HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2016 September 01.

Published in final edited form as:

Neuroimage. 2015 September ; 118: 613–627. doi:10.1016/j.neuroimage.2015.05.043.

FVGWAS: Fast Voxelwise Genome Wide Association Analysis of Large-scale Imaging Genetic Data [1]

Meiyan Huang[a], **Thomas Nichols**[b], **Chao Huang**[c], **Yu Yang**[d], **Zhaohua Lu**[c], **Qianjing Feng**[a], **Rebecca C Knickmeyer**[e], **Hongtu Zhu**[c], and **the Alzheimer's Disease Neuroimaging Initiative**[*,1]

[a]School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China

[b]Department of Statistics, University of Warwick, Coventry, UK

[c]Department of Biostatistics and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[d]Department of Statistics and Operation Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[e]Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract

More and more large-scale imaging genetic studies are being widely conducted to collect a rich set of imaging, genetic, and clinical data to detect putative genes for complexly inherited neuropsychiatric and neurodegenerative disorders. Several major big-data challenges arise from testing genome-wide ($N_C > 12$ million known variants) associations with signals at millions of locations ($N_V \sim 10^6$) in the brain from thousands of subjects ($n \sim 10^3$). The aim of this paper is to develop a Fast Voxelwise Genome Wide Association analysiS (FVGWAS) framework to e ciently carry out whole-genome analyses of whole-brain data. FVGWAS consists of three components including a heteroscedastic linear model, a global sure independence screening (G-SIS) procedure, and a detection procedure based on wild bootstrap methods. Specifically, for standard linear association, the computational complexity is $O(nN_V N_C)$ for voxelwise genome wide association analysis (VGWAS) method compared with $O((N_C + N_V)n^2)$ for FVGWAS. Simulation studies show that FVGWAS is an effcient method of searching sparse signals in an extremely large search space, while controlling for the family-wise error rate. Finally, we have successfully applied

The readers are welcome to request reprints from Dr. Hongtu Zhu. hzhu@bios.unc.edu; Phone: 919-966-7272..

[1]This work was partially supported by NIH grants MH086633, 1UL1TR001111, and MH092335, NSF grants SES-1357666 and DMS-1407655, and National Natural Science Funds of China (NSFC, No. 31371009). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

[*]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

FVGWAS to a large-scale imaging genetic data analysis of ADNI data with 708 subjects, 193,275 voxels in RAVENS maps, and 501,584 SNPs, and the total processing time was 203,645 seconds for a single CPU. Our FVG-WAS may be a valuable statistical toolbox for large-scale imaging genetic analysis as the field is rapidly advancing with ultra-high-resolution imaging and whole-genome sequencing.

Keywords

Computational complexity; Family-wise error rate; Heteroscedastic linear model; Voxelwise genome wide association; Wild bootstrap

1. Introduction

With the advent of both imaging and genotyping techniques, many large biomedical studies have been conducted to collect imaging and genetic data and associated data (e.g., clinical data) from increasingly large cohorts in order to delineate the complex genetic and environmental contributors to many neuropsychiatric and neurodegenerative diseases, such as schizophrenia. Understanding such genetic and environmental factors is an important step for the development of urgently needed approaches to the prevention, diagnosis, and treatment of these complex diseases. Such studies and research projects include the Philadelphia Neurodevelopmental Cohort (PNC), the Alzheimer's Disease Neuroimaging Initiative (ADNI), and the Longitudinal Study of Early Brain Development (LSEBD), among others (NIH; Durston, 2010; Shen et al., 2010; Satterthwaite et al., 2014; Gilmore et al., 2010; Knickmeyer et al., 2014). These initiatives have generated many high-dimensional and complex data sets, referred to as big data, whose size is beyond the ability of commonly used software tools to capture, manage, and process data within a tolerable elapsed time. The real-time and proper analysis of such big data requires the development of fast and efficient data analysis methods.

There are three groups of methods for jointly analyzing imaging measurements and genetic variations. The first group focuses on candidate phenotypes and/or candidate genotypes using pre-screen methods or variable selection methods (Braskie et al., 2011). To adopt these approaches, one must have prior knowledge of the disease pathology in order to choose proper region of interest in imaging data or potential genetic variation of interest. The second group of methods performs voxel-wise genomic-wide association analysis that repeatedly fits a univariate model (e.g., linear regression model) to each voxel and single-nucleotide polymorphism (SNP) (or gene) pair following with multiple comparison adjustment to control for false positive finding (Hibar et al., 2011; Shen et al., 2010; Ge et al., 2012a). The third group of methods is to fit a very big model accommodating all (or part of) genetic variation and imaging measurements (Vounou et al., 2010, 2012; Zhu et al., 2014; Wang et al., 2012a,b). These methods use penalization-based method and sparse regression techniques, such as Lasso, to select putative genetic markers and a ected voxels. N-evertheless, this group of methods often cannot provide *p*-values and it usually results in a relatively small number of scattered voxels in imaging space.

Running VGWAS poses significant computational challenges, including limited computer memory, finite CPU speed, and limited CPU nodes, since it usually runs genome-wide ($N_C \sim 10^6$ known variants) associations with signals at millions of locations ($N_V \sim 10^6$) in the brain. It leads to a total of $N_C N_V$ ($\sim 10^{12}$) massive univariate analyses and an expanded image×gene search space with $N_C N_V$ elements (Medland et al., 2014; Thompson et al., 2014; Liu and Calhoun, 2014). As demonstrated in Stein et al. (2010), it took 300 high performance CPU nodes running approximately 27 hours to perform VGWAS analysis based on simple linear models with only a few covariates to process an imaging genetic dataset with 448,293 SNPs and 31,622 voxels in the brain of each of 740 subjects. As demonstrated in Hibar et al. (2011), it took 80 high performance CPU nodes running approximately 13 days to perform VGWAS analysis based on simple linear models with only a few covariates to process an imaging genetic dataset with 18,044 genes and 31,622 voxels in the brain of each of 740 subjects. One can imagine the computational challenges associated with VGWAS when the imaging genetics is advanced to the use of both ultra-high-resolution imaging ($N_V \sim 10^7$) and whole-genome sequencing ($N_C \sim 10^8$). A critical question is whether any scalable statistical method can be used to perform VGWAS efficiently for both imaging and genetic big data obtained from thousands of subjects.

The aim of this paper is to develop a Fast Voxelwise Genome Wide Association analysiS (FVGWAS) framework to efficiently carry out voxel-wise genomic-wide association (VGWAS) analysis. A schematic overview of FVGWAS is given in Fig. 1. There are four methodological contributions in this paper. The first one is to use a heteroscedastic linear model, which does not assume the presence of homogeneous variance across subjects and allows for a large class of distributions in the imaging data. These features are desirable for the analysis of imaging measurements, because between-subject and between-voxel variability in the imaging measures can be substantial and the distribution of the imaging data often deviates from the Gaussian distribution (Salmond et al., 2002; Zhu et al., 2007). The second one is to develop an efficient global sure independence screening (GSIS) procedure based on global Wald-test statistics, while dramatically reducing the size of search space from $N_C N_V$ to $\sim N_0 N_V$, in which $N_0 << N_C$. The GSIS procedure extends the notorious sure independence screening method (Fan and Lv, 2008; Fan and Song, 2010) from univariate responses to ultra-high dimensional responses. The third one is to use wild-bootstrap methods to testing hypotheses of interest associated with image and genetic data. In addition, the wild bootstrap methods do not involve repeated analyses of simulated datasets and therefore is computationally cheap. Moreover, the wild bootstrap method requires neither complete exchangeability associated with the standard permutation methods nor strong assumptions associated with random field theory. The fourth one is to reduce the computational complexity from $O(nN_V N_C)$ for standard VGWAS in (Stein et al., 2010) to $O((N_C + N_V)n^2$ for FVGWAS. When n << min(N_C, N_V), we have $O((N_C + N_V)n^2 = O(nN_V N_C)\times(nN_C^{-1}+nN_V^{-1})$, leading to a computational gain at the order of $O(\min(N_C, N_V)/n)$. Such computational gain makes it possible to run VGWAS on a single CPU. Finally, we will develop companion software for FVGWAS and release it to the public through http://www.nitrc.org/ and http://www.bios.unc.edu/research/bias.

The paper is organized as follows. Section 2 describes the three components of FVGWAS including a heteroscedastic linear model in Section 2.1, a global sure independence screening (GSIS) procedure in Section 2.2, and a detection procedure based on wild bootstrap methods in Section 2.3. In Section 3, we evaluate the finite-sample performance and computational efficiency of FVGWAS by using simulation studies and a real data analysis. In Section 4, some conclusions and discussions are provided.

2. Method

Suppose we observe a set of imaging measurements, clinical variables, and genetic markers from n unrelated subjects. Let $\mathscr{V}$ be a selected brain region with N_V voxels and v be a voxel in $\mathscr{V}$. Let $\mathscr{C}$ be the set of N_C SNPs and c be a locus in $\mathscr{C}$. For each individual i ($i = 1, \ldots, n$), we observe an $N_V \times 1$ vector of imaging measurements, denoted by $Y_i = \{y_i(v) : v \in \mathscr{V}\}$, a $K \times 1$ vector of clinical covariates $x_i = (x_{i1}, \cdots, x_iK)^T$, and an $L \times 1$ vector $\mathbf{z}_i(c) = (z_{i1}(c), \cdots, z_i(c))^T$ for genetic data at the c-th locus. For notational simplicity, only univariate image measurement (e.g. no tensors) is considered here.

The objective of this paper is to develop FVGWAS to efficiently carry out voxel-wise genomic-wide association analysis (VGWAS). As discussed above, since standard VGWAS consists of N_VN_C massive univariate analyses for all possible combinations of (c, v), it is computationally challenging and intensive to compute all N_VN_C test statistics and to store and manage all N_C test statistic images in limited computer hard drive. To solve these computational bottlenecks, we propose FVGWAS with three major components including

- (I) a heteroscedastic linear model;
- (II) a global sure independence screening procedure;
- (III) a detection procedure based on wild bootstrap methods.

We elaborate on each of these components below.

2.1 FVGWAS (I): Heteroscedastic Linear Model

We consider a heteroscedastic linear model (HLM) consisting of a heteroscedastic linear model at each voxel and a very flexible covariance structure. At each voxel v in $\mathscr{V}$, $y_i(v)$ can be modeled as a heteroscedastic linear model given by

$$y_i(v) = \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{v}) + \mathbf{z_i}(\mathbf{c})^{\mathbf{T}} \boldsymbol{\gamma}(\mathbf{c}, \mathbf{v}) + \mathbf{e_i}(\mathbf{v}) \quad \text{for} \quad \mathbf{i} = \mathbf{1}, \cdots, \mathbf{n}, \quad (1)$$

where $\boldsymbol{\beta}(v), = (\beta_1(v), \cdots, \beta K(v))^T$ is a $K \times 1$ vector associated with non-genetic predictors, and $\boldsymbol{\gamma}(c, v) = (\gamma_1(c, v), \cdots, \gamma_L(c, v))^T$ is an $L \times 1$ vector of genetic fixed effects (e.g., additive or dominant). Moreover, $e_i(v)$ are measurement errors with zero mean and $\{e_i(v) : v \in \mathscr{V}\}$ are independent across i. The spatial covariance structure of HLM assumes that $\mathbf{e}_i = \{e_i(v) : v \in \mathscr{V}\}$ has zero mean and a heterogeneous covariance structure, that is, $\mathrm{Cov}(\mathrm{e}_i)$ may vary across subjects. Since we do not impose any smoothness assumption on the covariance matrix of e_i as a function of v, HLM should be desirable for the analysis of real-world imaging measurements, which commonly have large variation across the image×gene

search space. Therefore, the assumptions of HLM are much weaker than those of random field theory (Hayasaka et al., 2004; Worsley et al., 2004; Hayasaka and Nichols, 2003).

Most GWAS focuses on the use of test statistics for a given phenotype to test the null hypothesis of no association at each loci. Here, we need to test

$$H_0(c,v):\boldsymbol{\gamma}(\mathbf{c},\mathbf{v})=\mathbf{0} \quad \text{versus} \quad \mathbf{H_1}(\mathbf{c},\mathbf{v}):\boldsymbol{\gamma}(\mathbf{c},\mathbf{v})\neq\mathbf{0} \quad \text{for each} \quad (\mathbf{c},\mathbf{v}). \quad (2)$$

We introduce the standard Wald-test statistic as follows. Let $\mathbf{Y}(v) = (y_1(v), \cdots, yn(v))^T$ and $P_X = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ be the projection matrix of model (1), where $\mathbf{X} = (\mathrm{x}_1, \cdots, \mathrm{x}n)$ is a $K \times n$ matrix. Similar to Zhu et al. (2007), we calculate an ordinary least squares estimate of $\boldsymbol{\gamma}(c, v)$, denoted by $\tilde{\boldsymbol{\gamma}}(c,v)$, given by

$$\tilde{\boldsymbol{\gamma}}(c,v)=\left\{\boldsymbol{Z}_c^T(\mathbf{I}_n-P_X)\boldsymbol{Z}_c\right\}^{-1}\boldsymbol{Z}_c^T(\mathbf{I}_n-P_X)\boldsymbol{Y}(v), \quad (3)$$

where $\mathbf{I}_n$ is an $n \times n$ identity matrix, $\mathbf{Z}_c = (\mathbf{z}_1(c), \cdots, \mathbf{z}n(c))$ is an $L \times n$ matrix. Ignoring heteroscedasticity in model (1) leads to an approximation of $Cov(\tilde{\gamma}(c,v))$ given by

$$Cov(\tilde{\boldsymbol{\gamma}}(c,v))\approx\sigma_e^2(c,v)\left\{\boldsymbol{Z}_c^T(\mathbf{I}_n-P_X)\boldsymbol{Z}_c\right\}^{-1}, \quad (4)$$

where $\sigma_e^2(c,v)$ is the variance of $e_i(v)$ under the homogeneous assumption of model (1). To test whether $\boldsymbol{\gamma}(c, v) = \mathbf{0}$ or not, we calculate a Wald-type statistic as

$$\begin{aligned} W(c,v) &= \tilde{\boldsymbol{\gamma}}(c,v)^T\{Cov(\tilde{\boldsymbol{\gamma}}(c,v))\}^{-1}\tilde{\boldsymbol{\gamma}}(c,v) \\ &= tr\left\{\left\{\boldsymbol{Z}_c^T(\mathbf{I}_n-P_X)\boldsymbol{Z}_c\right\}^{-1}\boldsymbol{Z}_c^T(\mathbf{I}_n-P_X)\sigma_e^{-2}(c,v)\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T(\mathbf{I}_n-P_X)\boldsymbol{Z}_c\right\}. \end{aligned} \quad (5)$$

Under the heterogeneity assumption of model (1), one may not use standard approximations based on the $\chi^2(L)$ (or F) distribution to approximate the null distribution of $W(c, v)$. As shown below, we can use the wild bootstrap method to approximate the null distribution of $W(c, v)$ even under such assumption for model (1), which can be desirable for real-world imaging data.

