OXFORD

Bioimage informatics

Regional imaging genetic enrichment analysis

Xiaohui Yao[1], Shan Cong[1], Jingwen Yan (ORCID) **[2], Shannon L. Risacher[3], Andrew J. Saykin[3], Jason H. Moore[1] and Li Shen** (ORCID) **[1,*], UK Brain Expression Consortium[†], for the Alzheimer's Disease Neuroimaging Initiative[‡]**

[1]Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA, [2]Department of BioHealth Informatics, Indiana University School of Informatics and Computing, Indiana University and [3]Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN 46202, USA

*To whom correspondence should be addressed.

[†]See Supplementary UKBEC Acknowledgements section.

[‡]See Supplementary ADNI Acknowledgements section.

Associate Editor: Robert Murphy

Abstract

Motivation: Brain imaging genetics aims to reveal genetic effects on brain phenotypes, where most studies examine phenotypes defined on anatomical or functional regions of interest (ROIs) given their biologically meaningful interpretation and modest dimensionality compared with voxelwise approaches. Typical ROI-level measures used in these studies are summary statistics from voxelwise measures in the region, without making full use of individual voxel signals.

Results: In this article, we propose a flexible and powerful framework for mining regional imaging genetic associations via voxelwise enrichment analysis, which embraces the collective effect of weak voxel-level signals and integrates brain anatomical annotation information. Our proposed method achieves three goals at the same time: (i) increase the statistical power by substantially reducing the burden of multiple comparison correction; (ii) employ brain annotation information to enable biologically meaningful interpretation and (iii) make full use of fine-grained voxelwise signals. We demonstrate our method on an imaging genetic analysis using data from the Alzheimer's Disease Neuroimaging Initiative, where we assess the collective regional genetic effects of voxelwise FDG-positron emission tomography measures between 116 ROIs and 565 373 single-nucleotide polymorphisms. Compared with traditional ROI-wise and voxelwise approaches, our method identified 2946 novel imaging genetic associations in addition to 33 ones overlapping with the two benchmark methods. In particular, two newly reported variants were further supported by transcriptome evidences from region-specific expression analysis. This demonstrates the promise of the proposed method as a flexible and powerful framework for exploring imaging genetic effects on the brain.

Availability and implementation: The R code and sample data are freely available at https://github.com/lshen/RIGEA.

Contact: li.shen@pennmedicine.upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Brain imaging genetics is an emerging research field investigating the influence of genetic variants on brain imaging phenotypes. It examines associations between genetic variants, such as single-nucleotide polymorphisms (SNPs) and quantitative traits (QTs) extracted from brain imaging data, to gain novel insights into phenotypic characteristics and molecular mechanisms of complex brain disorders. These imaging QTs (iQTs) are measured based on either single voxels (Stein *et al.*, 2010) or regions of interest (ROIs) (Risacher *et al.*, 2010; Shen *et al.*, 2010; Yao *et al.*, 2017b) in the brain. An ROI is a pre-defined brain area containing a cluster of voxels with the same anatomical or functional annotation. Of note, the number of ROIs (e.g. dozens to hundreds) is much smaller than the number of voxels (e.g. tens of thousands to even millions) in the brain. Thus, imaging genetics studies typically investigate ROI-level phenotypes due to (i) modest dimensionality compared with voxelwise approaches for

increased statistical power and (ii) structural or functional annotation of ROIs that facilitates biologically meaningful interpretation.

Existing ROI-level imaging genetics studies often employ univariate strategies to evaluate the associations between individual genetic variants and ROI-level iQTs which are sometimes defined as summary statistics (e.g. mean) across all the voxelwise measures from each examined ROI. Genome-wide association study (GWAS), a widely used univariate analysis method, has been performed in a large number of studies of neurodegenerative diseases to discover genetic variants susceptible to brain iQTs. For example, GWAS of these iQTs have identified disease-relevant genetic associations with brain imaging measures in the studies of Alzheimer's disease (AD) [e.g. those reviewed in Shen *et al.* (2014)]. Targeted genetic association studies have also been performed on brain iQTs to associate candidate variants to brain ROIs for increasing statistical power and improving biological interpretation, based on valuable prior knowledge, such as genetic functional information (Rajagopalan *et al.*, 2012; Yao *et al.*, 2019). However, most ROI-based approaches simply collapse voxel measures into a single value, which reduces detection power to find weak signals (e.g. those that exist only in part of an ROI).

Recent advances in acquiring multi-modal neuroimaging technologies inherently provide detailed voxel-level information and thus offer enormous opportunities for examining fine-grained brain abnormalities. In brain imaging genetics, voxelwise strategies have been proposed to explore genetic effects on the voxel-based measures of the brain. Stein *et al.* (2010) proposed voxelwise GWAS (vGWAS) for exploring the pairwise associations between 448 293 SNPs with 31 622 voxels in an AD study. Hibar *et al.* (2011) further presented voxelwise gene-wide association study (vGeneWAS), which employed a multivariate model to relate the joint effect of multiple SNPs within a gene with voxel-level measures. Although a few genes have been suggested by vGWAS (e.g. *CSMD2* and *CADPS2*) and vGeneWAS (e.g. *GAB2*), neither approach identified any significant imaging genetic association surviving multiple testing correction. Clearly, voxelwise imaging genetics studies are facing a huge burden of multiple testing correction, given the ultra-high dimensionalities of the imaging and genetic data.

