FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data

Meiyan Huang, Thomas Nichols, Chao Huang, Yang Yu, Zhaohua Lu, Rebecca C. Knickmeyer, Qianjin Feng, Hongtu Zhu, for the Alzheimer's Disease Neuroimaging Initiative

**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 (\(N_C \geq 12 \) million known variants) associations with signals at millions of locations (\(N_V \sim 10^6\) in the brain from thousands of subjects (\(n \sim 10^2\)). The aim of this paper is to develop a Fast Voxelwise Genome Wide Association Analysis (FVGWAS) framework to efficiently carry out whole-genome analyses of whole-brain data. FVGWAS consists of three components including a heteroscedastic linear model, a global sure independence screening (GSIS) procedure, and a detection procedure based on wild bootstrap methods. Specifically, for standard linear association, the computational complexity is \(O\left(nN_vN_C\right)\) for voxelwise genome wide association analysis (VGWAS) method compared with \(O\left((N_C+N_v)^2\right)\) for FVGWAS. Simulation studies show that FVGWAS is an efficient method of searching sparse signals in an extremely large search space, while controlling for the family-wise error rate. Finally, we have successfully applied 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 s for a single CPU. Our FVGWAS 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.

© 2015 Elsevier Inc. All rights reserved.

**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 affected voxels. Nevertheless, 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 h 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 ($N_{C} \sim 10^{6}$). 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 analysisS (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 $N_{C}N_{V}$, in which $N_{0} \ll N_{C}$. 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 $n \ll \min{(N_{C}, N_{V})}$, we have $O((N_{C} + N_{V})n^{2}) = O(nN_{V}N_{C}) \times \frac{nN_{C}^{-1} + nN_{V}^{-1}}{n}$, leading to a computational gain at the order of $O(\min{(N_{C}, N_{V})} / n)$. Such computational gain makes it possible to run VGVAS 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

---

**Fig. 1. Schematic overview of FVGWAS.**

---

**Fast Voxelwise Genome Wide Association analysisS (FVGWAS)**

(I) **Heteroscedastic Linear Model**

(II) **Global Sure Independence Screening Procedure**

(III) **Detection Procedure**
and computational efficiency of FVGWAS by using simulation studies and a real data analysis. In Section 4, some conclusions and discussions are provided.

Method

Suppose we observe a set of imaging measurements, clinical variables, and genetic markers from n unrelated subjects. Let \( V \) be a selected brain region with \( N_v \) voxels and \( v \) be a voxel in \( V \). Let \( C \) be the set of \( N_c \) SNPs and \( c \) be a locus in \( C \). For each individual \( i (i = 1, \ldots, n) \), we observe an \( N_v \times 1 \) vector of imaging measurements, denoted by \( Y_i = (y_i(1), \ldots, y_i(N_v))^T \), a \( K \times 1 \) vector of clinical covariates \( Z_i = (x_{i1}, \ldots, x_{ik})^T \), and an \( L \times 1 \) vector \( Z_c(i) = (z_{ic}(1), \ldots, z_{ic}(L))^T \) 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 \( N_v N_c \) massive univariate analyses for all possible combinations of \((c, v)\), it is computationally challenging and intensive to compute all \( N_v N_c \) test statistics and to store and manage all \( N_v \) 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.

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 \( V \), \( y_i(v) \) can be modeled as a heteroscedastic linear model given by

\[
y_i(v) = X_i^T \beta_i(v) + Z_c(i)^T \gamma(c, v) + e_i(v) \quad \text{for} \quad i = 1, \ldots, n, \tag{1}
\]

where \( \beta_i(v) = (\beta_{i1}(v), \ldots, \beta_{ik}(v))^T \) is a \( K \times 1 \) vector associated with non-genetic predictors, and \( \gamma(c, v) = (\gamma_{c1}(v), \ldots, \gamma_{cl}(v))^T \) 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 \( \text{Cov}(e_i(v) : v \in V) \) has zero mean and a heterogenous covariance structure, that is, \( \text{Cov}(e_i(v)) \) 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 \( x \)-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

\[
H_0(c, v) : \gamma(c, v) = 0 \quad \text{versus} \quad H_1(c, v) : \gamma(c, v) \neq 0 \quad \text{for each} \quad (c, v). \tag{2}
\]