Several big-data challenges arise from the calculation of $W(c, v)$ as follows.

- (B1) Calculating $\sigma_e^2(c,v)$ across all (c, v)s' can be computationally intensive.
- (B2) Holding all $W(c, v)$ in the computer hard drive requires substantial computer resources.
- (B3) Speeding up the calculation of $W(c, v)$.

As shown below, the complexity of computing $\left\{\sigma_e^2(c,v), W(c,v)\right\}$ is at the order of $N_C N_V n^2$. Therefore, it is almost impossible to run a voxel-wise genome-wise association analysis in a single CPU.

To solve these computational bottlenecks, we propose two solutions as follows.

- (S1) Calculate $\sigma_e^2(c,v)$ under the null hypothesis $H_0(c, v)$ for each v and c.
- (S2) Develop a GSIS procedure to eliminate many 'noisy' loci based on a global Wald-type statistic.

By using (S1) and (S2), we are able to reduce the computational complexity from $O(N_C N_V n)$ to $O((N_C + N_V)n^2)$.

The key idea of (S1) is to estimate $\sigma_e^2(c,v)$ under the global null hypothesis $\boldsymbol{\gamma}(c, v) = \mathbf{0}$, which is similar to the well-known score test statistic. Under $H_0(c, v)$, we compute an unbiased estimate of $\sigma_e^2(c,v)$, denoted by $\hat{\sigma}_e^2(c,v)$, given by

$$\hat{\sigma}_e^2(c,v) = \boldsymbol{Y}(v)^T(\mathbf{I}_n - P_X)\boldsymbol{Y}(v)/(n-K). \quad (6)$$

Since $\hat{\sigma}_e^2(c,v)$ is invariant across all loci, we only need to calculate $\hat{\sigma}_e^2(c,v)$ at each voxel v and denote it as $\hat{\sigma}_e^2(v)$ from now on. The computational complexity of computing $\hat{\sigma}_e^2(v)$, is $O(n)$, and thus the total complexity of computing all $\left\{\hat{\sigma}_e^2(v)\right\}$ equals $O(N_V n)$. Therefore, computing $\left\{\hat{\sigma}_e^2(v)\right\}$ is about min(N_V, N_C) times faster than estimating $\sigma_e^2(c,v)$ under $H_1(c, v)$ for all possible (c, v). We will elaborate (S2) in the next subsection.

2. FVGWAS (II): A Global Sure Independence Screening Procedure

The key idea of (S2) is to extend the sure independence screening (SIS) procedure (Fan and Lv, 2008; Fan and Song, 2010; He and Lin, 2011). The key idea of GSIS is to first reduce the dimension from a very large scale to a moderate scale, and then select significant (c, v) pairs by using an approximation method. Specifically, we will use a global Wald-type statistic to eliminate many 'noisy' loci (no effect), since it is expected that only a small number of causal genetic markers contribute to the imaging phenotypic measures. The global Wald-type statistic at locus c is defined as

$$W(c) = N_V^{-1} tr\left\{\left\{\boldsymbol{Z}_c^T(\mathbf{I}_n - P_X)\boldsymbol{Z}_c\right\}^{-1}\boldsymbol{Z}_c^T(\mathbf{I}_n - P_X)\left\{\sum_{v\in\mathcal{V}}\hat{\sigma}_e(v)^{-2}\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T\right\}(\mathbf{I}_n - P_X)\boldsymbol{Z}_c\right\}. \quad (7)$$

The statistic $W(c)$ is an average of $W(c, v)$ across all v $\in V$ or an integration of $W(c, v)$ over $v \in V$. We choose $W(c)$ since detecting widespread genetic effects is more powerful and meaningful than testing for focal effects in neuroimaging. At a given locus c, $\mathcal{V}$ can be decomposed as the union of a true genetic effect region, denoted by $\mathcal{V}_S(c)$, and a false genetic effect region, denoted by $\mathcal{V}_{US}(c)$, such that $\mathcal{V} = \mathcal{V}_S(c) \cup \mathcal{V}_{US}(c)$ and $\mathcal{V}_S(c) \cap \mathcal{V}_{US}(c) = \varnothing$. If the volume of $\mathcal{V}_S(c)$ is relatively large and signals in $\mathcal{V}_S(c)$ are moderate, then the value of W(c) should be relatively large. Biologically, it is expected that important genetic markers should be associated with relatively large regions of interest (ROIs). However, a possible shortcoming of using $W(c)$ is that we may miss some loci with moderate signals in a small genetic effect region $\mathcal{V}_S(c)$. In contrast, observing large values of $W(c, v)$ in an extremely small effect region can be primarily caused by various noise

components, such as stochastic noise, susceptibility artifacts, or misalignment, in imaging data.

The complexity of computing $\{W(c)\}$ is at the order of $(N_C + N_V)n^2$, since $\left\{\sum_{v\in\mathcal{V}} \hat{\sigma}_e(v)^{-2}\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T\right\}$ is independent of c. In contrast, the complexity of computing $\{W(c, v)\}$ is at the order of $N_C N_V n$. Therefore, computing $\{W(c)\}$ is about $N_V N_C/\{(N_V + N_C)n\}$ times faster than computing all $\{W(c, v)\}$.

Our GSIS consists of the following steps:

- Step (II.1). Calculate $\Sigma_1 = (\mathbf{X}^T\mathbf{X})^{-1}$ with the computational complexity of $O(nK^2)$.
- Step (II.2). Calculate $\Sigma_2 = (\mathbf{I}_n - P_X)\left\{\sum_{v\in\mathcal{V}} \hat{\sigma}_e(v)^{-2}\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T\right\}(\mathbf{I}_n - P_X)$ with the computational complexity of $O(N_V n^2)$.
- Step (II.3). For the c–th locus, we do
  - Calculate $\boldsymbol{Z}_c^T\boldsymbol{Z}_c$ with the computational complexity of $O(L^2 n)$.
  - Calculate $\boldsymbol{Z}_c^T\mathbf{X}$ with the computational complexity of $O(LKn)$.
  - Calculate $\boldsymbol{Z}_c^T(\mathbf{I}_n - P_X)\boldsymbol{Z}_c = \left(\boldsymbol{Z}_c^T\boldsymbol{Z}_c\right) - \left(\boldsymbol{Z}_c^T\mathbf{X}\right)\Sigma_1\left(\mathbf{X}^T\boldsymbol{Z}_c\right)$ with the computational complexity of $O(L^2K^2)$.
  - Calculate $\boldsymbol{Z}_c^T\Sigma_2\boldsymbol{Z}_c$ with the computational complexity of $O(L^2n^2)$.
  - Calculate $W(c)$ with the computational complexity of $O(L^2)$.
- Step (II.4). Repeat Step (II.3) for all loci and calculate the p—value of $W(c)$, denoted by $p(c)$, across all loci by using an approximation method. Specifically, as shown in (Zhu et al., 2011; Zhang, 2005, 2011; Zhang and Chen, 2007), if $y_i(v)$ are treated as functional responses, then $W(c)$ asymptotically converges to a weighted χ^2 distribution as $n \to \infty$ when $H_0(c, v)$ holds for all (c, v) pairs. Let $\mathscr{K}_1(W)$, $\mathscr{K}_2(W)$, and $\mathscr{K}_3(W)$ be, respectively, the first three cumulants of $W(c)$. Therefore, following the reasonings in (Zhang, 2005), $W(c)$ can be approximated by a χ^2–type random variable $a_1\chi^2(a_2) + a_3$, where a_1, a_2, and a_3 are, respectively, given by

$$\alpha_1 = \frac{\mathscr{K}_3(W)}{4\mathscr{K}_2(W)}, \quad \alpha_2 = \frac{8\mathscr{K}_2^3(W)}{\mathscr{K}_3^2(W)}, \quad \text{and} \quad \alpha_3 = \mathscr{K}_1(W) - \frac{2\mathscr{K}_2^2(W)}{\mathscr{K}_3(W)}. \quad (8)$$

  We approximate $\{(\alpha_k, \mathscr{K}_k(W))\}_{k\le 3}$ by using the sample cumulants of $W(c)$ for k = 1, 2, 3. Finally, the p–value of $W(c)$ can be approximated by using $P\left(\chi^2(\alpha_2) \le [W(c) - \alpha_3]/\alpha_1\right)$. Note that the calculation of these p–values is not critical for the success of GSIS.
- Step (II.5). Sort the $-\log_{10}(p)$–values of all $W(c)$s' (or the values of $W(c)$s') according to their magnitudes and select the top N_0 loci (e.g., N_0 = 1000), denoted

by $\widetilde{\mathscr{C}}_0=\left\{\tilde{c}_1,\cdots,\tilde{c}_{N_0}\right\}$. From now on, we call $\widetilde{\mathscr{C}}_0$ as a *candidate significant locus set*.

There are some rationales of choosing $W(c)$ to determine $\widetilde{\mathscr{C}}_0$ and setting a relatively large N_0 in GSIS. As shown in simulations, if the volume of $\mathscr{V}_S(c)$ is relatively large and signal strength in $\mathscr{V}_S(c)$ is moderate, then $W(c)$ should put the c–th locus into $\widetilde{\mathscr{C}}_0$ This feature distinguishes VGWAS from eQTL analysis in the genetic literature (Sun, 2012; Shabalin, 2012). Moreover, we choose a relatively large N_0 so that the probability of all true positive loci contained in $\widetilde{\mathscr{C}}_0$ is relatively large. We will carry out simulations to evaluate such probability for different signal-to-noise ratios and sizes of $\mathscr{V}_{US}(c)$.

The accuracy of the χ^2–type approximation in Step (II.4) is not critical for the success of GSIS due to at least three reasons. First, since all loci share the same matrices P_X and $\sum_{v\in\mathscr{V}}\hat{\sigma}_e(v)^{-2}\boldsymbol{Y}(v)\boldsymbol{Y}(v)^T$ $W(c)$s slightly di er from each other only in term of $\mathbf{Z}_c$. Moreover, when $H_0(c, v)$ holds for all v for the c-th locus, the expectation of $W(c)$ is close to the dimension of $\mathbf{z}_i(c)$ (or L). Second, since it is expected that only a small number of causal genetic markers contribute to the imaging phenotypic measures, most $W(c)$ should roughly follow the same distribution and their empirical cumulants converge to $\mathscr{K}_k(W)$ under some mild conditions. Third, in ADNI data analysis presented in Section 3, we have found that such approximation is not only computationally simple, but also practically important. Specifically, for the whole brain analysis, the quantile-quantile (QQ) plots of $\{p(c)\}$ show a solid line matching expected=observed until it sharply curves at the end (representing the small number of true associations among thousands of unassociated SNPs). See Figs. 6 and 7 for details.

2.3. FVGWAS (III): A Detection Procedure Based on Wild Bootstrap Methods

Our detection procedure consists of two wild bootstrap methods:

(III.1) The first one is to detect significant voxel-locus pairs.

(III.2) The second one is to detect significant cluster-locus pairs.

The first wild bootstrap method is to simultaneously detect significant (locus, voxel) pairs. Conditional on the candidate significant locus set $\widetilde{\mathscr{C}}_0$ with the top N_0 loci, we calculate a maximum statistic over all voxels for the top N_0 loci as

$$W_{\mathscr{V},\widetilde{\mathscr{C}}_0}=\max_{\tilde{c}\in\widetilde{\mathscr{C}}_0,v\in\mathscr{V}}W(\tilde{c},v). \quad (9)$$

The maximum statistic $W_{\mathscr{V},\widetilde{\mathscr{C}}_0}$ plays a crucial role in controlling the family-wise error rate. The key idea of the first wild bootstrap method is to approximate the null distribution of $W_{\mathscr{V},\widetilde{\mathscr{C}}_0}$ under that the null hypothesis $H_0(c, v)$ holds for all $c\in\mathscr{C}$ and $v\in\mathscr{V}$.

We propose an efficient wild bootstrap procedure to detect significant $(\tilde{c},v)\in\widetilde{\mathscr{C}}_0\times\mathscr{V}$ as follows:

- Step (III.1.1). Calculate $W(\tilde{c}, v)$ for each pair $(\tilde{c}, v) \in \tilde{\mathscr{C}}_0 \times \mathscr{V}$ as

$$W(\tilde{c}, v) = tr\left\{\left\{\boldsymbol{Z}_c^T (\mathbf{I}_n - P_X) \boldsymbol{Z}_c\right\}^{-1} \boldsymbol{Z}_c^T (\mathbf{I}_n - P_X) \left\{\hat{\sigma}_e(v)^{-2} \boldsymbol{Y}(v) \boldsymbol{Y}(v)^T\right\} (\mathbf{I}_n - P_X) \boldsymbol{Z}_c\right\}.$$

The computational complexity is $O(N_V\, nN_0)$.

- Step (III.1.2). Calculate $W_{\mathscr{V}, \tilde{\mathscr{C}}_0}$.

- Step (III.1.3). Apply the first wild bootstrap method to the set $\tilde{\mathscr{C}}_0$.