Random field theory (RFT), implemented in the SPM software package (https://www.fil.ion.ucl.ac.uk/spm/), has been widely employed in brain imaging studies to alleviate multiple comparison burden in voxelwise analyses (Brett *et al.*, 2004). Recently, it has also been applied and extended to the imaging genetics studies. For example, Ge *et al.* (2012) presented a fast implementation of voxel- and cluster-wise inferences to take into consideration the spatial correlation in the imaging data. The proposed approach was applied to an AD study and reported several significant imaging genetic associations (e.g. associations between *GRIN2B* and volumetric changes in the parietal and temporal lobes). Note that cluster-wise inference here considers the spatial correlation among voxels but does not incorporate structural or functional annotation information into the model.

In genomic studies, pathway enrichment analysis has been widely performed where gene sets corresponding to biological pathways are examined for association with a phenotype, to help increase statistical power and interpret genomic findings with meaningful biological annotations. Two types of enrichment methods are often used in pathway analysis of GWAS findings: (i) threshold-based methods (e.g. Fisher's exact test, binomial z-test) that evaluate if the pathway is over-represented in a list of significant GWAS hits; and (ii) rank-based methods [e.g. GSEA-SNP (Holden *et al.*, 2008)] that employ Kolmogorov–Smirnov-like test to determine if the genes from a pathway are randomly distributed in GWAS results. In AD studies, several enrichment analyses have demonstrated that genes functioning in the same pathway can influence imaging traits collectively even when constituent SNPs do not show significant associations individually (Ramanan *et al.*, 2012; Yao *et al.*, 2017b).

Inspired by the above observations, in this work, we introduce enrichment analysis into the imaging domain and propose an enrichment-based framework for mining regional imaging genetic associations. We call this strategy as 'regional imaging genetic enrichment analysis' (RIGEA), which uses meaningful brain anatomical information as prior knowledge to estimate the collective effect of a given genetic variant on all the voxels within a pre-defined brain ROI.

Compared with traditional ROI-wise or vGWAS methods and the RFT-based cluster-wise inference approaches, the advantage of the proposed framework is 3-fold: (i) To the best of our knowledge, this is the first study to apply the enrichment analysis method widely used in the genomic domain to solving similar problems in the brain imaging domain. An ROI is a set of voxels with the same structural or functional annotation. A pathway contains a set of genes with certain functional annotation. Thus, the ROI in the imaging domain plays the same role as the pathway in the genomic domain. This enables a natural expansion of enrichment analysis from the genomic domain to the imaging domain. (ii) The RIGEA framework, by exploring the collective effect of weak voxelwise signals, properly addresses the limited power of ROI-based method which collapses voxel measures into a single value. (iii) The RIGEA framework is powerful and flexible. The implementation of enrichment not only significantly reduces the multiple comparison burden and increases statistical power, but also can be applied to ROIs with various structural and functional annotations to enable biological interpretation.

To show the effectiveness of RIGEA, we compare it with traditional ROI-based and voxelwise approaches via an imaging genetics study in AD, as well as with an RFT-based cluster-wise inference approach implemented in the SPM software package. Furthermore, we interpret and validate the RIGEA findings in brain- and AD-relevant studies, including brain tissue-specific expression QT locus (eQTL) analysis, genetic association study of AD and functional annotation analysis using AD pathways.

2 Materials and methods

We first discuss the data and relevant analyses used in this study, then introduce the detailed methods of the proposed RIGEA framework and finally, describe the strategies for performance evaluation and results validation.

2.1 Imaging and genotyping data

To demonstrate the power of the proposed RIGEA framework, we apply it to the fluorodeoxyglucose [^{18}F]FDG-positron emission tomography (PET) imaging genetics analysis in the study of AD. FDG-PET has been used to measure cerebral metabolic rates of glucose, and its change occurs early in AD (Mosconi *et al.*, 2010).

The imaging and genotyping data used for GWAS were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, PET, other biological markers and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

Preprocessed [^{18}F]FDG-PET scans were downloaded from the LONI website (adni.loni.usc.edu), and then aligned to each participant's same visit scan and normalized to the Montreal Neurological Institute (MNI) space as $2 \times 2 \times 2$ mm^3 voxels. Due to the computational and space cost of vGWAS, we further resliced each high-resolution scans to $4 \times 4 \times 4$ mm^3 using nearest neighbour interpolation strategy. FDG measurements of 26 951 voxels were extracted and 116 ROIs were further computed using the mean of voxel measures within each ROI based on the MarsBaR AAL atlas as previously described in Yao *et al.* (2017a). The number of voxels within 116 ROIs ranges from 7 to 788. Nine hundred and ninety-eight non-Hispanic Caucasian participants (Supplementary Table S1) with complete baseline voxel-level and ROI-level FDG measurements were studied. To evaluate the down-sampling quality, we randomly selected 20 (out of 998) participants and 20 (out of 116) ROIs. For

each participant, the Kolmogorov–Smirnov test was performed on each ROI, to evaluate if there is significant difference between the distributions of voxel-level values from raw imaging data and down-sampled imaging data. Over 98% participant-ROI pairs (i.e. 393 out of 400) demonstrated no significant difference, indicating an acceptable down-sampling quality. The diagnostic information was further used in the evaluation of RIGEA findings for associations with AD status (Section 2.6.4).