We introduce the standard Wald-test statistic as follows. Let \( Y(v) = (y_{i1}(v), \ldots, y_{in}(v))^T \) and \( P_X = X(X'X)^{-1}X' \) be the projection matrix of model (1), where \( X = (x_{11}, \ldots, x_{kn}) \) is a \( K \times n \) matrix. Similar to Zhu et al. (2007), we calculate an ordinary least squares estimate of \( \gamma(c, v) \), denoted by \( \hat{\gamma}(c, v) \), given by

\[
\hat{\gamma}(c, v) = \left(Z_c^T(I_n - P_X)Z_c\right)^{-1}Z_c^T(I_n - P_X)Y(v). \tag{3}
\]

where \( I_n \) is an \( n \times n \) identity matrix, \( Z_c = (z_{c1}, \ldots, z_{cn}) \) is an \( L \times n \) matrix. Ignoring heteroscedasticity in model (1) leads to an approximation of \( \text{Cov}(\hat{\gamma}(c, v)) \) given by

\[
\text{Cov}(\hat{\gamma}(c, v)) \approx \sigma^2(c, v)\left(Z_c^T(I_n - P_X)Z_c\right)^{-1}, \tag{4}
\]

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

\[
W(c, v) = \gamma(c, v)^T\left[Cov(\gamma(c, v))\right]^{-1}\gamma(c, v) = \left[Z_c^T(I_n - P_X)\sigma^{-2}(c, v)Y(v)Y(v)^T(I_n - P_X)Z_c\right]. \tag{5}
\]

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 \( \sigma^2(c, v) \) 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 \( \{\sigma^2(c, v), W(c, v)\} \) is at the order of \( N_v N_c n^2 \). Therefore, it is almost impossible to run a voxel-wise genome-wide association analysis in a single CPU.

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

- (S1) Calculate \( \sigma^2(c, v) \) under the null hypothesis \( H_0(c, v) \) for each \( c \) and \( v \).
- (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 \( O(N_v N_c n) \) to \( O(N_v + N_c n^2) \).

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

\[
\hat{\sigma}^2(c, v) = Y(v)^T(I_n - P_X)Y(v)/(n-K). \tag{6}
\]

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

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,
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

\[
W(c) = N_c^{-1} \text{tr} \left\{ \left( Z^c I_n - P_c \right) Z^c \left( I_n - P_c \right) \left( \sum_{v} \hat{\alpha}_v (v)^2 Y(v) Y(v)^T \right) \left( I_n - P_c \right) Z^c \right\}
\]  

(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 \(c\), \(V\) can be decomposed as the union of a true genetic effect region, denoted by \(V_3(c)\), and a false genetic effect region, denoted by \(V_4(c)\), such that \(V = V_3(c) \cup V_4(c)\) and \(V_3(c) \cap V_4(c) = \emptyset\). If the volume of \(V_3(c)\) is relatively large and signals in \(V_3(c)\) 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 \(V_3(c)\). In contrast, observing large values of \(W(c, v)\) in an extremely small effect region can primarily be caused by various noisy components, such as stochastic noise, susceptibility artifacts, or misalignment, in imaging data.

The complexity of computing \([W(c)]\) is at the order of \((N_c + N_0)n^2\), since \(\sum_{v} \hat{\alpha}_v (v)^2 Y(v) Y(v)^T\) is independent of \(c\). In contrast, the complexity of computing \([W(c, v)]\) is at the order of \(N_c n^2\). Therefore, computing \([W(c)]\) is about \(N_c N_0/(N_c + N_0)\) times faster than computing all \([W(c, v)]\).

Our GSIS consists of the following steps:

- Step (II.1). Calculate \(\Sigma_1 = (X'X)^{-1}\) with the computational complexity of \(O(nK^2)\).
- Step (II.2). Calculate \(\Sigma_2 = (I_n - P_k) \left( \sum_{v} \hat{\alpha}_v (v)^2 Y(v) Y(v)^T \right) (I_n - P_k)\) with the computational complexity of \(O(N_k n^2)\).
- Step (II.3). For the \(c\)-th locus, we do:
  - Calculate \(Z^c Z^c\) with the computational complexity of \(O(L^2 n)\).
  - Calculate \(Z'X\) with the computational complexity of \(O(1K n)\).
  - Calculate \(Z^c (I_n - P_c) Z^c = (Z^c Z^c) \Sigma_1 (X'X)\) with the computational complexity of \(O(L^2 K^2)\).
  - Calculate \(Z^c Z^c\) with the computational complexity of \(O(L^2 n^2)\).
  - 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) and Zhang and Chen (2007), if \(y_i(v)\) are treated as functional responses, then \(W(c)\) asymptotically converges to a weighted \(\chi^2\) distribution as \(n \rightarrow \infty\) when \(H_0(c, v)\) holds for all \((c, v)\) pairs. Let \(K_3(W)\), \(K_2(W)\), and \(K_1(W)\) be, respectively, the first three cumulants of \(W(c)\). Therefore, following the reasonings in Zhang (2005), \(W(c)\) can be approximated by a \(\chi^2\)-type random variable \(\alpha_1 X^2(\alpha_2) + \alpha_3 X^2(\alpha_4)\), where \(\alpha_1, \alpha_2, \alpha_3, \alpha_4\) are, respectively, given by

\[
\alpha_1 = \frac{K_3(W)}{4K_2(W)}, \quad \alpha_2 = \frac{8K_3(W)}{K_2(W)} \quad \text{and} \quad \alpha_3 = K_1(W) \frac{2K_3(W)}{K_3(W)}
\]  

(8)

We approximate \(\{(\alpha_k, K_k(W))\}_{k=3}^{\infty}\) 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 \(P(X^2(\alpha) \leq W(c) - \alpha_3/\alpha_2)\). 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 \(c_0 = \{c_1, \ldots, c_{N_0}\}\). From now on, we call \(c_0\) as a candidate significant locus set.

There are some rationales of choosing \(W(c)\) to determine \(c_0\) and setting a relatively large \(N_0\) in GSIS. As shown in simulations, if the volume of \(V_3(c)\) is relatively large and signal strength in \(V_3(c)\) is moderate, then \(W(c)\) should put the \(c\)-th locus into \(c_0\). 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 \(c_0\) is relatively large. We will carry out simulations to evaluate such probability for different signal-to-noise ratios and sizes of \(V_n(c)\).

The accuracy of the \(\chi^2\)-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 \(P_k\) and \(\sum_{v} \hat{\alpha}_v (v)^2 Y(v) Y(v)^T\), \(W(c)\) is slightly different from each other only in term of \(Z_c\). 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 \(Z_c\) (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 \(K_4(W)\) 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.

FVGWAS (III): a detection procedure based on wild bootstrap methods

Our detection procedure consists of two wild bootstrap methods:

- Step (III.1). The first one is to detect significant voxel–locus pairs.
- Step (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 \(c_0\) with the top \(N_0\) loci, we calculate a maximum statistic over all voxels for the top \(N_0\) loci as

\[
W_{v,c_0} = \max_{v \in c_0} W(\tilde{c}, v).
\]  

(9)

The maximum statistic \(W_{v,c_0}\) 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 \(W_{v,c_0}\) under the null hypothesis \(H_0(c, v)\) holds for all \(c \in C\) and \(v \in V\).

We propose an efficient wild bootstrap procedure to detect significant \((\tilde{c}, v)\) pairs as follows:

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

\[
W(\tilde{c}, v) = \text{tr} \left\{ \left( Z^c (I_n - P_c) Z^c \right)^{-1} \left( I_n - P_c \right) \hat{\alpha}_v (v)^2 Y(v) Y(v)^T \right\}.
\]

The computational complexity is \(O(N_0 n N_0)\).

- Step (III.1.2). Calculate \(W_{v,c_0}\).
- Step (III.1.3). Apply the first wild bootstrap method to the set \(\tilde{c}_0\).
  - Fit a linear model \(y_i(v) = x_i' \beta(v) + e_i(v)\) to imaging data and calculate \(\hat{\beta}(v) = y_i(v) - x_i' \hat{\beta}(v)\) for all \(i\) and \(v\), where \(\hat{\beta}(v) = (\sum_{i=1}^{N} x_i x_i')^{-1} \sum_{i=1}^{N} x_i y_i(v)\). Generate \(G\) bootstrap samples \(y_i^{(g)}(v) = x_i' \beta(v) + \eta_i^{(g)} \hat{e}_i(v)\) 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 $\gamma(c, v) = 0$ for all $(c, v) \in C \times V$, 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 $(\tilde{y}^{(g)}(\tilde{e}(v)), \tilde{e}(v))_{i, s} \in \mathcal{V}$ and the residuals $(\tilde{y}^{(g)}(\tilde{e}(v)))_{i, s} \in \mathcal{V}$ given the raw imaging data is given by

$$n^{-1} \sum_{i=1}^{n} \text{Cov} \left( \tilde{y}^{(g)}(\tilde{e}(v)), \tilde{y}^{(g)}(\tilde{e}(v)) \right) = n^{-1} \sum_{i=1}^{n} \tilde{e}(v) \tilde{e}(v)^{T}.$$ 