  – Fit a linear model $y_i(v) = \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{v}) + \mathbf{e_i}(\mathbf{v})$ to imaging data and calculate $\hat{e}_i(v) = y_i(v) - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}(v)$ for all *i* and *v*, where $\tilde{\boldsymbol{\beta}}(v) = \left(\sum_{i=1} \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \sum_{i=1}^n \mathbf{x}_i y_i(v)$. Generate *G* bootstrap samples $\mathbf{y}_i^{(g)}(v) = \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}(v) + \eta_i^{(g)} \hat{e}_i(v)$ for all *i* and *v*, where $\eta_i^{(g)}$ are independently generated from a *N*(0, 1) generator. The key idea of this step is to generate imaging data from model (1) satisfying $\boldsymbol{\gamma}(c, v) = 0$ for all $(c, v) \in \mathscr{C} \times \mathscr{V}$, while asymptotically preserving the spatial correlation structure of imaging data. We can show that this data generating process can asymptotically preserve the spatial dependence structure among the imaging data under the null hypotheses. Specifically, the average conditional covariance between the residuals $\left\{\eta_i^{(g)} \hat{e}_i(v)\right\}_{i \le n}$ at voxel v and the residuals $\left\{\eta_i^{(g)} \hat{e}_i\left(v'\right)\right\}_{i \le n}$ at v′ given the raw imaging data is given by

$$n^{-1} \sum_{i=1}^n Cov\left(\eta_i^{(g)} \hat{e}_i(v), \quad \eta_i^{(g)} \hat{e}_i\left(v'\right) | \text{Data}\right) = n^{-1} \sum_{i=1}^n \hat{e}_i(v) \hat{e}_i\left(v'\right).$$

It follows from the law of large number that $n^{-1} \sum_{i=1}^n \hat{e}_i(v) \hat{e}_i\left(v'\right)$ converges to the spatial covariance between voxels *v* and v′.

– Let $\boldsymbol{\eta}^{(g)} = \left(\eta_1^{(g)}, \cdots, \eta_n^{(g)}\right)^T$ and $\hat{E}(v) = diag\left(\hat{e}_1(v), \cdots, \hat{e}_n(v)\right)$, we calculate $W(\tilde{c}, v)^{(g)}$ given by

$$tr\left[\left\{\boldsymbol{Z}_c^T (\mathbf{I}_n - P_X) \hat{E}(v)^2 (\mathbf{I}_n - P_X) \boldsymbol{Z}_c\right\}^{-1} \boldsymbol{Z}_c^T (\mathbf{I}_n - P_X) \hat{E}(v) \boldsymbol{\eta}^{(g)} \boldsymbol{\eta}^{(g)T} \hat{E}(v) (\mathbf{I}_n - P_X) \boldsymbol{Z}_c\right] \quad (10)$$

for all $(\tilde{c}, v) \in \mathscr{C}_0 \times \mathscr{V}$, which leads to $W^{(g)}_{\mathscr{V}, \tilde{\mathscr{C}}_0}$.

– For all $c \in \mathscr{C}$, we calculate $W(c)^{(g)} = \sum_{v \in \mathscr{V}} W(c, v)^{(g)} = \boldsymbol{\eta}^{(g)T} S(c) \boldsymbol{\eta}^{(g)}$, where $S(c)$ is given by

$$\sum_{v\in\mathscr{V}}\hat{E}(v)(\mathbf{I}_n-P_X)\boldsymbol{Z}_c\left\{\boldsymbol{Z}_c^T(\mathbf{I}_n-P_X)\hat{E}(v)^2(\mathbf{I}_n-P_X)\boldsymbol{Z}_c\right\}^{-1}\boldsymbol{Z}_c^T(\mathbf{I}_n-P_X)\hat{E}(v). \quad (11)$$

– Sort all $W(c)^{(g)}$s according to their magnitudes and select the top N_0 loci to form $\widetilde{\mathscr{C}}_0^{(g)}$. Calculate $W^{(g)}_{\mathscr{V},\widetilde{\mathscr{C}}_0^{(g)}}=max_{\widetilde{c}\in\widetilde{\mathscr{C}}_0^{(g)},v\in\mathscr{V}}\left\{W(\widetilde{c}.v)^{(g)}\right\}$

- Step (III.1.4). Calculate approximated Chi-squared distributions based on the bootstrapped samples $\left\{W(\tilde{c},v)^{(g)}\right\}_{g=1,\cdots,G}$ and calculate the uncorrected *p*–values of $W(\widetilde{c},v)$ across $(\widetilde{c},v)\in\widetilde{\mathscr{C}_0}\times\mathscr{V}$.
- Step (III.1.5). Calculate the family-wise error (FWE) corrected *p*–values of $W(\widetilde{c},v)$ across all $(\widetilde{c},v)\in\widetilde{\mathscr{C}_0}\times\mathscr{V}$ based on the empirical distribution of $\left\{W^{(g)}_{\mathscr{V},\widetilde{\mathscr{C}}_0^{(g)}}:g=1,\cdots,G\right\}$. At given significance level a, we can detect significant $(\widetilde{c},v)$ pairs in $\widetilde{\mathscr{C}_0}\times\mathscr{V}$. Since the number of pairs N_CN_V in $\mathscr{C}\times\mathscr{V}$ is much larger than the sample size, we choose a significance level, say a = 0.5.

There are three key advantages of using the wild bootstrap method in (III.1). First, it is robust to several key assumptions of normal linear model, such as Gaussian noise and homogeneous variance. See extensive simulations in Zhu et al. (2007) for the evaluation of the wild bootstrap method at the voxel level. Second, it automatically accounts for spatial correlations among imaging data and those among genetic data. Then, based on all bootstrapped samples at each locus c_0, we use the same approximation method in Step (II.4) to approximate the null distribution of the test statistic $W(c, v)$. By using such parametric approximation, we are able to obtain *p*-values that are better behaved (i.e., that are not necessarily multiples of 1/*G*, as would be the case if $\left\{W(c,v)^{(g)}\right\}_{g=1,\cdots,G}$ were used directly). Third, since S(c) is independent of $\boldsymbol{\eta}^{(g)}$, it is computationally efficient to calculate $W(c)^{(g)}$ and sort them in order to compute $\widetilde{\mathscr{C}}_0^{(g)}$.

The second wild bootstrap method is to simultaneously detect significant cluster-locus pairs. In neuroimaging, cluster size inference has been widely used to assess the significance of all numbers of interconnected voxels greater than a given threshold, say a_I = 0.005 or 0.001 (Ge et al., 2012b; Salimi-Khorshidi et al., 2011; Smith and Nichols, 2009; Hayasaka et al., 2004). For the *c*-th locus, let $N(c, _I)$ be the largest cluster size at a given threshold a_I based on the *p*-values of $\{W(c,v):v\in\mathscr{V}\}$. To detect significant (locus, cluster) pairs, we consider a maximum cluster size statistic and its approximation as

$$N(\mathscr{C},\alpha_1)=\max_{c\in\mathscr{C}}N(c,\alpha_I)\approx N\left(\widetilde{\mathscr{C}_0},\alpha_I\right)=\max_{c\in\widetilde{\mathscr{C}_0}}N(c,\alpha_I). \quad (12)$$

Given $\widetilde{\mathscr{C}_0}$ and the definition of $W(c)$ (Eq. (7)), it is expected that $N(\mathscr{C},\alpha_I)$ is very close to $N\left(\widetilde{\mathscr{C}_0},\alpha_I\right)$ both in terms of both size and distribution.

We propose an efficient wild bootstrap procedure to detect significant cluster-locus pairs as follows:

- Step (III.2.1). For a given a_I, we use the wild bootstrap method in Step (I-II.1.3) to generate $\left\{W(c,v)^{(g)}: c \in \widetilde{\mathscr{C}}_0^{(g)}, v \in \mathscr{V}, g=1,\cdots,G\right\}$ and calculate $N\left(\widetilde{\mathscr{C}}_0^{(g)}, \alpha_I\right)^{(g)}$ for each wild bootstrap sample. For computational efficiency, we suggest to directly compare $W(c, v)^{(g)}$ with the $100(1 - aI)$th percentile of the F distribution in order to determine clusters at each locus c.
- Step (III.2.2). For each locus $c \in \widetilde{\mathscr{C}}_0$, we calculate all possible clusters and their associated FWE-corrected p–values based on the empirical distribution of $\left\{N\left(\widetilde{\mathscr{C}}_0^{(g)}, \alpha_I\right)^{(g)}\right\}_{g=1,\cdots,G^*}$

3. Simulation Studies and ADNI Data Analysis

In this section, we use Monte Carlo simulations and a real example to evaluate the finite-sample performance of FVGWAS. All computations for these numerical examples were done in Matlab on a Dell C6100 server. The computation for FVGWAS is efficient even for large scale imaging genetic data and its computational time can be further reduced by using other computer languages, such as C++.

3.1. Simulation Studies

We simulated imaging data at $N_V = 3,355$ pixels in the brain region of a 128×128 image, which is a middle slice of a brain volume obtained from the public accessible data of the Alzheimer's Disease Neuroimage Initiative (ADNI). More information on the ADNI data used in the current study will be given in the ADNI Data Analysis Section. We assumed that the genetic e ect of SNPs is additive and homogeneous such that $y_i(v)$ were generated from:

$$y_i(v) = \mathbf{x}_i^T \boldsymbol{\beta}(\mathbf{v}) + \sum_{\mathbf{j}=\mathbf{1}}^{\mathbf{N_C}} \boldsymbol{\gamma}(\mathbf{c_j}, \mathbf{v})\, \mathbf{z_i}(\mathbf{c_j}) + \mathbf{e_i}(\mathbf{v}), \quad (13)$$

where $e_i(v) \sim N(0, \sigma^2)$, $z_i(c_j)$ were simulated genetic data as described below, and $x_i = (1, x_{i1}, \ldots, x_{i9})^T$ were designed to mimic the covariates used in ADNI data analysis and were generated from either $U(0, 1)$ or the Bernoulli distribution with success probability 0.5. The true values of $\beta(v)$ were set to be those of estimated $\beta(v)$ by fitting model (1) without genetic covariates to real ADNI data set in the real data analysis section. The elements in $\gamma(c_j, v)$ corresponding to the pre-specified pairs of causal SNPs and e ected Regions Of Interest (ROI) were set to e ect magnitude γ_*, zero otherwise. In addition, the e ected ROI associated with the causal SNPs was pre-fixed as a $r \times r$ region (Fig. 2).

We simulated genetic data $z_i(c_j)$ 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 LD blocks, n subjects were simulated by randomly combining haplotypes of HapMap CEU subjects. We used PLINK to determine the LD blocks based on these subjects. We randomly selected 2,000 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, and thus we had N_C = 20, 000 SNPs for each subject. Moreover, we chose the first q SNPs as the causal SNPs. We set the sample size (n), the number of causal SNPs (q), the standard deviation of measurement error (σ), and the size of effected ROI ($r \times r$) to be 1000, 100, 1, and 10×10, respectively. 100 Monte Carlo realisations were used.

First, we evaluate the finite sample performance of the proposed GSIS for $\gamma_* = 0.005$, 0.010, 0.015, 0.020, 0.025, and N_0's ranging between 200 and 2000. Moreover, we set $q = 100$. We measure the causal SNP rate, which is defined as the ratio of the number of causal SNPs in $\widetilde{\mathscr{C}}_0$ over the total number of causal SNPs. Table 1 includes the causal SNP rates corresponding to different top N_0 SNPs and γ_* values. As expected, the causal SNP rate increases as the number of top N_0 SNPs and γ_* increase. However, the causal SNP rate is low for $N0 = 100$. One may use large N_0 in order to increase the probability of including all causal SNPs in $\widetilde{\mathscr{C}}_0$. Specifically, when N_0 was set as 2000, almost all causal SNPs are included in the set C_0 even for small γ_*, such as $\gamma_* = 0.005$. See Table 1 for more details.

Second, we evaluate the finite sample performance of FVGWAS in the detection of the causal SNPs and voxels in the a ected ROIs as N_0 varies from 100, 500, to 1000. Moreover, parameter q was set to 100. The panels in the first row of Fig. 3 show Receiver Operating Characteristic (ROC) curves corresponding to different γ_* and N_0 values. As expected, large γ_* values representing larger genetic effects lead to a higher probability of detecting the causal SNPs and their associated voxels in the effected ROIs. For $\gamma_* \geq 0.015$, the ROC curves maintain high true positive rates and low false positive rates when $N_0 = 1000$. Moreover, a larger N_0 usually leads to a higher true positive rate for different γ_* values, whereas a larger N_0 can lead to a higher false positive rate. Then, we set γ_* to be 0.010, which is a moderate signal, in order to investigate the effects of different ROI sizes, σ, and n on signal detection. As expected, in the second row of Fig. 3, the true positive rate increases as the size of ROI increases; in the third row of Fig. 3, the true positive rate decreases as the value of σ increases; in the fourth row of Fig. 3, the true positive rate decreases with the sample size.

Third, we evaluate the finite sample performance of FVGWAS in detecting the causal SNP and cluster pairs. We set $n = 1000$, $q = 100$, $\sigma = 1$, $\gamma_* = 0.01$, and $r = 10$. Moreover, we used an uncorrected 0.01 p-value threshold to identify clusters of contiguous supra-threshold pixels. If the pixels in a thresholded cluster overlaps with some pixels in the effected ROI at a causal SNP, we call these pixels as “true positive pixels”. If a thresholded cluster does not overlap with any pixels of the effected ROI at any causal SNP, we call a cluster as a “false positive” cluster. We summarized results by using the dice overlap ratio (DOR), the number of false positive clusters, and the size in the number of pixels in false positive clusters. DOR is the ratio between the number of true positive pixels over the size of the effected ROI. Thus, the higher DOR means the higher probability of detecting the effected ROI. As shown in Fig. 4, there is no false positive cluster 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 method with the Matrix eQTL method (Shabalin, 2012) for pixel-wise inference. For a fair comparison, we applied both the Matrix eQTL and FVGWAS to the same simulated data sets. We set γ_* to be either 0.005 or 0.01. Fig. 5 presents the ROC curves corresponding to different N_0 and γ_* values. The proposed method outperforms the Matrix eQTL method when $\gamma_* = 0.005$, indicating that the proposed method is more capable than the Matrix eQTL method to detect small genetic effects. For $\gamma_* = 0.01$, the true positive rates of the proposed method with $N_0 = 100$ and $N_0 = 500$ are lower than those of the Matrix eQTL method, whereas the true positive rates of the proposed method with $N_0 = 1000$ are higher than those of the Matrix eQTL method. Specifically, the false positive rate of FVGWAS is lower than that of the Matrix eQTL method for diffrent N_0 and γ_* values. This result is primarily attributed to the GSIS procedure used in FVGWAS.