Genotyping data were also obtained from the ADNI database and were quality controlled (QCed) as described in Yao *et al.* (2019). Briefly, genotyping was performed on all ADNI participants following manufacturer's protocol using blood genomic DNA samples and Illumina GWAS arrays (610-Quad; OmniExpress, or HumanOmni2.5-4v1) (Saykin *et al.*, 2010). QC was performed in PLINK v1.90 (Purcell *et al.*, 2007) using the following criteria: (i) call rate per marker ≥95%, (ii) minor allele frequency ≥5%, (iii) Hardy–Weinberg equilibrium test $P \geq 1.0E-6$ and (iv) call rate per participant ≥95%. Significant relatedness pairs with PI_HAT >0.45 were identified, and thereafter one individual from each pair was randomly excluded. In total, 565 373 SNPs were obtained for 998 participants based on the ADNI-1 610-Quad panel.

2.2 GWAS of FDG-PET imaging

We performed both voxelwise and ROI-wise GWAS on FDG-PET imaging measures, using linear regression under an additive genetic model in PLINK (Purcell *et al.*, 2007), with age, gender and education as covariates. *Post hoc* analysis used Bonferroni correction for adjusting both the number of SNPs and the number of iQTs (i.e. voxel number for voxelwise analysis and ROI number for ROI-wise analysis).

For performance evaluation purpose, we further constructed a novel ROI-level P-value using a summarized statistic from the voxel-level P-values, named as 'voxel set analysis', which we borrowed the idea of gene set analysis that assigned the k-th smallest SNP-level P-value across all SNPs inside the gene as the gene-level P-value (Nam *et al.*, 2010). Here, we chose the second-best voxel-level P-value across all voxels inside an ROI to represent the ROI-level P-value, to avoid spurious associations from the best P-value.

2.3 Genome-wide meta-analysis of AD

A landmark large-scale genome-wide meta-analysis of clinically diagnosed AD and AD-by-proxy was recently conducted by Jansen *et al.* (2019). Three phases were included as summarized in Yao *et al.* (2019). Briefly, Phase 1 was a meta-analysis of multi-cohort GWAS findings of the AD status with total of 79 154 samples from three major AD genetics consortia including AD working group of the Psychiatric Genomics Consortium, the International Genomics of Alzheimer's Project and the AD Sequencing Project. Phase 2 was a GWAS of the AD-by-proxy status with 376 113 samples from the UK Biobank. Phase 3 was a meta-analysis of Phases 1 and 2 findings, with a total of 455 238 samples (71 880 cases and 383 378 controls). In this study, we downloaded the summary statistics of the Phase 3 analysis (available at https://ctg.cncr.nl/software/summary_statistics) to examine if our RIGEA findings were associated with AD.

2.4 RIGEA framework

Pathway analyses have been widely used in genomic studies to examine gene sets corresponding to biological pathways for their associations with studied phenotypes (Ramanan *et al.*, 2012). In this work, we expand the scope of enrichment analysis from genomics to brain imaging and propose a novel framework for mining regional imaging genetic associations via voxelwise enrichment analysis. Briefly, we treat brain anatomical regions as pathways, each of which contains a set of voxels and perform enrichment analysis on the voxelwise statistics to identify ROIs over-represented by top voxelwise findings. Below, we describe the details of the proposed RIGEA framework (Fig. 1).

We obtained the voxelwise genetic association results from Section 2.2, including P-values between $S = 565\,373$ SNPs and $N = 26\,951$ voxels. Considering the computational cost resulted from the large number of SNPs, we chose threshold-based strategy for enrichment analysis. Specifically, we employed Fisher's exact test with Bonferroni correction for multiple tests. Given a SNP S_i, the imaging genetic findings are a list L_i of significant SNP-voxel associations with P-values passing a pre-defined threshold. Given an ROI R_k that contains total r_k voxels $V_k = \{v_{k,1}, \ldots, v_{k,r_k}\}$, RIGEA aims to determine whether the set of voxels V_k from the targeted ROI R_k is enriched in L_i.

Now, we present our RIGEA framework. Given a SNP S_i, we have N S_i-voxel associations from voxelwise imaging genetic association analysis, out of which $n_i = |L_i|$ (the set L_i) are significant ones with P-value passing a pre-defined threshold. Within ROI R_k, there are $r_k = |V_k|$ (out of N) S_i-voxel associations, of which l_i significant ones are from L_i. Applying the Fisher's exact test for independence, the enrichment P-value for the ROI R_k associated with SNP S_i is as follows:

$$P_{i,k} = \Pr\left(|V_k \cap L_i| \geq l_i\right) = \sum_{j \geq l_i} \frac{\binom{r_k}{j} \times \binom{N - r_k}{n_i - j}}{\binom{N}{n_i}}. \quad (1)$$

Here, $Pr(\cdot)$ is the probability function.

2.5 Evaluation of the RIGEA framework

We evaluated the statistical power of RIGEA on discovering imaging genetic associations by comparing it with ROI-based GWAS and voxel set analysis (second-best P-value strategy) as described in Section 2.2.

We also compared RIGEA with the RFT-based cluster-wise inference implemented in SPM. Due to the computational time cost, it was infeasible to perform cluster-wise inference for each of the 565 373 SNPs. Thus, we narrowed down the comparison from examining all the genome-wide variants to just the meaningful RIGEA findings (see Section 2.6 for detailed information), to check if RFT-based approach could have comparable statistical power on detecting the biologically validated SNP-ROI associations. Specifically, the additive effects of the identified SNPs from RIGEA were assessed at each voxel using SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/) under one-way ANOVA test with age, gender and education as covariates. The statistical analysis results were assessed at family-wise error (FWE) corrected P-value <0.05. The genetic effects were then mapped onto the brain via cluster-wise analysis results.

