003. Loss of muscarinic m4 receptors in hippocampus of alzheimer patients. *Brain Res.* 960 (1–2), 259–262.
- Nicolae, D.L., 2016. Association tests for rare variants. *Annu. Rev. Genomics Hum. Genet.* 17, 117–130.
- Plaisier, S.B., Taschereau, R., Wong, J.A., Graeber, T.G., 2010. Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res.* 38 (17-e169), 1–17.
- Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., Willer, C.J., 2010. Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26 (18), 2336–2337.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., et al., 2007. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81 (3), 559–575.
- Rabl, U., Meyer, B.M., Diers, K., Bartova, L., Berger, A., Mandorfer, D., Popovic, A., Scharinger, C., Huemer, J., Kalcher, K., Pail, J., Haslacher, H., Perkmann, T., Windischberger, C., Brocke, B., Sitte, H.H., Pollak, D.D., Dreher, J., Kasper, S., Praschak-Rieder, N., Moser, E., Esterbauer, H., Pezawas, L., 2014. Additive gene–environment effects on hippocampal structure in healthy humans. *J. Neurosci.* 34 (30), 9917–9926.
- Ramsay, J.O., Silverman, B.W., 2005. *Functional Data Analysis*. Springer-Verlag, New York.
- Reimherr, M., Nicolae, D., 2014. A functional data analysis approach for genetic association studies. *Ann. Appl. Stat.* 8 (1), 406–429.
- Shen, L., Kim, S., Risacher, S.L., Nho, K., Swaminathan, S., West, J.D., Foroud, T., Pankratz, N., Moore, J.H., Sloan, C.D., Huentelman, M.J., Craig, D.W., DeChairo, B.M., Potkin, S.G., Jack Jr., C.R., Weiner, M.W., Saykin, A.J., ADNI, 2010. 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.
- Shi, F., Liu, B., Zhou, Y., Yu, C., Jiang, T., 2009. Hippocampal volume and asymmetry in mild cognitive impairment and alzheimer's disease: meta-analyses of mri studies. *Hippocampus* 19 (11), 1055–1064.
- Shi, J., Thompson, P.M., Gutman, B., Wang, Y., 2013. Surface fluid registration of conformal representation: application to detect disease burden and genetic influence on hippocampus. *NeuroImage* 78, 111–134.
- Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* 44, 83–98.
- Smith, S.M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T.E., Mackay, C.E., Watkins, K.E., Ciccarelli, O., Cader, M.Z., Matthews, P.M., Behrens, T.E., 2006. Tractbased spatial statistics: voxelwise analysis of multi-subject diffusion data. *NeuroImage* 31, 1487–1505.
- Styner, M., Lieberman, J.A., McClure, D.R., Weinberger, R.K., Jones, D.W., Gerig, G., 2005. Morphometric analysis of lateral ventricles in schizophrenia and healthy controls regarding genetic and disease-specific factors. *Proc. Natl. Acad. Sci. U. S. A.* 102, 4872–4877.
- Thompson, P.M., Ge, T., Glahn, D.C., Jahanshad, N., Nichols, T.E., 2013. Genetics of the connectome. *NeuroImage* 80, 475–488.
- Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., et al., 2014. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging Behav.* 8 (2), 153–182.
- Tzeng, J., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M.I., Sale, M.M., Worrall, B.B., Hsu, F.C., Thomas, D.C., Sullivan, P.F., 2011. Studying gene and gene–environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am. J. Hum. Genet.* 89 (2), 277–288.
- Wang, J.L., Chiou, J.M., Muller, H.G., 2016. Functional data analysis. *Annu. Rev. Stat. Appl.* 3 (1), 257–295.
- Wang, Y., Song, Y., Rajagopalan, P., An, T., Liu, K., Chou, Y.Y., Gutman, B., Toga, A.W., Thompson, P.M., 2011. Surface-based tbm boosts power to detect disease effects on the brain: an n = 804 adni study. *NeuroImage* 56, 1993–2010.
- Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., Lin, X., 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89 (1), 82–93.
- Wu, R., Lin, M., 2006. Functional mapping — how to map and study the genetic architecture of dynamic complex traits. *Nat. Rev. Genet.* 7, 229–237.
- Yan, J., Kim, S., Nho, K., Chen, R., Risacher, S.L., Moore, J.H., Saykin, A.J., Shen, L., 2015. Hippocampal transcriptome-guided genetic analysis of correlated episodic memory phenotypes in alzheimer's disease. *Front. Genet.* 6.
- Yushkevich, P.A., Zhang, H., Simon, T.J., Gee, J.C., 2008. Structure-specific statistical mapping of white matter tracts. *Neuroimage* 41, 448–461.
- Zhang, J., Chen, J., 2007. Statistical inference for functional data. *Ann. Stat.* 35, 1052–1079.
- Zhang, Y.W., Xu, Z.Y., Shen, X.T., Pan, W., 2014. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage* 96, 309–325.
- Zhao, Y., Castellanos, F.X., 2016. Annual research review: discovery science strategies in studies of the pathophysiology of child and adolescent psychiatric disorders: promises and limitations. *J. Child Psychol. Psychiatry* 57, 421–439.
- Zhu, H., Kong, L., Li, R., Styner, M., Gerig, G., Lin, W., Gilmore, J.H., 2011. Fadtts: functional analysis of diffusion tensor tract statistics. *NeuroImage* 56, 1412–1425.
- Zhu, H., Fan, J., Kong, L., 2014. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *J. Am. Stat. Assoc.* 109, 977–990.
- Zhu, H.T., Li, R.Z., Kong, L.L., 2012. Multivariate varying coefficient model for functional responses. *Ann. Stat.* 40, 2634–2666.
- Zipunnikov, V., Caffo, B.S., Yousem, D.M., Davatzikos, C., Schwartz, B.S., Crainiceanu, C., 2011. Functional principal components model for high-dimensional brain imaging. *Neuroimage* 58, 772–784.