Fifth, we set $\boldsymbol{\gamma}(c, v) = \mathbf{0}$ for all (c, v) in order to assess the overall Type I error rates. We calculated the family-wise error rate (FWER) for the Type I error rates at both the voxel-locus and cluster-SNP levels (Dudoit et al., 2003; Shaffer, 1995). The significance level was varied from 0.1 to 0.5, and 1000 replications were used to estimate FWERs. For a fixed α, if the FWER is smaller than α, then the test is conservative, whereas if the FWER is greater than α_1, then the test is anticonservative, or liberal (Hayasaka and Nichols, 2003). Moreover, α_I was set to 0.005. Table 2 lists the FWERs corresponding to different N_0 α and values. For detecting significant SNP and voxel pairs, the rejection rates of the proposed method are accurate with large N_0 and α values. Moreover, for relatively small $\alpha = 0.1$, no significant SNP and cluster pairs are detected.

3.2. ADNI Data Analysis

To illustrate the usefulness of FVGWAS, we considered anatomical MRI data collected at the baseline by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. “Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.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 organizations, as a $60 million, 5-year publicprivate 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 Alzheimer's disease (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. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of e orts 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 initial goal of ADNI was to recruit 800 subjects but ADNI 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 option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org."

The brain MRI data were provided by the ADNI database, which can be downloaded from http://adni.loni.usc.edu/. We considered 708 MRI scans of AD, MCI, and healthy controls (186 AD, 388 MCI, and 224 healthy controls) from ADNI1 in this data analysis. These scans on 462 males and 336 females (age 75.42 ± 6.83 years) were performed on a 1.5 T MRI scanners using a sagittal MPRAGE sequence. The typical protocol includes the following parameters: repetition time (TR) = 2400 ms, inversion time (TI) = 1000 ms, flip angle = 8°, and field of view (FOV) = 24 cm with a 256 × 256 × 170 acquisition matrix in the *x*–, *y*–, and *z*–dimensions, which yields a voxel size of 1.25 × 1.26 × 1.2 mm^3.

We processed the MRI data by using standard steps including anterior commissure and posterior commissure correction, skull-stripping, cerebellum removal, intensity inhomogeneity correction, segmentation, and registration (Shen and Davatzikos, 2004). After segmentation, we segmented the brain data into four different tissues: grey matter (GM), white matter (WM), ventricle (VN), and cerebrospinal fluid (CSF). We used the deformation field to generate RAVENS maps (Davatzikos et al., 2001) to quantify the local volumetric group differences for the whole brain and each of the segmented tissue type (GM, WM, VN, and CSF), respectively. Moreover, we automatically labeled 93 ROIs on the template and transferred the labels following the deformable registration of subject images (Wang et al., 2011). We computed the volumes of all ROIs for all subjects.

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 749 Caucasians from all 818 subjects with complete imaging measurements at baseline. Quality control procedures include (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 the minor allele frequency, and (vi) population stratification. The second line preprocessing steps include removal of SNPs with (i) more than 5% missing values, (ii) minor allele frequency 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.

We consider both ROI volumes and RAVENS maps to illustrate the wide applicability of FVGWAS. We carried out two different FVGWAS analyses: one is to use the volumes of 93 ROIs as multivariate phenotypic vectors and the other is to use RAVENS maps as whole-brain phenotypic vectors. In both analyses, we used model (1) and included an intercept, gender, age, whole brain volume, and the top 5 Principal Components scores in SNPs. Then, we tested the additive effect of each of 501,584 SNPs on either 93 ROI volumes or RAVENS maps. In particular, for RAVENS maps, with 708 subjects, 193,275 voxels, 501,584 SNPs, and $N_0 = 1000$, the total processing time was 203,645 s, of which 116 s was allotted for the GSIS procedure and 203,529 s was allotted for determining significant

voxel-locus and cluster-SNP pairs. Finally, as a comparison, we applied the Matrix eQTL to RAVENS maps to carry out VGWAS.

3.2.1. ROI Volumes—The Manhattan and *QQ* plots of GWAS for all volumes of 93 ROIs are shown in Fig. 6. In Fig. 6 (a), only SNP in TOMM40 in chromosome 19 passes the threshold 5×10^{-8} commonly used in GWAS. In the *QQ* plot (Fig 6 (b)), the observed *p*–values fit the expected *p*–values from the null hypothesis well for most of the *p*–values. The *p*–values in the upper tail of the distribution do show a significant deviation suggesting strong associations between these SNPs and the univariate image measures. Fig. 6 (c) and (d) shows the number of significant SNP-ROI pairs with different numbers of top N_0 SNPs. The number of significant SNP-ROI pairs decreases when the number of N_0 increases. These results may indicate that more important information (significant SNP pairs) can be identified when N_0 is small. Therefore, we can just select a small N_0 value in the first screen step, which is a huge save of both computational time and memory.

To test the effect of SNPs on the volumes of 93 ROIs, we first set N_0 as 1,000 and 2,000. In this case, we can only detect significant ROI-locus pairs by using Step (III.1). Specifically, we generated 1,000 bootstrapped samples $W^{(g)}_{\mathscr{V},\widetilde{\mathscr{C}}_0}$ for $g = 1, \ldots, G = 1, 000$ and then calculated the corrected *p*–values of $W(\tilde{c}, v)$ across all $(\tilde{c}, v) \in \widetilde{\mathscr{C}}_0 \times \mathscr{V}$. By setting the 0.5 significance level, we are able to detect 2 and 1 significant ROI-locus pairs for $N_0 = 1, 000$ and 2, 000, respectively. These 2 significant ROI-locus pairs are rs2075650(TOMM40) and hippocampal formation left and amygdala right, respectively.

We selected several ROIs that are known to be meaningful biomarkers for Alzheimer's disease: Hippocampus Left/Right (HL/HR) and Amygdala Left/Right (AL/AR). Then, we carried out GWAS for each of the four ROIs. The SNPs associated with volumes of ROIs are reported in Table 3, together with their corresponding chromosome numbers, genomic coordinates, and *p*–values. Among the identified SNP sets in Table 3, the famous ApoE and TOMM40 in chromosome (Chrs) 19 are known to be associated with Alzheimer's disease.

3.2.2. RAVENS Maps—Fig. 7 (a) and (b) shows the Manhattan and *QQ* plots of the GWAS results for RAVENS maps and Table 4 includes the top 30 SNPs associated with the whole brain. Fig. 8 (a) shows the density of the global Wild-type statistic and its Chi-squared approximation for the whole brain. These two curves are very close to each other, indicating the accuracy of the χ^2 approximation. At the 10^{-5} significance level, 21 SNPs were detected to be associated with the whole brain in the GSIS analyses. For instance, these 21 SNPs include four SNPs in chromosome 10 (rs11815438, rs1060373, rs2480271, and rs2935713) and 2 SNPs (rs11891634 and rs13419007) in chromosome 2. Moreover, among the top $N_0 = 1, 000$ SNPs, we able to detect several important SNPs including rs2480271 on gene GLRX3 (chr 10), rs1534446 on gene PCEF1 (chr6), rs12436472 on gene NOVA1 (chr14), rs6116375 20 on gene PRNP (chr 20), rs4746622 on gene CTNNA3 (chr 10), rs4296809 on gene FGF10 (chr 5), rs439401 on gene APOE (chr 19), rs2075650 on gene TOMM40 (chr 19), rs3826810 on gene LDLR (chr 19), rs2679098 on gene NTRK3 (chr 15), and rs6896317 on gene TRIO (chr 5). Gene PRNP, gene CTNNA3, and gene LDLR are related to the Alzheimer's desease (Golanska et al., 2009; Miyashita et al., 2007; Gopalraj et

al., 2005). Gene NOVA1 is associated with aging and neurodegeneration (Tollervey et al., 2011). Gene NTRK3 is related to schizophrenia, bipolar disorder, and obsessive-compulsive disorder hoarding (Braskie et al., 2013). Gene TRIO is also related to schizophrenia (Stelzer et al., 2011). Further information about all top 1,000 SNPs will be available at http://www.bios.unc.edu/research/bias.

In Step (III.1), we first calculated the raw *p*-values of $W(c, v)$ in order to detect significant voxel-locus pairs. We set N_0 as either 1,000 or 2,000 and then generated 1,000 bootstrapped samples $\left\{W(c,v)^{(g)}\right\}$ for $g = 1, \ldots, G = 1, 000$. By using χ^2 approximation, we calculated the raw *p*–values of $W(\tilde{c}, v)$ across all $(\tilde{c}, v) \in \widetilde{\mathscr{C}_0} \times \mathscr{V}$. At the 10^{-5} significance level, Fig. 7 (c) and (d) show the number of significant voxel-locus pairs based on the raw *p*–values of $W(\tilde{c}, v)$ against the top N_0 SNPs in $\widetilde{C}_0$.

Second, we calculated the corrected *p*-values of $W(c, v)$ in order to detect significant voxel-locus pairs by correcting for multiple comparisons. Fig. 8 (c) and (d) show the density plots of $W_{V,\widetilde{C}_0}$ for $N_0 = 1, 000$ and $N_0 = 2, 000$. It can be seen that these two densities are quite close to each other. Moreover, Fig. 8 (b) shows that the density plot of $W_{V,\widetilde{C}_0}$ for $N_0 = 1, 000$ is close to its Chi-squared approximation. Subsequently, we calculated the corrected *p*–values of $W(\tilde{c}, v)$. Figures 7 (e) and (f) show the number of significant voxel-locus pairs based on the corrected *p*–values of $W(\tilde{c}, v)$ against the top $N_0 = 1, 000$ SNPs at the 0.5 and 0.8 significance, respectively. Table 5 includes 3 selected SNPs, who have the 3 largest numbers of significant voxel-locus pairs.

In Step (III.2), we set $\alpha_I = 0.005$ and then calculated all possible clusters and their associated *p*-values for the top N_0 SNPs in order to detect significant voxel-cluster pairs. Fig. 9 (a) and (b) show the density plots of $N\left(\widetilde{C}_0, \alpha_I = 0.005\right)$ for $N_0 = 1, 000$ and $N_0 = 2, 000$, respectively. Fig. 9 (c) and (d) show the numbers of significant voxel-locus pairs based on the corrected *p*–values of all clusters corresponding to the top $N_0 = 1, 000$ and $N_0 = 2, 000$ SNPs. Table 5 includes 3 selected SNPs, who have the 3 largest numbers of significant cluster-locus pairs.

Figure 10 shows some selected slices maps of $-\log_{10}(p)$ values for significant clusters corresponding to a selected SNP (rs2480271). Inspecting significant clusters in Figure 10 shows symmetric clustering across the left and right hemispheres. Since brain structures are highly symmetric between hemispheres, at least for most brain regions, it may be biologically plausible to observe symmetric associations for the SNPs and clusters. Several major clusters include major ROIs including superior temporal gyrus, inferior temporal gyrus, anterior cingulate gyrus, hippocampus, putamen, and fusiform. Among them, the superior temporal gyrus is an essential structure involved in auditory processing, in social cognition processes, as well as in the function of language. The inferior temporal gyrus is one of the higher levels of the ventral stream of visual processing. The anterior cingulate gyrus is involved in rational cognitive functions, such as reward anticipation, decision-making, empathy, impulse control, and emotion. The hippocampus is known to be

associated with memory and cognition. The fusiform is associated with color recognition, word and body recognition and the putamen is associated with motor skills.

We used the Matrix eQTL to carry out VGWAS by calculating the raw p–values of standard *t* statistics based on the normal linear model across all voxel-locus pairs. We selected those voxel-locus pairs, whose raw *p*–values are smaller than 10^{-7}, and then calculated their corresponding FWE corrected *p*–values by using the first wild bootstrap method described in Steps (III.1.1)-(III.1.5). Fig. 11 (a) shows the raw – $\log_{10}(p)$-values of all selected voxel-locus pairs corresponding to our method and the standard *t* test. It can be seen that the – $\log_{10}(p)$-values of our method are approximately proportional to those of the standard method, indicating an agreement between our proposed method and the standard method. However, for all selected pairs, the – $\log_{10}(p)$-value of our method is smaller than that of the standard *t* test. It may indicate that some of key assumptions (e.g., homogeneous variance) for the normal linear model is problematic in these voxels.

Similar the FVGWAS results presented in Fig. 7 and Table 5, we used the FWE corrected *p*–value 0.5 as the significant level to determine important voxel-locus pairs obtained from the Matrix eQTL results. Note that each SNP may have multiple voxels with their raw *p*–values smaller than 10^{-7}. Fig. 11 (b) shows the number of significant voxel-locus pairs corresponding to all unique SNPs obtained from the Matrix eQTL results at the 0.5 significant level. Then, for each voxel-locus pair, whose FWE corrected *p*-value is smaller than 0.5, we took its $3 \times 3 \times 3$ neighborhood in the image space and then calculated the percentage of neighboring voxels, whose raw *p*–values were smaller than 10^{-5}. We found that such percentage of neighboring voxels is 0 for all voxel-locus pairs, which indicates that these significant voxels are isolated in the image space. Such isolated voxel-locus pairs may be biologically meaningless.

Subsequently, we selected all SNPs with more than 20 significant voxel-locus pairs based on the Matrix eQTL results in order to detect the significant cluster-locus pairs. For each of such SNPs, we used the same setting for the cluster-locus pairs used for FVGWAS. Figs. 11 (c) and (d) show the maximum cluster size for each SNP and its corresponding corrected *p*–value. From Fig. 11 (d), two significant cluster-locus pairs are detected at the 0.5 significant level. These two SNPs are rs11815438 and rs7001339, which are included in our detected significant cluster-locus pairs listed in Table 5. This result demonstrates that our proposed method may be able to identify important significant cluster-locus pairs.

4. Conclusion and Discussions

We have developed a FVGWAS pipeline for efficiently carrying out whole-genome analyses of whole-brain data. Our FVGWAS consists of a heteroscedastic linear model, a global sure independence screening (GSIS) procedure, and a detection procedure based on wild bootstrap methods. Two key advantages of using FVGWAS include a much smaller computational complexity $O((N_C + N_V)n^2)$ for FVGWAS compared with $O(nN_VN_C)$ for VGWAS and GSIS for screening many noisy SNPs. We have used simulations to evaluate the finite sample performance of each component of FVGWAS. Finally, we have successfully applied FVGWAS to imaging genetic data of ADNI study. Our FVGWAS may

be a valuable statistical toolbox for fast large-scale imaging genetic analysis particularly when the field is rapidly advancing with ultra-high-resolution imaging and whole-genome sequencing.