2.6 Biological interpretation and validation of the RIGEA findings

To determine the biological functions and molecular mechanisms of the RIGEA imaging genetic findings, we test whether the genetic findings from RIGEA are functionally relevant to AD and explore the effect of identified SNPs on gene expression of specific brain tissues and direct effect on AD status.

2.6.1 Functional annotation

Given that FDG glucose metabolism measures in brain regions are promising AD biomarkers, the imaging genetic findings from the FDG study could have direct or indirect connection with AD. Thus, we proposed to use AD pathways to evaluate the functional relevance of our findings. We collected the AD-related pathways from three publicly available databases (Qiagen RT^2 Profiler PCR Arrays, KEGG pathway database and MSigDB) and further kept those with size ranging from 10 to 400 (Ramanan *et al.*, 2012). We then mapped the RIGEA identified SNPs to their closest genes (within ±20 kb), and grouped these genes by their associated ROIs (i.e. for each ROI, formed a set of its associated genes). Finally, we tested the enrichment of the AD-relevant pathways in each set of ROI-associated genes.
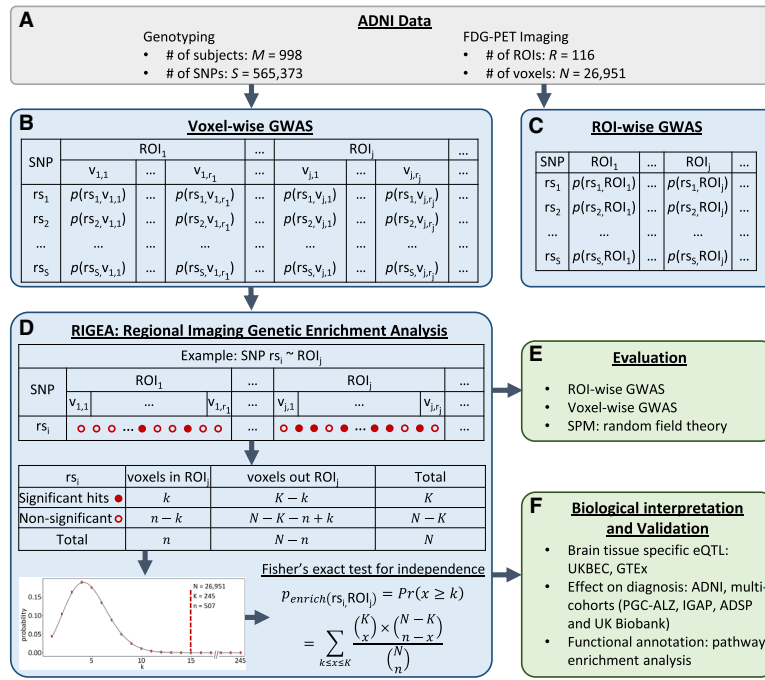
Fig. 1. RIGEA. The workflow illustrates the data and methods used in the development of RIGEA. (**A**) Briefly summaries the statistics of the genotyping and imaging data from the ADNI cohort. **B** and **C** describe the voxelwise and ROI-wise GWAS, respectively. (**D**) Shows the detailed implementation of RIGEA, demonstrated by an example that examines association between SNP rs_i and ROI_j. (**E**) Compares the performance of RIGEA with other strategies, for evaluating the statistical power of our proposed method. (**F**) Validates the imaging genetic findings from RIGEA

2.6.2 UKBEC brain tissue-specific eQTL analysis

We validated the novel SNP-ROI findings in brain tissue-specific eQTL analysis. Specifically, we used eQTL dataset available at BRAINEAC (http://www.braineac.org/), a web server for data from the UK Brain Expression Consortium (UKBEC) (Ramasamy *et al.*, 2014). This dataset contains 12 brain tissues from 134 neuropathologically normal subjects. For each RIGEA identified significant SNP-ROI association, we first mapped the ROI to the corresponding UKBEC brain tissue according to the anatomical location, and then assessed the altered gene expression of the SNP in this specific brain tissue. Here, we focused on the cis-effect and examined expressions of genes located within ±100 kb of the SNP. Barplot was employed to illustrate the gene expression stratified by the eQTL findings. These tissue-specific eQTL results can help provide novel insights into neurogenetic mechanisms of how genetic variants affect brain QTs via regulating gene expression and offer molecular evidence for better mechanistic understanding of the studied disease.

2.6.3 GTEx brain tissue-specific eQTL analysis

To further validate the regulatory relationships among SNPs, genes and brain tissues reported from the UKBEC eQTL analysis, we used another brain tissue gene expression dataset to examine the effect of identified eQTLs on brain tissue-specific gene expressions. Specifically, we used brain tissue-specific gene expression data from the Genotype-Tissue Expression (GTEx) project (https://gtexportal.org/home/), an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples in GTEx were collected from 53 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including whole genome sequencing, whole-exome sequencing and RNA-Seq. There are 13 brain tissues included in GTEx, with sample sizes ranging from 80 to 154. Given a significant UKBEC eQTL finding (i.e. a combination of SNP, gene and ROI), we first mapped the ROI to the corresponding GTEx brain tissues according to the anatomical location, and then assessed the association between the SNP and the gene expression in the specific brain tissue. Violin plot was employed to illustrate the altered expression level among genetic groups.