It follows from the law of large number that $n^{-1} \sum_{i=1}^{n} \tilde{e}(v) \tilde{e}(v)^{T}$ converges to the spatial covariance between voxels $v$ and $v'$. Let $\tilde{y}^{(g)}(\tilde{e}(v), \tilde{e}(v'))$ be a vector of spatial covariances between voxels $v$ and $v'$ under the null hypotheses. Then, under the null hypotheses, $\tilde{y}^{(g)}(\tilde{e}(v), \tilde{e}(v'))$ is asymptotically normal with mean zero and variance $\sigma_{W}^{2}$. The family-wise error (FWE) corrected $p$-values of $\tilde{y}^{(g)}(\tilde{e}(v), \tilde{e}(v'))$ can be calculated using the bootstrap method at the voxel level. Second, it automatically accounts for multiple simulations in Zhu et al. (2007) for the evaluation of the wild bootstrap procedure for detection of significant clusters.

Step (III.1.5). Calculate the family-wise error (FWE) corrected $p$-values of $\tilde{y}^{(g)}(\tilde{e}(v), \tilde{e}(v'))$ given by

$$\text{tr} \left[ \left( \mathbf{Z}^{(g)}(\mathbf{I}_{m} - P_{x}) \tilde{E}(v)^{T} \right) \left( \mathbf{I}_{n} - P_{x} \right) \tilde{E}(v) \right] \left( \mathbf{Z}^{(g)}(\mathbf{I}_{m} - P_{x}) \tilde{E}(v)^{T} \right) \left( \mathbf{I}_{n} - P_{x} \right) \tilde{E}(v).$$

Step (III.1.4). Calculate approximated Chi-squared distributions based on the bootstrapped samples $\left\{ W(c, v)^{(g)} \right\}_{g=1, \ldots, G}$ and calculate the uncorrected $p$-values of $\tilde{y}^{(g)}(\tilde{e}(v), \tilde{e}(v'))$ across all $(c, v) \in \mathcal{C} \times \mathcal{V}$. The elements in $\mathcal{C}$ and $\mathcal{V}$ are given by

$$\mathcal{C} = \{0, \ldots, n_{C}\} \quad \text{and} \quad \mathcal{V} = \{0, \ldots, n_{V}\}.$$

Step (III.1.5). Calculate the family-wise error (FWE) corrected $p$-values of $\tilde{y}^{(g)}(\tilde{e}(v), \tilde{e}(v'))$ across all $(c, v) \in \mathcal{C} \times \mathcal{V}$ based on the empirical distribution of $\left\{ W(c, v)^{(g)} : g = 1, \ldots, G \right\}$. At a given significance level $\alpha$, we can detect significant $(c, v)$ pairs in $\mathcal{C} \times \mathcal{V}$. Since the number of pairs $N_{C}N_{V} \in \mathcal{C} \times \mathcal{V}$ is much larger than the sample size, we choose a significance level, say $\alpha = 0.5$.

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 homogeneity 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 $c_{o}$, 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 $\left\{ W(c, v)^{(g)} \right\}_{g=1, \ldots, G}$ were used directly). Third, since $S(c)$ is independent of $\tilde{y}^{(g)}$, it is computationally efficient to calculate $W(c, v)^{(g)}$ and sort them in order to compute $\mathcal{C}^{(g)}$.