Many important issues need to be addressed in future research. First, the heteroscedastic linear model in FVGWAS is a standard voxel-wise method. However, as discussed in (Li et al., 2011; Polzehl et al., 2010), the voxel-wise methods are not optimal since they ignore a spatial feature of imaging data. Imaging data are spatially correlated in nature and contain spatially contiguous regions with rather sharp edges due to the inherent biological structure and function of objects. Such spatial information can be important for accurate estimation and prediction. Although it is common to use Gaussian smoothing with a prefixed bandwidth, it may introduce substantial bias in the statistical results (Li et al., 2012, 2013). Although several multi-scale adaptive regression models (MARMs) have been developed for the group analysis of imaging data (Li et al., 2011; Skup et al., 2012; Li et al., 2012; Polzehl et al., 2010), these methods are not computationally feasible even for thousands of SNPs. It is critically important to develop some novel methods to explicitly incorporate the spatial feature of the imaging data in FVGWAS, while achieving computational efficiency for ultra-high-resolution imaging and whole-genome sequencing.

Second, the effectiveness of GSIS strongly depends on the behavior of the global Wald-type statistics $\{W(c)\}_c$. Although, as shown in simulations, GSIS can perform reasonably well for moderate and strong signals, we expect that it can su er some difficulties in the detection of weak genetic effects on ROIs. We may consider two strategies. One is to explicitly incorporate the spatial feature of the image data as discussed above. The other is to propose other global statistics to increase the power of detecting weak genetic effects on ROIs (Zhang et al., 2014; Chen and Qin, 2010; Sun et al., 2015).

Third, FVGWAS is still a single SNP analysis framework (Hibar et al., 2011; Shen et al., 2010). However, it is well known that the power of genome-wide association studies (GWAS) for mapping complex traits with single SNP analysis may be undermined by modest SNP effect sizes, unobserved causal SNPs, correlation among adjacent SNPs, and SNP-SNP interactions (Tzeng 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 (Schaid et al., 2002; Vounou et al., 2010; Ge et al., 2012a; Thompson et al., 2013). Therefore, it is definitely interesting and important to extend FVGWAS to carry out marker-set and whole-brain association mapping. We expect many challenging issues arising from such development.

Acknowledgement

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers 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. Ho mann-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). 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.

References

Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of ld and haplotype maps. Bioinformatics. 2005; 21(2):263–265. [PubMed: 15297300]

Braskie MN, Kohannim O, Jahanshad N, Chiang MC, Barysheva M, Toga AW, Ringman JM, Montgomery GW, McMahon KL, de Z. Greig I. et al. Relation between variants in the neurotrophin receptor gene, ntrk3, and white matter integrity in healthy young adults. NeuroImage. 2013; 82:146–153. [PubMed: 23727532]

Braskie, Meredith N.; Jahanshad, Neda; Stein, Jason L.; Barysheva, Marina; McMahon, Katie L.; de Zubicaray, Greig I.; Martin, Nicholas G.; Wright, Margaret J.; Ringman, John M.; Toga, Arthur W., et al. Common alzheimer's disease risk variant within the clu gene affects white matter microstructure in young adults. The Journal of Neuroscience. 2011; 31(18):6764–6770. [PubMed: 21543606]

Chen SX, Qin YL. A two-sample test for high-dimensional data with applications to gene-set testing. Annals of Statistics. 2010; 38:808–835.

Davatzikos C, Genc A, Xu D, Resnick SM. Voxel-based morphometry using the ravens maps: Methods and validation using simulated longitudinal atrophy. NeuroImage. 2001; 14:1361–1369. [PubMed: 11707092]

Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Statist. Sci. 2003; 18:71–103.

Durston S. Imaging genetics in adhd. NeuroImage. 2010; 53:832–838. [PubMed: 20206707]

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70(5):849–911. [PubMed: 19603084]

Fan J, Song R. Sure independence screening in generalized linear models with np-dimensionality. The Annals of Statistics. 2010; 38(6):3567–3604.

Gabriel SB. The structure of haplotype blocks in the human genome. Science. 2002; 296(5576):2225–2229. [PubMed: 12029063]

Ge T, Feng J, Hibar DP, Thompson PM, Nichols TE. Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures. NeuroImage. 2012a; 63:858–873. [PubMed: 22800732]

Ge T, Feng J, Hibar DP, Thompson PM, Nichols TE. Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. NeuroImage. 2012b; 63(2):858–873. [PubMed: 22800732]

Gilmore JH, Schmitt JE, Knickmeyer RA, Smithm JK, Lin W, Styner M, Gerig G, Neale MC. Genetic and environmental contributions to neonatal brain structure: a twin study. Human Brain Mapping. 2010; 31:1174–1182. [PubMed: 20063301]

Golanska E, Hulas-Bigoszewska K, Sieruta M, Zawlik I, Witusik M, Gresner SM, Sobow T, Styczynska M, Peplonska B, Barcikowska M, et al. Earlier onset of alzheimer's disease: risk polymorphisms within prnp, prnd, cyp46, and apoe genes. Journal of Alzheimer's Disease. 2009; 17(2):359–368.

Gopalraj RK, Zhu H, Kelly JF, Mendiondo M, Pulliam JF, Bennett DA, Estus S. Genetic association of low density lipoprotein receptor and alzheimer's disease. Neurobiology of aging. 2005; 26(1):1–7. [PubMed: 15585340]

Hayasaka S, Nichols T. Validating cluster size inference: random field and permutation methods. NeuroImage. 2003; 20

Hayasaka S, Phan LK, Liberzon I, Worsley KJ, Nichols T. Nonstationary cluster-size inference with random field and permutation methods. NeuroImage. 2004; 22:676–687. [PubMed: 15193596]

He QC, Lin DY. A variable selection method for genome-wide association studies. Bioinformatics. 2011; 27(1):1–8. [PubMed: 21036813]

Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, Kim S, Pankratz N, Foroud T, Huentelman MJ, Potkin SG, Jack CR, Weiner MW, Toga AW, Thompson PM, ADNI. Voxelwise gene-wide association study (vgenewas): multivariate gene-based association testing in 731 elderly subjects. Neuroimage. 2011; 56:1875–1891. [PubMed: 21497199]

Knickmeyer RC, Wang JP, Zhu HT, Geng X, Woolson S, Hamer RM, Konneker T, Lin WL, Styner M, Gilmore JH. Common variants in psychiatric risk genes predict brain structure at birth. Cerebra Cortex. 2014; 24:1230–1246.

Li Y, Zhu H, Shen D, Lin W, Gilmore JH, Ibrahim JG. Multiscale adaptive regression models for neuroimaging data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011; 73:559–578. [PubMed: 21860598]

Li Y, Gilmore JH, Shen D, Styner M, Lin W, Zhu HT. Multiscale adaptive generalized estimating equations for longitudinal neuroimaging data. NeuroImage. 2013; 72:91–105. [PubMed: 23357075]

Li YM, Gilmore JH, Wang JP, Styner M, Lin WL, Zhu HT. Two-stage multiscale adaptive regression methods of twin neuroimaging data. IEEE Transactions on Medical Imaging. 2012; 31:1100–1112. [PubMed: 22287236]

Liu JY, Calhoun VD. A review of multivariate analyses in imaging genetics. Frontiers in neuroinformatics. 2014; 8

Medland SE, Jahanshad N, Neale BM, Thompson PM. Whole-genome analyses of whole-brain data: working within an expanded search space. Nature neuroscience. 2014; 17(6):791–800. [PubMed: 24866045]

Miyashita A, Arai H, Asada T, Imagawa M, Matsubara E, Shoji M, Higuchi S, Urakami K, Kakita A, Takahashi H, et al. Genetic association of ctnna3 with late-onset alzheimer's disease in females. Human molecular genetics. 2007; 16(23):2854–2869. [PubMed: 17761686]

Polzehl J, Voss HU, Tabelow K. Structural adaptive segmentation for statistical parametric mapping. NeuroImage. 2010; 52:515–523. [PubMed: 20420928]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, Bakker PID, Daly MJ, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics. 2007; 81(3):559–575. [PubMed: 17701901]

Salimi-Khorshidi G, Smith SM, Nichols TE. Adjusting the effect of nonstationarity in cluster-based and tfce inference. Neuroimage. 2011; 54(3):2006–2019. [PubMed: 20955803]

Salmond CH, Ashburner J, Vargha-Khadem F, Connelly A, Gadian DG, Friston KJ. Distributional assumptions in voxel-based morphometry. NeuroImage. 2002; 17:1027–1030. [PubMed: 12377176]

Satterthwaite TD, Elliott MA, Ruparel K, Loughead J, Prabhakaran K, Calkins ME, Hopson R, Jackson C, Keefe J, Riley M, et al. Neuroimaging of the philadelphia neurodevelopmental cohort. Neuroimage. 2014; 86:544–553. [PubMed: 23921101]

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. The American Journal of Human Genetics. 2002; 70(2):425–434. [PubMed: 11791212]

Shabalin AA. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. Bioinformatics. 2012; 28(10):1353–1358. [PubMed: 22492648]

Shaffer JP. Multiple hypothesis testing. Annual review of psychology. 1995; 46(1):561–584.

Shen DG, Davatzikos C. Measuring temporal morphological changes robustly in brain mr images via 4-dimensional template warping. NeuroImage. 2004; 21(4):1508–1517. [PubMed: 15050575]

Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud TM, Pankratz ND, Moore JH, Sloan SD, Huentelman MJ, Craig DW, DeChairo BM, Potkin SG, Jack CR, Weiner MW, Saykin AJ, ADNI. 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. 2010; 53:1051–1063. [PubMed: 20100581]

Skup M, Zhu HT, Zhang HP. Multiscale adaptive marginal analysis of longitudinal neuroimaging data with time-varying covariates. Biometrics. 2012; 68:1083–1092. [PubMed: 22551084]

Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage. 2009; 44:83–98. [PubMed: 18501637]

Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, Saykin AJ, Shen L, Foroud T, Pankratz N, Huentelman MJ, Craig DW, Gerber JD, Allen AN, Corneveaux JJ, Dechairo BM, Potkin SG, Weiner MW, Thompson PM, ADNI. Voxelwise genome-wide association study (vgwas). Neuroimage. 2010; 53:1160–1174. [PubMed: 20171287]

Stelzer G, Dalah I, Stein TI, Satanower Y, Rosen N, Nativ N, Oz-Levi D, Olender T, Belinky F, Bahir I, et al. In-silico human genomics with genecards. Hum Genomics. 2011; 5(6):709–17. [PubMed: 22155609]

Sun Q, Zhu HT, Liu YF, Ibrahim JG. Sprem: sparse projection regression model for high-dimensional linear regression. Journal of American Statistical Association. 2015 in press.

Sun W. A statistical framework for eqtl mapping using rna-seq data. Biometrics. 2012; 68:1–11. [PubMed: 21838806]

Thompson PM, Ge T, Glahn DC, Jahanshad N, Nichols TE. Genetics of the connectome. NeuroImage. 2013; 80:475–488. [PubMed: 23707675]

Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, Toro R, Jahanshad N, Schumann G, Franke B, et al. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain imaging and behavior. 2014; 8(2):153–182. [PubMed: 24399358]

Tollervey JR, Wang Z, Hortobágyi T, Witten JT, Zarnack K, Kayikci M, Clark TA, Schweitzer AC, Rot G, Curk T, et al. Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. Genome research. 2011; 21(10):1572–1582. [PubMed: 21846794]

Tzeng JY, Zhang DW, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. The American Journal of Human Genetics. 2011; 89(2):277–288. [PubMed: 21835306]

Vounou M, Nichols TE, Montana G, the Alzheimer's Disease Neuroimaging Initiative. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. Neuroimage. 2010; 53:1147–1159. [PubMed: 20624472]

Vounou M, Janousova E, Wolz R, Stein JL, Thompson PM, Rueckert D, Montana G, ADNI. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in alzheimer's disease. Neuroimage. 2012; 60:700–716. [PubMed: 22209813]

Wang H, Nie F, Huang H, Kim S, Nho K, Risacher SL, Saykin AJ, Shen L, ADNI. Identifying quantitative trait loci via group-sparse multi-task regression and feature selection: An imaging genetics study of the adni cohort. Bioinformatics. 2012a; 28:229–237. [PubMed: 22155867]

Wang H, Nie F, Huang H, Risacher SL, Saykin AJ, Shen L, ADNI. I-dentifying disease sensitive and quantitative trait relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multi-modal multi-task learning. Bioinformatics. 2012b; 28:127–136. [PubMed: 22088842]

Wang, Y.; Nie, J.; Yap, PT.; Shi, F.; Guo, L.; Shen, D. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer; 2011. Robust deformable-surface-based skull-stripping for large-scale studies.; p. 635-642.

Worsley KJ, Taylor JE, Tomaiuolo F, Lerch J. Unified univariate and multivariate random field theory. NeuroImage. 2004; 23:189–195.

Zhang HP. Statistical analysis in genetic studies of mental illnesses. Statistical Science. 2011; 26:116–129. [PubMed: 21909187]

Zhang J, Chen J. Statistical inference for functional data. The Annals of Statistics. 2007; 35:1052–1079.

Zhang JT. Approximate and asymptotic distributions of chi-squared -type mixtures with applications. J. Amer. Statist. Assoc. 2005; 100:273–285.

Zhang YW, Xu ZY, Shen XT, Pan W. Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. NeuroImage. 2014; 96:309–325. [PubMed: 24704269]

Zhu H, Kong L, Li R, Styner M, Gerig G, Lin W, Gilmore JH. Fadtts: functional analysis of di usion tensor tract statistics. NeuroImage. 2011; 56:1412–1425. [PubMed: 21335092]

Zhu HT, Ibrahim JG, Tang N, Rowe DB, Hao X, Bansal R, Peterson BS. A statistical analysis of brain morphology using wild bootstrapping. IEEE Trans Med Imaging. 2007; 26:954–966. [PubMed: 17649909]

Zhu HT, Khondker ZH, and Lu ZS, Ibrahim JG. Journal of American Statistical Association. 2014; 109(507):977–990.