2.6.4 Genetic association analysis of AD diagnosis

In addition, to examine whether the validated genetic findings have a direct effect on AD, we further evaluated their associations with the AD diagnostic status on two datasets as described in Sections 2.1 and 2.3. SNPs identified from RIGEA and further validated in eQTL analysis were assessed for associations with AD diagnosis.

3 Results

We applied our RIGEA framework to the vGWAS of the FDG-PET measures in an AD study, for mining the regional imaging genetic associations. We compared the performance of proposed framework with traditional ROI-wise GWAS and a proposed voxel set analysis (second-best strategy), as well as an RFT-based cluster-wise inference. We further evaluated the molecular mechanisms of the identified SNPs in their associated brain regions and examined the effects of these genetic variants on disease. In the below sections, we report and discuss our results.

3.1 GWAS of FDG-PET iQTs

GWAS was performed on both ROI-wise (i.e. the mean of all voxel measures within the ROI) and voxelwise FDG-PET measures, to examine the imaging genetic associations of 565 373 SNPs with FDG measures of 116 ROIs and 26 951 voxels, respectively. We further performed the voxel set analysis on the vGWAS results by assigning the second-best voxel-level P-value to each ROI and generated a list of 116 novel ROI-level summary P-values.

In the vGWAS, we used the Bonferroni corrected P-value $< 0.05/(26\,951 \times 565\,373) = 3.28\mathrm{e}-12$, and identified 123 significant associations between two SNPs and 12 ROIs. In both the ROI-wise analysis and the second-best voxel set analysis, we used the same Bonferroni corrected P-value $< 0.05/(116 \times 565\,373) = 7.62\mathrm{e}-10$ as the threshold (i.e. corrected for both the number of ROIs and the number of SNPs). With ROI-wise approach, we identified nine SNP-ROI associations covering two SNPs and seven ROIs. With second-best voxel set approach, we identified 46 SNP-ROI associations covering eight SNPs and 24 ROIs. Note that all nine

Table 1. Significant eQTL results from UKBEC and GTEx databases

SNP					RIGEA		eQTL			
Chr	rsID	bp	A1/A2	Closest gene	ROI	*P*-value	Brain tissue	Beta	*P*-value	Regulated gene
2	rs2863242	113989236	T/C	*PAX8*	Lingual_R	3.54E−07	UKBEC OCTX	−0.33	2.18E−08	*PAX8-AS1* (*LOC654433*)
							GTEx Cortex	−0.67	7.00E−06	*PAX8*
							GTEx Cortex	−0.67	9.60E−06	*PAX8-AS1*
17	rs4668	9814423	T/C	*GAS7*	Frontal_Mid_R	1.06E−02	UKBEC FCTX	0.50	2.80E−15	*RCVRN*
							GTEx Frontal Cortex	1.10	3.00E−27	*RCVRN*

Note: UKBEC uses the symbol *LOC654433*, which has already been replaced by *PAX8-AS1* in the NCBI gene database (https://www.ncbi.nlm.nih.gov/gene/?term=LOC654433). We choose to use *PAX8-AS1* in this paper for consistency.

SNP-ROI associations from ROI-wise GWAS were identified by the second-best voxel set analysis. Of note, two SNPs rs769449 and rs4420638 reported from both voxelwise and ROI-wise GWAS are located in well-known AD genes (rs769449 is located in *APOE* and rs4420638 is located in *APOC1*) and are in linkage disequilibrium (*D'* = 0.954 calculated from 1000 Genome Phase 3 CEU population). Detailed findings of vGWAS, ROI-wise GWAS and second-best voxel set analysis were listed in Supplementary Tables S2–4.

3.2 RIGEA imaging genetic discoveries

For each SNP, we obtained a list of 26 951 SNP-voxel associations across all voxels in the brain. Given a SNP S_i, for each ROI, we assessed the collective effect of S_i on all voxels located inside the ROI by calculating the enrichment score, to relate S_i to ROI. We employed $P \leq 1e - 5$ as the threshold to determine the list of significant SNP-voxel associations for RIGEA, to avoid missing individually moderate while collectively significant signals. We obtained enrichment *P*-values between all $565\,373$ SNPs with 116 ROIs, among which 4435 SNP-ROI pairs were significant after correcting for both the number of SNPs and the number of ROIs (i.e. *P*-value $< 0.05/(116 \times 565\,373) = 7.62e - 10$). To avoid spurious, over-representation associations resulted from a small number of significant hits across all voxels [i.e. small L_i value in Equation (1)], we kept significant findings with $L_i > 10$ and obtained 2979 (out of 4435) associations. These RIGEA findings covered 2100 SNPs and 112 unique ROIs. We further mapped the identified SNPs to their closest genes (± 20 kb). A full list of RIGEA findings is shown in Supplementary Table S5.

3.3 Performance evaluation

We compared RIGEA significant findings with ROI-wise GWAS and second-best voxel set analysis results. RIGEA identified 2946 novel imaging genetic associations in addition to 33 SNP-ROI pairs that were also found by the other two methods. Note that the nine significant associations from ROI-wise GWAS were all captured by RIGEA. As expected, RIGEA not only conserved high concordance with findings from ROI-wise and second-best voxel set strategies, but also reported novel SNP-ROI associations. This indicates that integrating fine-grained association statistics with brain ROI information could help identify high-level imaging genetic associations with meaningful biological interpretation. Supplementary Table S6 shows the 33 overlapped findings.