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 inter-connected voxels greater than a given threshold, say $\alpha_{c} = 0.005$ 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 $N(c, \alpha_{c})$ be the largest cluster size at a given threshold $\alpha_{c}$ based on the $p$-values of $\left\{ W(c, v) : v \in \mathcal{V} \right\}$. To detect significant (locus, cluster) pairs, we consider a maximum cluster size statistic and its approximation as

$$N(c, \alpha_{c}) = \max_{c \in \mathcal{C}} N(c, \alpha_{c}) \approx N(\mathcal{C}_{0}, \alpha_{c}) = \max_{c \in \mathcal{C}_{0}} N(c, \alpha_{c}).$$

Given $\mathcal{C}_{0}$ and the definition of $W(c)$ (Eq. (7)), it is expected that $N(c, \alpha_{c})$ is very close to $N(\mathcal{C}_{0}, \alpha_{c})$ 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 $\alpha_{o}$, we use the wild bootstrap method in Step (III.1.3) to generate $\left\{ W(c, v)^{(g)} : c \in \mathcal{C}, v \in \mathcal{V}, g = 1, \ldots, G \right\}$ and calculate $N(\mathcal{C}_{0}, \alpha_{o})$ for each wild bootstrap sample. For computational efficiency, we suggest to directly compare $W(c, v)^{(g)}$ with the 100(1 $- \alpha_{o})$th percentile of the $F$ distribution in order to determine clusters at each locus $c$.

- **Step (III.2.2).** For each locus $c \in \mathcal{C}_{0}$, we calculate all possible clusters and their associated FWE-corrected $p$-values based on the empirical distribution of $\left\{ N(\mathcal{C}_{0}, \alpha_{o}) \right\}_{g=1, \ldots, G}$.

### 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 FVCWAS. All computations for these numerical examples were done in Matlab on a Dell C6100 server. The computation for FVCWAS 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++.

**Simulation studies**

We simulated imaging data at $N_{V} = 3355$ pixels in the brain region of a 128 $\times$ 128 image, which is a middle slice of a brain volume obtained from the public accessible data of the Alzheimer’s Disease Neuroimaging 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 effect of SNPs is additive and homogeneous such that $y_{i}(v)$ were generated from:

$$y_{i}(v) = \sum_{j=1}^{N_{c}} (\beta_{j}^{(c)} v_{i,j}) + \epsilon_{i}(v).$$

where $\epsilon_{i}(v) \sim N(0, \sigma^{2})$, $z_{i}(c)$ were simulated genetic data as described below, and $\beta_{j}^{(c)} (c_{j}, v_{i,j})$ 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 $\beta^{(c)}$ were set to be those of estimated $\beta^{(c)}$ by fitting model (1) without genetic covariates to real ADNI dataset in the real ADNI data analysis section. The elements in $\gamma(c, v)$ corresponding to the pre-specified pairs of causal SNPs and effected Regions Of Interest (ROI) were set to effect magnitude $\gamma$, zero otherwise. In addition, the effected ROI associated with the causal SNPs was pre-fixed as a $r \times r$ region (Fig. 2).

We simulated genetic data $z_{i}(c)$ 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 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, and thus we had $N_0 = 20,000$ 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 ($\sigma$), and the size of affected ROI ($r \times r$) to be 1000, 100, 1, and 10 $\times$ 10, respectively. 100 Monte Carlo realizations were used.

First, we evaluate the finite sample performance of the proposed GSIS for $\gamma = 0.005, 0.010, 0.015, 0.020$, and 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 $C_0$ over the total number of causal SNPs. Table 1 includes the causal SNP rates corresponding to different top $N_0$ SNPs and $\gamma$ values. As expected, the causal SNP rate increases as the number of top $N_0$ SNPs and $\gamma$ increase. However, the causal SNP rate is low for $\gamma = 0.005$, 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 different $N_0$ and $\gamma$ values. This result is primarily attributed to the GSIS procedure used in FVGWAS.

Second, we evaluate the finite sample performance of FVGWAS in the detection of the causal SNPs and voxels in the affected 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 $\gamma$ and $N_0$ values. As expected, large $\gamma$ values representing larger genetic effects lead to a higher probability of detecting the causal SNPs and their associated voxels in the affected 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 $\gamma$ values, whereas a larger $N_0$ can lead to a higher false positive rate. Then, we set $\gamma$ to be 0.010, which is a moderate signal, in order to investigate the effects of different ROI sizes, $\sigma$, 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 $\sigma$ 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 overlap with some pixels in the affected 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 affected 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 affected ROI. Thus, the higher DOR means the higher probability of detecting the affected 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 $\gamma$- to be either 0.005 or 0.01. Fig. 5 presents the ROC curves corresponding to different $N_0$ and $\gamma$ 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 different $N_0$ and $\gamma$ values. This result is primarily attributed to the GSIS procedure used in FVGWAS.