Highlights

- Develop a FVGAWS for adaptive analysis of large-scale imaging genetic data.
- An efficient global sure independence screening
- Develop companion software for FVGWAS

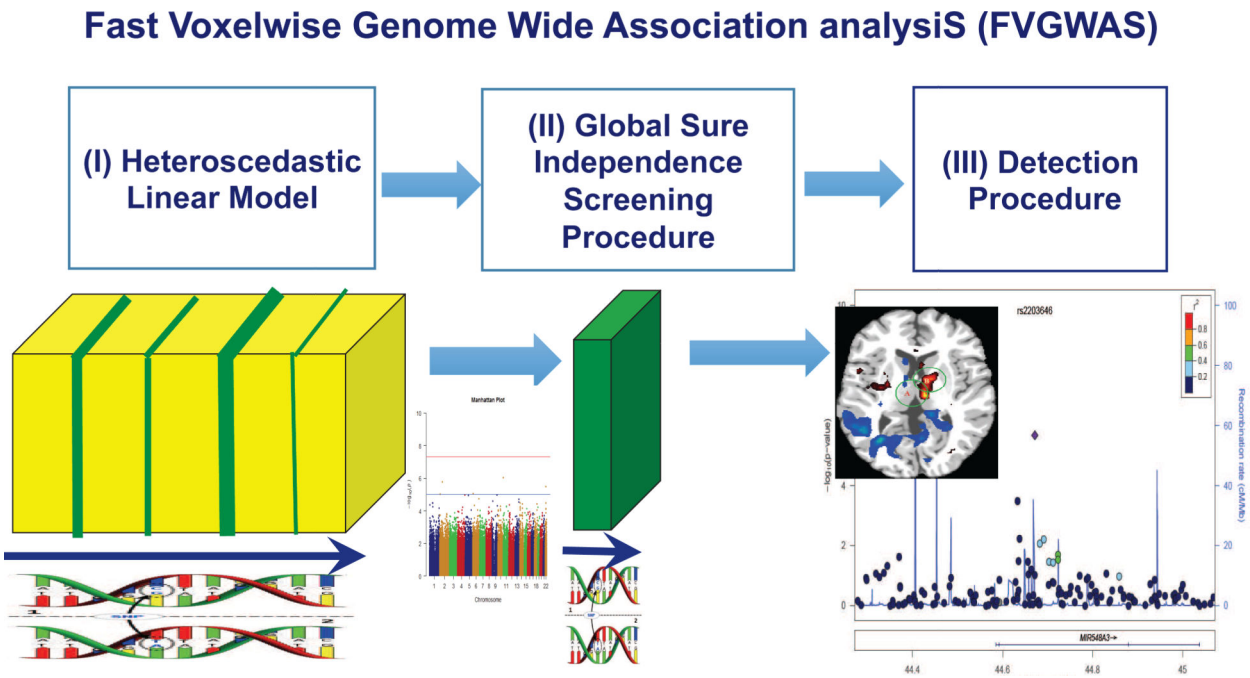

Fig. 1.
Schematic overview of FVGWAS

Fig. 2.

Simulation settings: the dark, gray, and white regions in each panel, respectively, represent background, brain region, and the effected ROI associated with the causal SNPs. From the left to the right, the sizes of the effected ROI are, respectively, set as 5 × 5, 10 × 10, and 20 × 20.

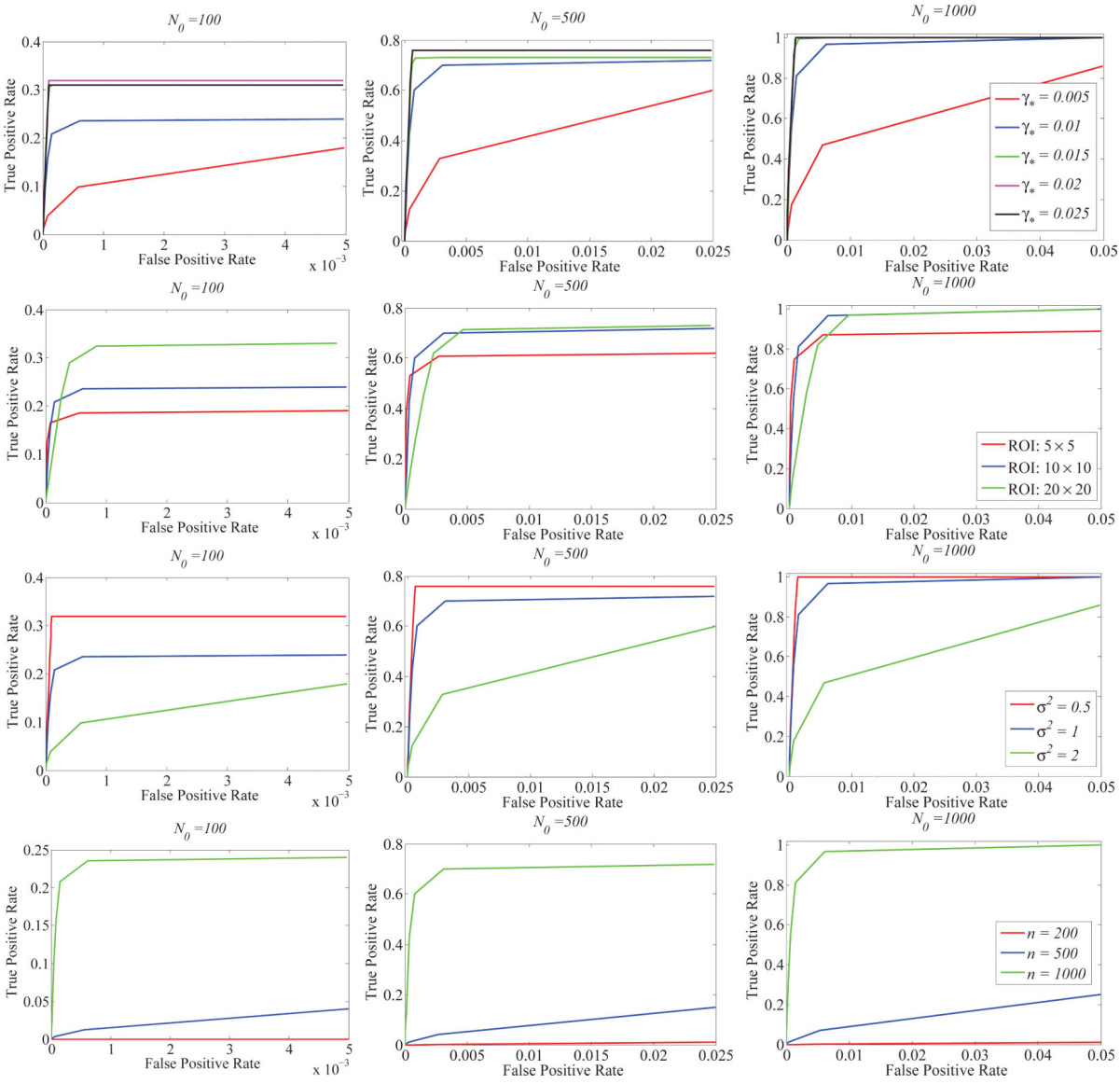

Fig. 3.

Simulation results for the association between SNPs and voxels: the first row contains ROC curves with varying γ_* values (corresponding to the causal SNPs' effect magnitude) and the number of the top N_0 SNPs included in the selection procedure. Parameters r, σ^2, and n are set to 10, 1, and 1000, respectively. The second row contains ROC curves with different ROIs. Parameters γ_*, σ^2, and n are set to 0.01, 1, and 1000, respectively. The third row contains ROC curves with varying σ. Parameters γ_*, r, and n are set to 0.01, 10, and 1000, respectively. The fourth row contains ROC curves with varying n. Parameters γ_*, σ^2, and r are set to 0.01, 1, and 10, respectively.

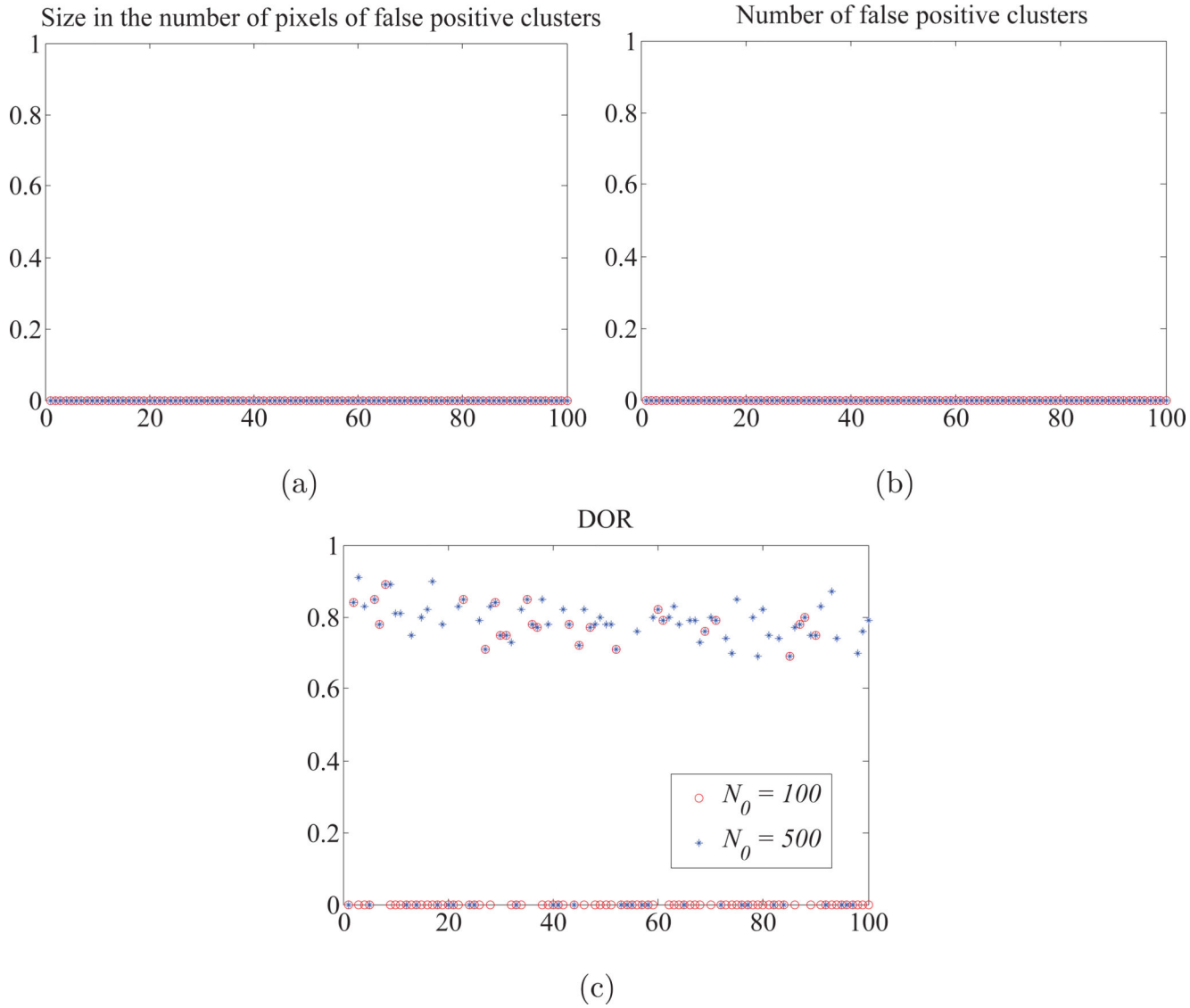

Fig. 4.

Simulation results for the association between SNPs and clusters: (a) the size in the number of pixels of false positive clusters in each causal SNP; (b) number of false positive clusters in each causal SNP; and (c) dice overlap ratio (DOR) in each causal SNP. Parameters γ_*, σ^2, n, and r are set to 0.01, 1, 1000, and 10, respectively.

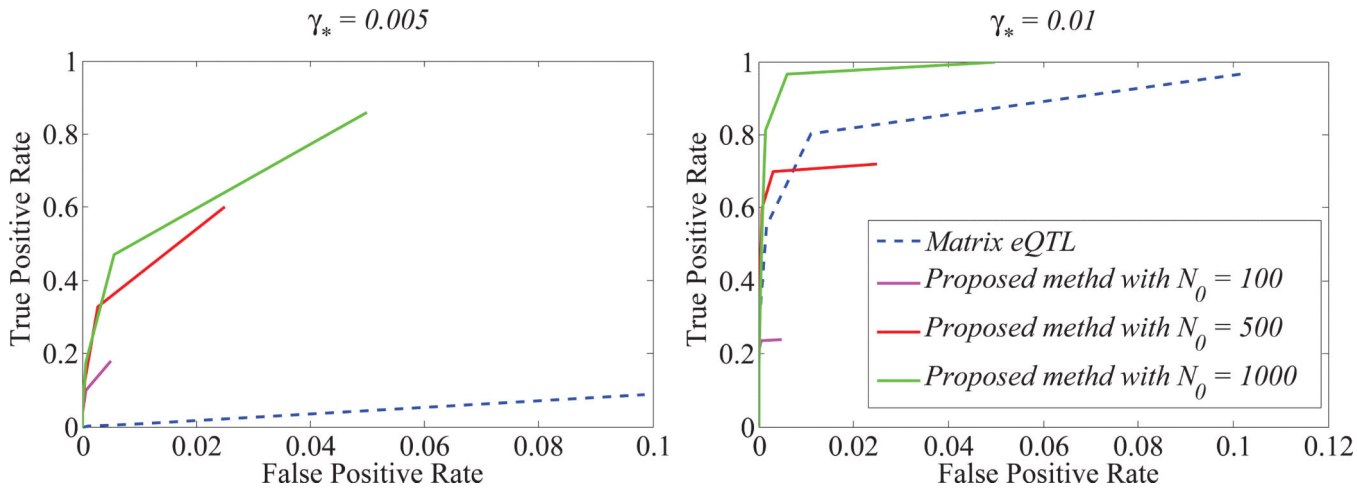

Fig. 5.

Simulation results for comparisons between FVGWAS and the Matrix eQTL in identifying significant voxel-locus pairs: ROC curves of the proposed method with N_0 = 100, 500, and 1, 000, and the Matrix eQTL method at γ_* = 0.005 and γ_* = 0.01. Parameters σ^2, n, and r are set to 1, 1000, and 10, respectively.