3.4 Biological interpretation and validation

3.4.1 Functional annotation of the RIGEA findings

We extracted totally 18 AD relevant pathways from three resources: 10 from Qiagen RT^2 Profiler PCR Arrays human AD, one from KEGG AD pathway and eight from MSigDB regulation analysis of AD. In our functional annotation analysis, we excluded five pathways with too small or too large sizes, and included 13 pathways with sizes ranging from 10 to 400.

After mapping the 2100 unique SNPs from RIGEA to the closest genes, we obtained 829 unique genes associated with 110 ROIs and then tested the enrichment of 13 AD pathways in each of 110 gene sets for functional relevance to AD. Supplementary Table S8 shows the significantly enriched AD pathways under Bonferroni corrected *P*-value of $0.05/13/110 = 3.50e - 5$ (i.e. corrected for both the number of gene sets and the number of pathways). Seven AD pathways are significantly enriched in 31 ROI-specific genetic findings, indicating the power of RIGEA for identification of disease risk factors.

3.4.2 UKBEC brain tissue-specific eQTL analysis

Identified genetic variants can influence the disease-relevant imaging phenotypes through gene expression regulation. Thus, we further examine the biological significance of 2946 new SNP-ROI findings from RIGEA, through brain tissue-specific eQTL analysis using genotyping and expression data of 12 brain tissues from UK Brain Expression Consortium (UKBEC) (Ramasamy *et al.*, 2014).

There were totally 2098 unique SNPs and 112 unique ROIs covered by 2946 new RIGEA hits. After mapping ROIs to UKBEC brain tissues (not every ROI can be mapped to a UKBEC tissue), 1006 SNP-ROI pairs remained, covering 735 SNPs, 72 ROIs and seven UKBEC brain tissues. We assessed the cis-effects of these 735 SNPs on brain tissue-specific expression levels of genes located within ± 100 kb from the SNPs. We obtained a list of 3121 tissue-specific eQTLs (see Supplementary Table S7 for a full list of UKBEC eQTL result) among which 179 are nominally significant with *P*-value < 0.05. We further restricted the significance threshold to $0.05/3121 = 1.60e - 5$, and identified two cis-eQTLs.

Table 1 lists the details of the two identified SNPs, including the RIGEA results and eQTL results from UKBEC analysis. The SNP rs2863242, located in gene *PAX8*, is significantly associated with the right lingual FDG glucose metabolism (corrected *P* = 3.54e − 07). The right lingual gyrus is located in occipital cortex (i.e. UKBEC OCTX), in which rs2863242 is significantly associated with the expression of gene *PAX8-AS1* (*P* = 2.14e − 08). Another significant finding is rs2863242, which is located in gene *GAS7* and is significantly associated with right middle frontal FDG glucose metabolism (corrected *P* = 1.06e − 02). The right middle frontal gyrus is located in frontal cortex (i.e. UKBEC FCTX), in which rs4668 regulates the gene expression of *RCVRN* with *P* = 2.80e − 15.

Figure 2 shows the associations and effects of the identified eQTLs on expression of regulated genes. Figure 2A illustrates the differential expression level of *PAX8-AS1* across three rs2863242 genotype groups (i.e. CC, CT and TT), showing that the presence of minor allele T of rs2863242 has an additive effect on decreasing expression of *PAX8-AS1*. Figure 2B shows the expression of *RCVRN* stratified by three rs4668 groups (i.e. CC, CT and TT), indicating the additive effect of minor allele T of rs4668 on increasing expression of *RCVRN*.

The region-specific eQTL results shown above demonstrate the ROI-relevance of the identified SNPs and their roles on molecular regulation in the corresponding brain regions. This indicates that the
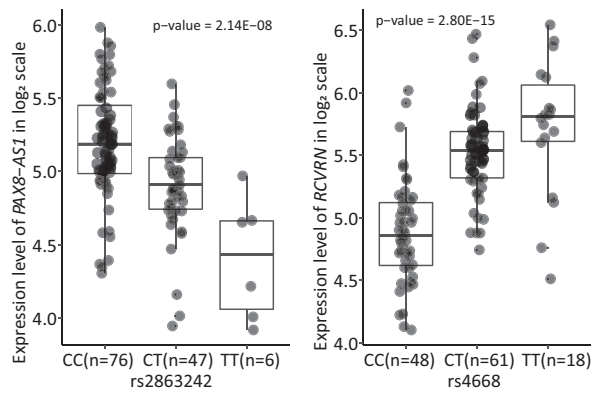
Fig. 2. UKBEC brain tissue-specific cis-eQTL results. (**A**) The expression level of *PAX8-AS1* in the temporal cortex stratified by rs2863242 groups. *P*-value indicates the association significance of rs2863242 with the expression of *PAX8-AS1* in the occipital cortex. *P*-value is calculated from linear regression with gender as covariate. Presence of minor allele T of rs2863242 suggests an additive effect of decreasing gene expression of *PAX8-AS1*. (**B**) The expression level of *RCVRN* in the frontal cortex stratified by rs4668 groups. *P*-value indicates the association significance of rs4668 with the expression of *RCVRN* in the frontal cortex. *P*-value is calculated from linear regression with gender as covariate. Presence of minor allele T of rs4668 suggests an additive effect of increasing gene expression of *RCVRN*.

proposed framework can promote the identification of the imaging genetic associations with evidence manifest at the brain transcriptome level, with the potential to help interpret the underlying molecular mechanisms.