Fifth, we set $\gamma(c, v) = 0$ 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 $\alpha$ was varied from 0.1 to 0.5, and 1000 replications were used to estimate FWERs. For a fixed $\alpha$, if the FWER is smaller than $\alpha$, then the test is conservative, whereas if the FWER is greater than $\alpha$, then the test is anticonservative, or liberal (Hayasaka and Nichols, 2003). Moreover, $\alpha$ was set to 0.005. Table 2 lists the FWERs corresponding to different $N_0$ and $\alpha$ values. For detecting significant SNP and voxel pairs, the rejection rates of the proposed method are accurate with large $N_0$ and $\alpha$ values. Moreover, for relatively small $\alpha = 0.1$, no significant SNP and cluster pairs are detected.

**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 nonprofit organizations, as a $60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether serial magnetic

**Table 1**

<table>
<thead>
<tr>
<th>$N_0$</th>
<th>$\gamma$</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
<th>1200</th>
<th>1400</th>
<th>1600</th>
<th>1800</th>
<th>2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.005</td>
<td>0.18</td>
<td>0.3</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.71</td>
<td>0.79</td>
<td>0.83</td>
<td>0.84</td>
<td>0.86</td>
<td>0.92</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>0.010</td>
<td>0.24</td>
<td>0.43</td>
<td>0.57</td>
<td>0.66</td>
<td>0.72</td>
<td>0.8</td>
<td>0.87</td>
<td>0.95</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.015</td>
<td>0.31</td>
<td>0.46</td>
<td>0.59</td>
<td>0.68</td>
<td>0.73</td>
<td>0.82</td>
<td>0.88</td>
<td>0.95</td>
<td>0.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.020</td>
<td>0.31</td>
<td>0.5</td>
<td>0.6</td>
<td>0.68</td>
<td>0.76</td>
<td>0.84</td>
<td>0.9</td>
<td>0.96</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.025</td>
<td>0.32</td>
<td>0.5</td>
<td>0.6</td>
<td>0.68</td>
<td>0.76</td>
<td>0.84</td>
<td>0.9</td>
<td>0.96</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
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 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 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 (164 AD, 346 MCI and 198 healthy controls) from ADNI1 in this data analysis. These scans on 421 males and 287 females (age 75.61 ± 6.76 years) were performed on a 1.5 T MRI scanners with 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 mm³.

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: gray 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 component 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 N₀ = 1000, the total processing time was 203,645 s, of which 116 s was

![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⁎, σ², n, and r are set to 0.01, 1, 1000, and 10, respectively.](image4.png)

![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 N₀ = 100, 500, and 1000, and the Matrix eQTL method at γ = 0.005 and γ = 0.01. Parameters σ², n, and r are set to 1, 1000, and 10, respectively.](image5.png)
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.

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 $5 \times 10^{-8}$ 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. Figs. 6(c) and (d) show 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 1000 and 2000. In this case, we can only detect significant ROI–locus pairs by using Step (III.1). Specifically, we generated 1000 bootstrapped samples $W_g(c, v)$ for $g = 1, ..., G = 1000$ and then calculated the corrected $p$-values of $W(c, v)$ across all $(c, v) \in \tilde{C}_0 \times V$. By setting the 0.5 significance level, we are able to detect 2 and 1 significant ROI–locus pairs for $N_0 = 1000$ and 2000, 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.

---

### Table 2

<table>
<thead>
<tr>
<th>$N_0$</th>
<th>$\alpha$</th>
<th>Voxel and SNP pairs</th>
<th>Cluster and SNP pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.04</td>
<td>0.575</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.12</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.24</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.74</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.94</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### Table 3

<table>
<thead>
<tr>
<th>ROI</th>
<th>Best SNP</th>
<th>CHR</th>
<th>BP</th>
<th>$\alpha$</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>HL</td>
<td>rs2075650</td>
<td>19</td>
<td>45395619</td>
<td>1.4E-07</td>
<td>TOMM40</td>
</tr>
<tr>
<td>HR</td>
<td>rs2075650</td>
<td>19</td>
<td>45395619</td>
<td>2.7E-07</td>
<td>TOMM40</td>
</tr>
<tr>
<td>AL</td>
<td>rs2075650</td>
<td>19</td>
<td>45395619</td>
<td>1.5E-05</td>
<td>TOMM40</td>
</tr>
<tr>
<td>AR</td>
<td>rs2075650</td>
<td>19</td>
<td>45395619</td>
<td>1.4E-08</td>
<td>TOMM40</td>
</tr>
</tbody>
</table>