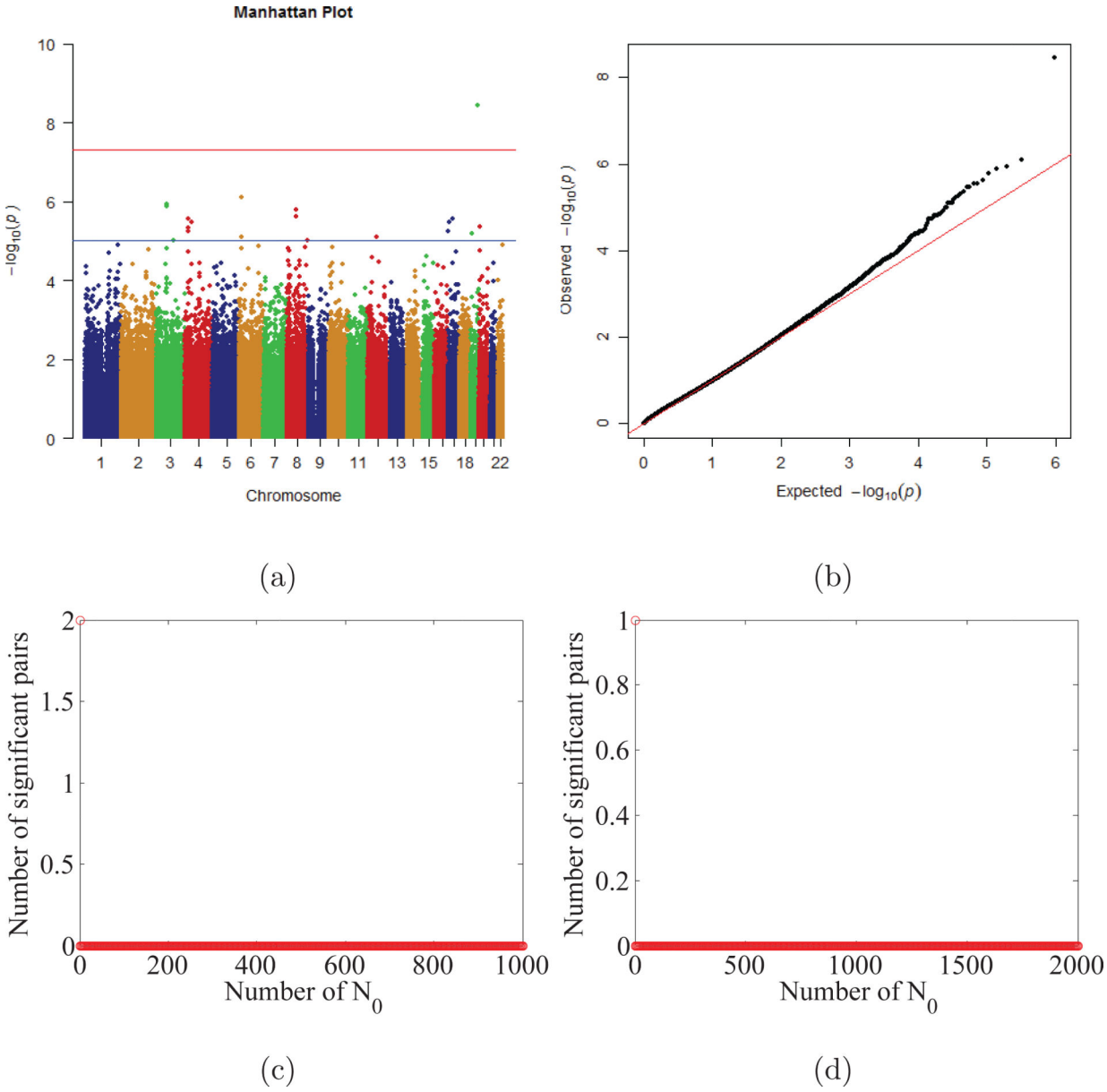

Fig. 6.

ADNI ROI volume GWAS: (a) Manhattan plot; (b) *QQ* plot; and the numbers of significant SNP-ROI pairs based on the corrected *p*–values of $W(c, v)$ at the 0.5 significance level corresponding to the top (c) N_0 = 1, 000 and (d) N_0 = 2, 000 SNPs;

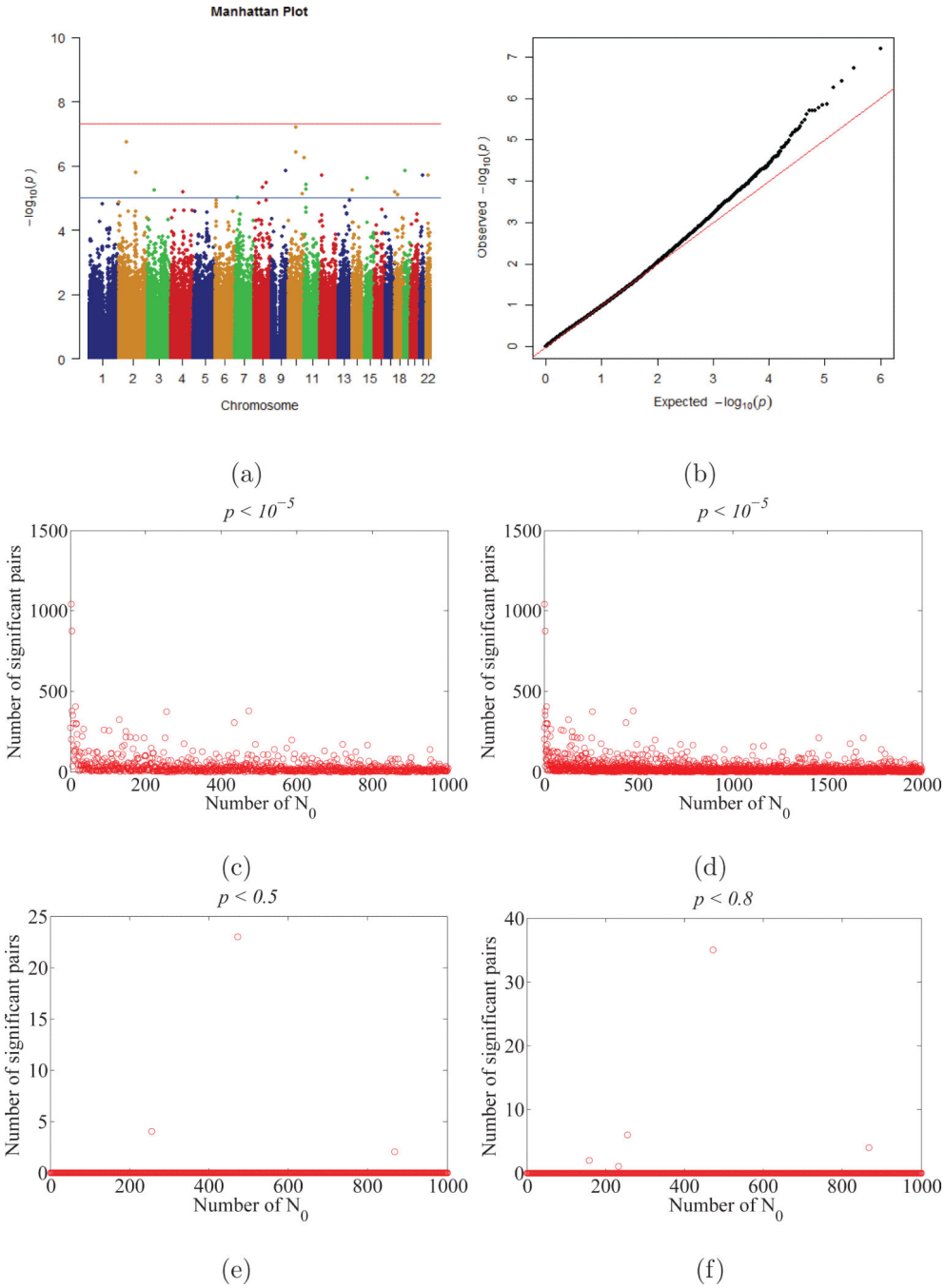

Fig. 7.

ADNI whole-brain GWAS: (a) Manhattan plot; (b) *QQ* plot; the numbers of significant voxel-locus pairs based on the raw *p*–values of $W(c, v)$ at the 10^{-5} significance level corresponding to the top (c) $N_0 = 1,000$ and (d) $N_0 = 2,000$ SNPs; the numbers of significant voxel-locus pairs based on the corrected *p*–values of $W(c, v)$ at the (e) 0.5 or (f) 0.8 significance level corresponding to the top $N_0 = 1,000$ SNPs.

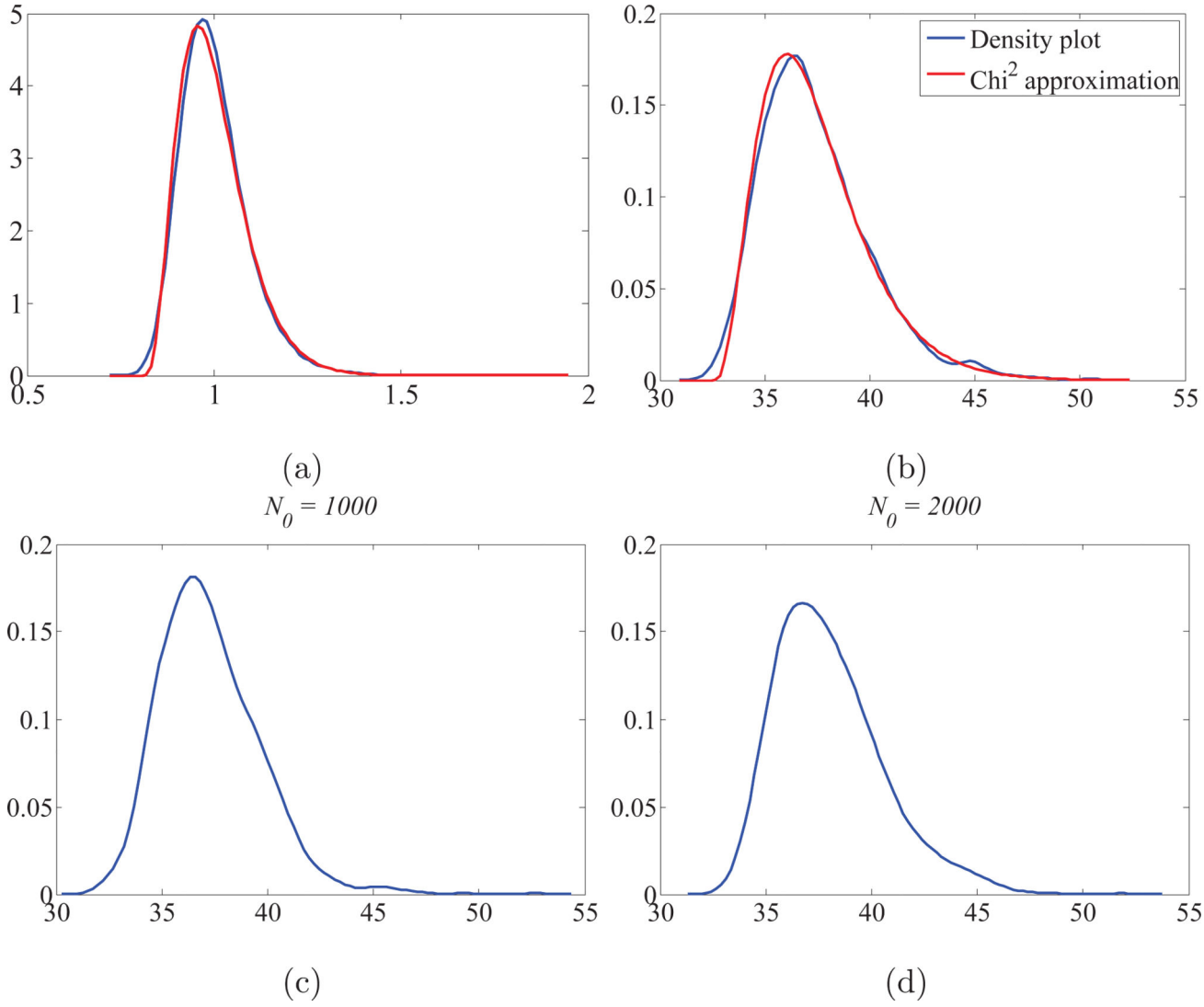

Fig. 8.

ADNI whole-brain GWAS: (a) the density plot of $W(c)$ and its χ^2 approximation; (b) the density plot of $W_{V,\widetilde{C}_0}$ and its χ^2 approximation; the density plots of $W_{V,\widetilde{C}_0}$ for N_0 = 1, 000 (c) and N_0 = 2, 000 (d).