3.4.3 GTEx brain tissue-specific eQTL analysis

For further validation of the eQTL findings, we used tissue-specific expression data in GTEx to examine the effects of rs2863242 and rs4668 on the expression of *PAX8-AS1* and *RCVRN* in the corresponding brain tissues, respectively. Table 1 shows the GTEx eQTL results of rs2863242 and rs4668. Given the 13 available GTEx brain tissues, right lingual gyrus and right frontal middle gyrus are mapped to cortex (GTEx Brain-Cortex) and frontal cortex [GTEx Brain-Frontal Cortex (BA9)], respectively. Supplementary Figure S1 shows the expressions of genes stratified by genetic groups. Minor allele T of rs2863242 down-regulates the expression of *PAX8* and *PAX8-AS1* with *P*-values of $7.00e-06$ and $9.60e-06$. Minor allele T of rs4668 is associated with the up-regulation of *RCVRN* expression ($P = 3.00e-27$). This concordance between GTEx and UKBEC eQTL results further confirms the potential regulatory roles that identified variants play on the expression of the relevant genes in the corresponding brain regions.

3.4.4 Association of the RIGEA findings with AD

To examine whether the genetic findings from RIGEA are directly associated with the AD status, we first leveraged the valuable results from the most recent large-scale genome-wide meta-analysis of AD by Jansen *et al.* (2019). According to the summary statistics of Phase 3, rs2863242 exhibited a significant association with AD ($P = 1.32e-2$; $N = 458\,744$). The corresponding effect size is -0.0062, indicating a protective role of the minor allele rs2863242-T in AD. No significant association between rs4668 and AD ($P = 0.20$) was observed.

We further checked the AD association of these two SNPs in the ADNI data, using the participants involved in our RIGEA analysis. There was a significant association between rs4668 and AD ($P = 2.42e-2$; $N = 998$) with effect size of -0.071, suggesting a protective effect of minor allele rs4668-T for AD. No significant association between rs2863242 and AD ($P = 0.83$) was observed.

3.4.5 Comparison with RFT-based approach

As mentioned in Section 2.5, we did a comparative study between RIGEA and the RFT-based analyses implemented in the SPM

Table 2. SPM cluster-wise analysis of effect of SNPs on FDG glucose metabolism

SNP	FWE p	Number of voxel	x	y	z	ROI
rs4668	0.002	38	−44	48	8	Front Mid R
	0.009	16	−52	30	18	Front Inf Tri R
	0.012	13	52	36	16	Front Inf Tri L
	0.023	5	10	70	−6	Front Med Orb L
	0.029	3	44	58	−2	—
rs2863242	—	—	—	—	—	—

Note: x, y and z represent the coordinates of cluster centres in the MNI space.

software package on the RIGEA significant findings. Due to the computational cost, we performed the comparison only on two identified SNPs, rs2863242 and rs4668. Table 2 shows the significant clusters identified from voxelwise analysis of the additive effect of SNPs with FWE $P < 0.05$. There are five clusters identified to be associated with rs4668 with the number of voxels ranging from 3 to 38, four out of which are located in the left and right frontal lobe. Supplementary Figure S2 illustrates the additive effect of rs4668 on FDG glucose metabolism, where the minor allele of rs4668 is associated with higher FDG glucose metabolism. There is no significant finding for rs2863242. *Note that the FWE P-value from SPM voxelwise analysis is corrected for the clusters, while not corrected for the number of SNPs. These findings will face a major burden of multiple comparison correction if a huge number of genome-wide SNPs are examined.*

4 Discussion and conclusions

We have proposed a RIGEA framework via an innovative application of enrichment analysis to the imaging domain, to explore the effect of each genetic variant on an ROI by aggregating fine-grained voxelwise imaging genetic associations using anatomically or functionally annotated ROI information. We have demonstrated the effectiveness of RIGEA using the ADNI imaging genetics data. In addition to associations identified by traditional ROI-wise and voxel set analyses, our approach has reported novel SNP-ROI findings, some of which have been revealed as eQTLs regulating genes expressed in the corresponding regions. Our approach further outperforms the RFT-based image analysis strategy through achieving more significant *P*-values on biologically eQTLs. These findings demonstrate the power of the presented method on identifying individually modest while collectively substantial imaging genetic signals.

Our approach effectively leverages neuro-anatomical knowledge and voxelwise statistics. The neuro-anatomical knowledge guides the construction of voxel sets (i.e. analogy to pathways in the genomic domain) that facilitate meaningful interpretation of the enrichment findings. The use of fine-grained voxel-level statistics can promote the identification of signals which are individually modest while jointly substantial. Our biologically validated associations can help provide valuable information for revealing the molecular pathway from genetic variants, gene expression and brain region to disease and further help improve the understanding of complex disease pathology.

The two RIGEA discovered region-specific eQTLs, rs2863242 and rs4668, are located in gene *PAX8* and *GAS7*, respectively. *PAX8* has been studied for its critical role in pattern formation in central nervous system development (Song *et al.*, 1996) and shown its differential methylation between schizophrenics and healthy controls in human brain cortex (Siegmund *et al.*, 2007). The *GAS7*, growth arrest-specific protein 7, has been widely studied for its primary expression in brain cells and functional involvement in the AD progression (Akiyama *et al.*, 2009; Gotoh *et al.*, 2013; Hidaka *et al.*, 2012). Given these biological evidences, it warrants further

investigation to examine the regulatory roles of the identified SNPs on the expression level of the corresponding genes in relevant brain regions, and their relationship with AD pathology.