---

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) $N_0 = 1000$ and (d) $N_0 = 2000$ SNPs.
RAVENS maps

Figs. 7(a) and (b) show 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 Wildtype statistic and its Chi-squared approximation for the whole brain. These two curves are very close to each other, indicating the accuracy of the $\chi^2$ 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, rs10605973, rs2480271, and rs29357713) and 2 SNPs (rs11891634 and rs13419007) in chromosome 2. Moreover, among the top $N_0 = 1000$ SNPs, we detected several important SNPs including rs2480271 on gene GLRX3 (chr 10), rs1534446 on gene PCEF1 (chr6), rs12436472 on gene NOVA1 (chr14), rs611637520 on gene PRNP (chr 20), rs47466222 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 disease (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 1000 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 1000 or 2000 and then generated 1000 bootstrapped samples $\{W(c, v)(g)\}$ for $g = 1, \ldots, G = 1000$. By using $\chi^2$ approximation, we calculated the raw $p$-values of $W(c, v)$ across all $(c, v) \in C_0 \times V$. At the $10^{-5}$ significance level, Figs. 7(c) and (d) show the number of significant voxel–locus pairs based on the raw $p$-values of $W(c, v)$ against the top $N_0$ SNPs in $C_0$.

Second, we calculated the corrected $p$-values of $W(c, v)$ in order to detect significant voxel–locus pairs by correcting for multiple
comparisons. Figs. 8(c) and (d) show the density plots of $W_{V,C_0}$ for $N_0 = 1000$ and $N_0 = 2000$. It can be seen that these two densities are quite close to each other. Moreover, Fig. 8(b) shows that the density plot of $W_{V,C_0}$ for $N_0 = 1000$ is close to its Chi-squared approximation. Subsequently, we calculated the corrected $p$-values of $W_{V,C_0}(v)$. Figs. 7(e) and (f) show the number of significant voxel–locus pairs based on the corrected $p$-values of $W_{V,C_0}(v)$ against the top $N_0 = 1000$ 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 $\alpha = 0.005$ 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. Figs. 9(a) and (b) show the density plots of $N(C_0, \alpha = 0.005)$ for $N_0 = 1000$ and $N_0 = 2000$, respectively. Figs. 9(c) and (d) show the numbers of significant voxel-cluster pairs based on the corrected $p$-values of all clusters corresponding to the top $N_0 = 1000$ and $N_0 = 2000$ SNPs. Table 5 includes 3 selected SNPs, who have the 3 largest numbers of significant cluster–locus pairs.

Fig. 10 shows some selected slices maps of $-\log_{10}(p)$ values for significant clusters corresponding to a selected SNP (rs2480271). Inspecting significant clusters in Fig. 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.
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).

<table>
<thead>
<tr>
<th>SNP</th>
<th>Number of voxel-locus pairs</th>
<th>SNP</th>
<th>Number of cluster-SNP pairs</th>
<th>Max cluster size</th>
<th>$p$-Value of the max cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2075650</td>
<td>23</td>
<td>rs11815438</td>
<td>1</td>
<td>7906</td>
<td>0.11</td>
</tr>
<tr>
<td>(TOMM40)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs9490103</td>
<td>4</td>
<td>rs2480271</td>
<td>1</td>
<td>7365</td>
<td>0.23</td>
</tr>
<tr>
<td>rs2244634</td>
<td>2</td>
<td>rs7001339</td>
<td>1</td>
<td>6864</td>
<td>0.45</td>
</tr>
</tbody>
</table>

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.

**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 $O(nN_C)$ 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

![Fig. 9. ADNI whole-brain GWAS: density plots of $N(c, 0.005)$ for $N_C = 1000$ (a) and $N_C = 2000$ (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 $N_0 = 1000$ (c) and $N_0 = 2000$ (d) SNPs.](image-url)
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 \( W(c) \). Although, as shown in simulations, GSIS can perform reasonably well for moderate and strong signals, we expect that it can suffer 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.

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

![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.](image)
Acknowledgment

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institute of Aging 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; Aracron Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; Emulonnm; 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; Neurologics Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc. and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is grateful for 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