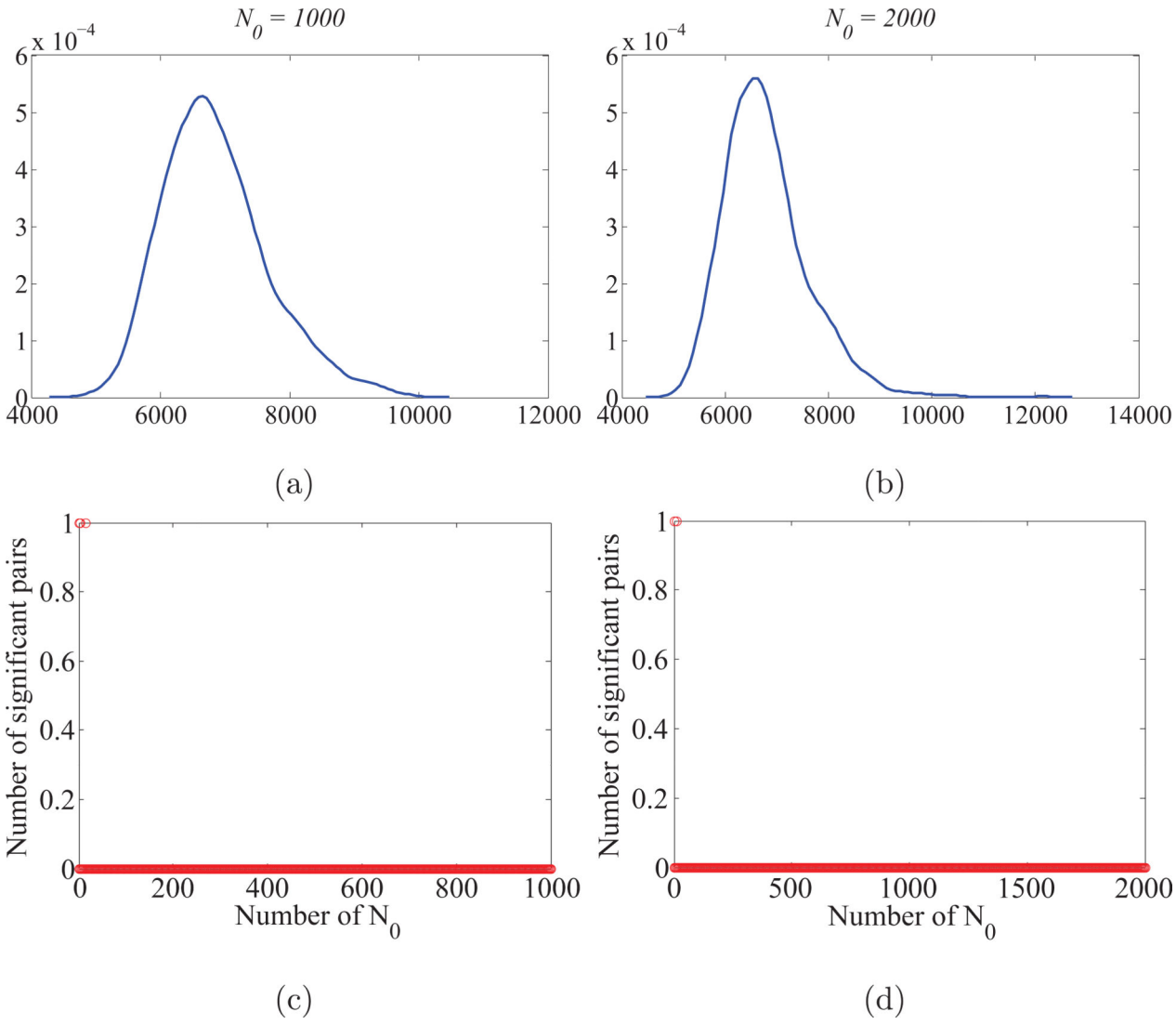

Fig. 9.

ADNI whole-brain GWAS: density plots of $N\left(\widetilde{C}_0, \alpha_I = 0.005\right)$ for N_0 = 1, 000 (a) and N_0 = 2, 000 (b); the numbers of significant voxel-locus pairs based on the corrected *p*–values of $W(c, v)$ at the 0.5 significance level corresponding to the top N_0 = 1, 000 (c) and N_0 = 2, 000 (d) SNPs.

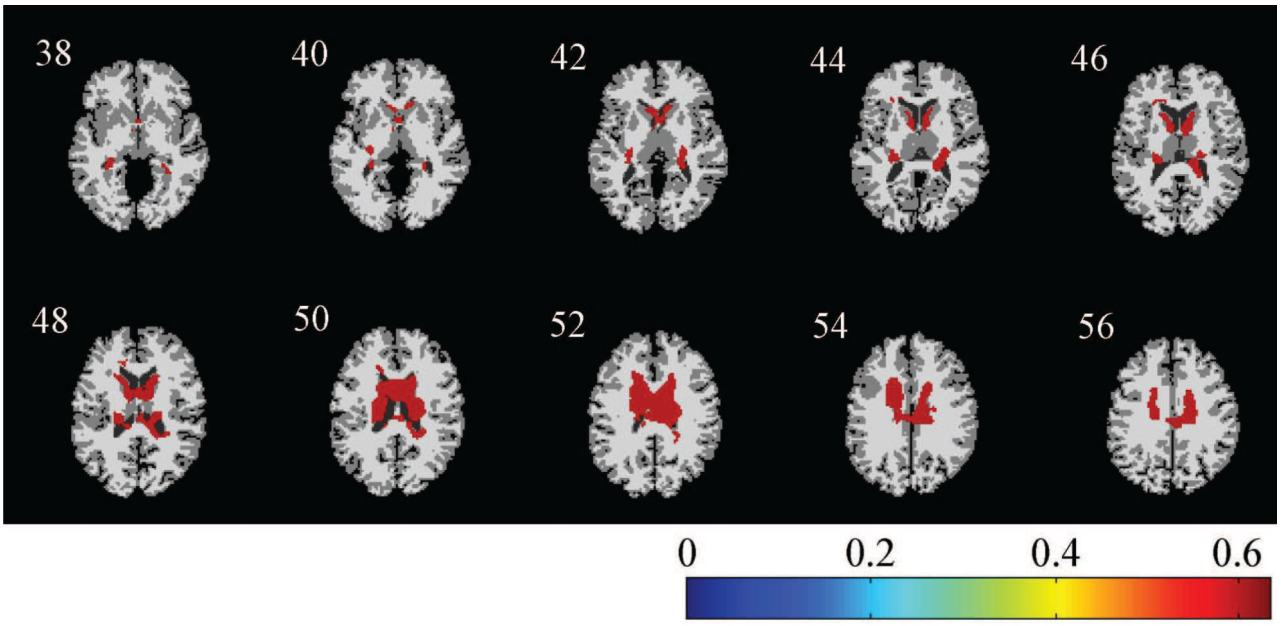

Fig. 10.

ADNI whole-brain GWAS: selected slices of $-\log_{10}(p)$ for significant clusters corresponding to a SNP (rs2480271).

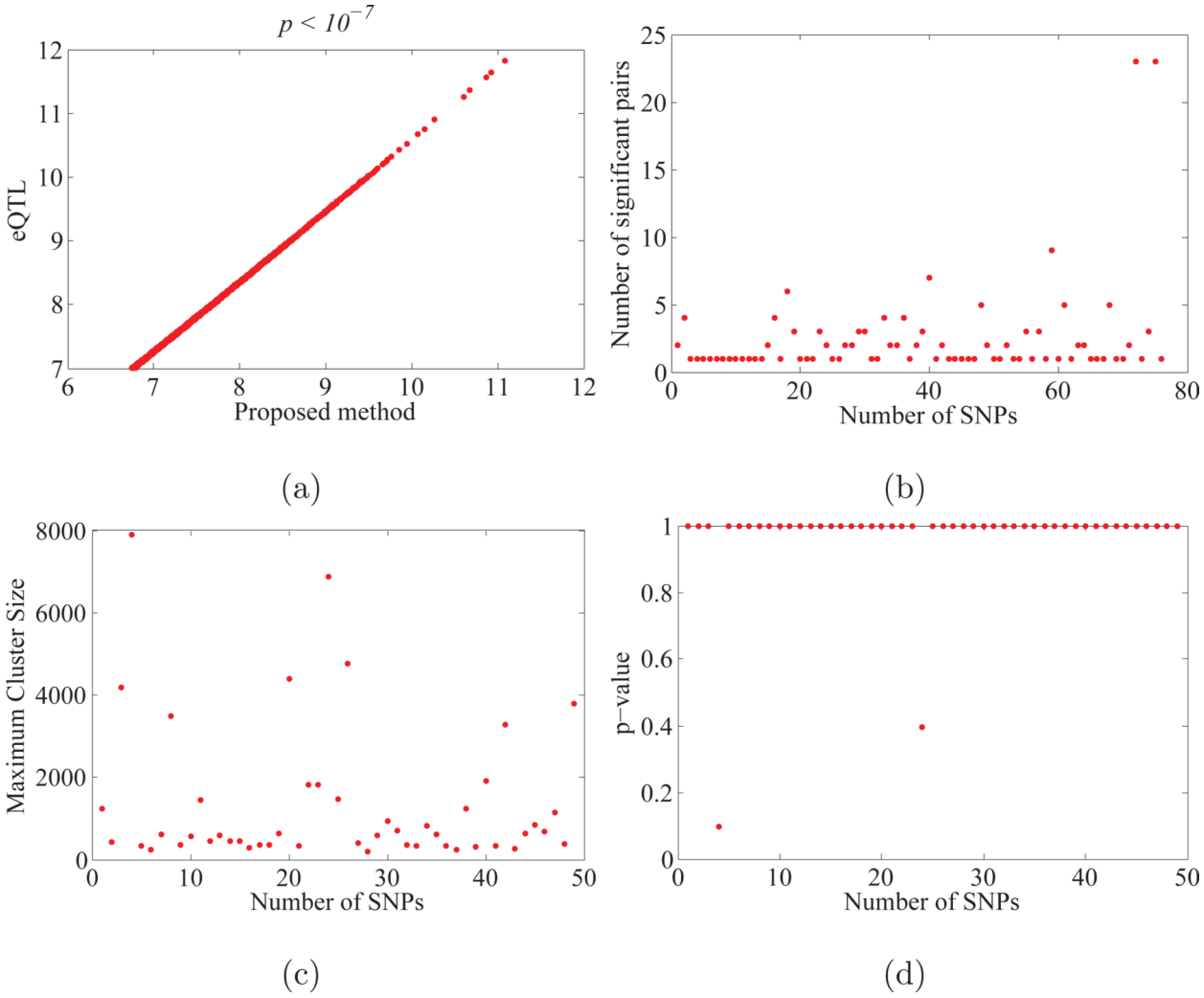

Fig. 11.

ADNI FVGWAS versus VGWAS: (a) raw – $\log_{10}(p)$-values of all selected voxel-locus pairs corresponding to our method and the standard *t* test; (b) number of significant voxel-locus pairs at the 0.5 significant level; (c) maximum cluster sizes of all selected SNPs obtained from the Matrix eQTL results; (d) the *p*–values of the maximum clusters corresponding to all selected SNPs.

Table 1

Simulation results: causal SNP rates correspond to different (N_0, γ_*) values in the effected ROI with size 10 × 10. The causal SNP rate is defined as the ratio of the number of causal SNPs in $\widetilde{\mathscr{C}_0}$ over the total number of causal SNPs.

	N_0														
γ_*	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**	**1200**	**1400**	**1600**	**1800**	**2000**
0.005	0.18	0.3	0.4	0.5	0.6	0.71	0.79	0.83	0.84	0.86	0.92	0.96	0.97	0.98	1
0.010	0.24	0.43	0.57	0.66	0.72	0.8	0.87	0.95	0.98	1	1	1	1	1	1
0.015	0.31	0.46	0.59	0.68	0.73	0.82	0.88	0.95	0.98	1	1	1	1	1	1
0.020	0.31	0.5	0.6	0.68	0.76	0.82	0.88	0.96	0.99	1	1	1	1	1	1
0.025	0.32	0.5	0.6	0.68	0.76	0.84	0.9	0.96	0.99	1	1	1	1	1	1

Table 2

Percentage of times of significant voxel and SNP pairs or cluster and SNP pairs at different thresholds (total times of significant pairs/repeat times).

	N_0	α				
		0.1	**0.2**	**0.3**	**0.4**	**0.5**
voxel and SNP pairs	500	0	0.04	0.12	0.24	0.74
	1000	0.06	0.1	0.17	0.48	0.52
cluster and SNP pairs	500	0	0.575	1	1	1
	1000	0	0	0.38	0.94	1

Table 3

ADNI ROI volume GWAS: selected top SNPs associated with volumes of HL/HR - Hippocampus left/right and AL/AR - Amygdala left/right.

ROI	Best SNP	CHR	BP	*p*–value	Gene
HL	rs2075650	19	45395619	1.4E-07	TOMM40
	rs6896317	5	142949513	5.5E-05	TRIO
	rs439401	19	45414451	7.6E-04	APOE
HR	rs2075650	19	45395619	2.7E-07	TOMM40
	rs6896317	5	142949513	5.5E-05	TRIO
	rs439401	19	45414451	1.2E-03	APOE
AL	rs2075650	19	45395619	1.5E-05	TOMM40
	rs6896317	5	142949513	5.8E-05	TRIO
	rs405509	19	45408836	1.4E-03	APOE
AR	rs2075650	19	45395619	1.4E-08	TOMM40
	rs6896317	5	142949513	4.7E-07	TRIO
	rs405509	19	45408836	1.1E-03	APOE

Table 4

ADNI whole-brain GWAS: selected top 30 SNPs associated with the whole brain

SNP	CHR	BP	*p*–value	SNP	CHR	BP	*p*–value
rs11815438	10	62501737	6.5E-08	rs17182599	14	22051519	5.8E-06
rs11891634	2	65926939	1.9E-07	rs11717277	3	54220871	5.9E-06
rs1060373	10	62554500	3.8E-07	rs971752	4	103224534	6.5E-06
rs2480271	10	132061197	5.6E-07	rs11872654	18	2164155	6.7E-06
rs10402592	19	11256887	1.4E-06	rs2935713	10	123432188	7.6E-06
rs12001550	9	120672883	1.5E-06	rs4129156	18	25437752	8.1E-06
rs13419007	2	145043653	1.7E-06	rs10261484	7	22583326	1.0E-05
rs2834077	21	34422738	2.0E-06	rs522793	6	10802955	1.2E-05
rs9645752	12	12544266	2.0E-06	rs14067	13	114110660	1.2E-05
rs5994978	22	34988594	2.0E-06	rs2443568	8	99254045	1.2E-05
rs4924156	15	37688630	2.5E-06	rs1448575	2	6386393	1.4E-05
rs2514323	8	99236899	3.3E-06	rs2697880	8	37337905	1.5E-05
rs1852755	11	13996686	3.9E-06	rs9382934	6	14040480	1.5E-05
rs7001339	8	69855507	4.8E-06	rs472276	1	244112606	1.5E-05
rs1528663	11	13967222	5.4E-06	rs1767282	1	112357101	1.5E-05

Table 5

RAVENS map GWAS: significant voxel-locus pairs at the 0.5 significance level (left) and significant cluster-SNP pairs at the 0.5 significance level (right)

SNP	number of voxel-locus pairs	SNP	number of cluster-SNP pairs	max cluster size	*p*–value of the max cluster
rs2075650 (TOMM40)	23	rs11815438	1	7906	0.11
rs9490103	4	rs2480271	1	7365	0.23
rs2244634	2	rs7001339	1	6864	0.45