The real power of RIGEA, however, can be affected by several factors. First, Fisher's test requires a pre-defined threshold to determine the list of significant SNP-voxel pairs. Although this makes the framework more flexible in practice for tightening or relaxing voxel-level effects, it considers only the number of significant pairs without taking into account the full spectrum of association statistics. Rank-based enrichment strategies [e.g. (Subramanian *et al.*, 2005)] can be employed in our framework to overcome these limitations, while their computational burden is considerable. Another issue is that RIGEA requires to compute voxel-level associations in advance, which is both time and space demanding, especially given millions of SNPs in GWAS data. Therefore, another direction is to design scalable computational framework for accelerating the voxel-level GWAS.

In sum, we have proposed the RIGEA framework by expanding the scope of pathway enrichment analysis from genomics to imaging domain and demonstrated its performance with an imaging genetics study in AD. The imaging genetic findings conserve both regional specificity and functional relevance in the brain, showing the promise of the proposed method. This work can be further expanded towards several future directions. For example, one direction is to employ rank-based strategy for enrichment test to make use of the full spectrum of voxelwise statistics. Another direction is to apply this framework to other annotations like functional annotation, brain connectivity network for integrating brain regional information into vGWAS.

References

Akiyama,H. *et al.* (2009) A novel role for hGas7b in microtubular maintenance: possible implication in Tau-associated pathology in Alzheimer disease. *J. Biol. Chem.*, **284**, 32695–32699.

Brett,M. *et al.* (2004) Introduction to random field theory, Chapter 44. In: Frackowiak, R.S.J. *et al.* (eds.) *Human Brain Function*, 2nd edn. Academic Press, pp. 867–879. doi: 10.1016/B978-012264841-0/50046-9.

Ge,T. *et al.* (2012) Increasing power for voxel-wise genome-wide association studies: the random field theory, least square Kernel machines and fast permutation procedures. *Neuroimage*, **63**, 858–873.

Gotoh,A. *et al.* (2013) Gas7b (growth arrest specific protein 7b) regulates neuronal cell morphology by enhancing microtubule and actin filament assembly. *J. Biol. Chem.*, **288**, 34699–34706.

Hibar,D.P. *et al.* (2011) Voxelwise gene-wide association study (vGeneWas): multivariate gene-based association testing in 731 elderly subjects. *Neuroimage*, **56**, 1875–1891.

Hidaka,M. *et al.* (2012) Alzheimer's disease-related protein hGas7b interferes with kinesin motility. *J. Biochem.*, **151**, 593–598.

Holden,M. *et al.* (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics*, **24**, 2784–2785.

Jansen,I.E. *et al.* (2019) Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.*, **51**, 404–413.

Mosconi,L. *et al.* (2010) Pre-clinical detection of Alzheimer's disease using FDG-PET, with or without amyloid imaging. *J. Alzheimers Dis. 2010*, **20**, 843–854.

Nam,D. *et al.* (2010) GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res.*, **38**, W749–W754.

Purcell,S. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Rajagopalan,P. *et al.* (2012) Common folate gene variant, MTHFR C677T, is associated with brain structure in two independent cohorts of people with mild cognitive impairment. *Neuroimage Clin.*, **1**, 179–187.

Ramanan,V.K. *et al.* (2012) Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.*, **28**, 323–332.

Ramasamy,A. *et al.*; UK Brain Expression Consortium. (2014) Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat. Neurosci.*, **17**, 1418–1428.

Risacher,S.L. *et al.* (2010) Longitudinal MRI atrophy biomarkers: relationship to conversion in the ADNI cohort. *Neurobiol. Aging*, **31**, 1401–1418.

Saykin,A.J. *et al.* (2010) Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimers Dement.*, **6**, 265–273.

Shen,L. *et al.* (2010) Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: a study of the ADNI cohort. *Neuroimage*, **53**, 1051–1063.

Shen,L. *et al.*; for the Alzheimer's Disease Neuroimaging Initiative. (2014) Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers. *Brain Imaging Behav.*, **8**, 183–207.

Siegmund,K.D. *et al.* (2007) DNA methylation in the human cerebral cortex is dynamically regulated throughout the life span and involves differentiated neurons. *PLoS One*, **2**, e895–e899.

Song,D. *et al.* (1996) Two Pax-binding sites are required for early embryonic brain expression of an Engrailed-2 transgene. *Development*, **122**, 627–635.

Stein,J.L. *et al.* (2010) Voxelwise genome-wide association study (vGWAS). *Neuroimage*, **53**, 1160–1174.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Yao,X. *et al.*; for the Alzheimer's Disease Neuroimaging Initiative. (2017a) Tissue-specific network-based genome wide study of amygdala imaging phenotypes to identify functional interaction modules. *Bioinformatics*, **33**, 3250–3257.

Yao,X. *et al.* (2017b) Two-dimensional enrichment analysis for mining high-level imaging genetic associations. *Brain Inform.*, **4**, 27–37.

Yao,X. *et al.* (2019) Targeted genetic analysis of cerebral blood flow imaging phenotypes implicates the *INPP5D* gene. *Neurobiol. Aging*, **81**, 213–221.