



Neuroimaging PheWAS (Phenome-Wide Association Study): A Free Cloud-Computing Platform for Big-Data, Brain-Wide Imaging Association Studies

Lu Zhao¹ · Ishaan Batta¹ · William Matloff¹ · Caroline O'Driscoll¹ · Samuel Hobel¹ · Arthur W. Toga¹

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Large-scale, case-control genome-wide association studies (GWASs) have revealed genetic variations associated with diverse neurological and psychiatric disorders. Recent advances in neuroimaging and genomic databases of large healthy and diseased cohorts have empowered studies to characterize effects of the discovered genetic factors on brain structure and function, implicating neural pathways and genetic mechanisms in the underlying biology. However, the unprecedented scale and complexity of the imaging and genomic data requires new advanced biomedical data science tools to manage, process and analyze the data. In this work, we introduce Neuroimaging PheWAS (phenome-wide association study): a web-based system for searching over a wide variety of brain-wide imaging phenotypes to discover true system-level gene-brain relationships using a unified genotype-to-phenotype strategy. This design features a user-friendly graphical user interface (GUI) for anonymous data uploading, study definition and management, and interactive result visualizations as well as a cloud-based computational infrastructure and multiple state-of-art methods for statistical association analysis and multiple comparison correction. We demonstrated the potential of Neuroimaging PheWAS with a case study analyzing the influences of the apolipoprotein E (APOE) gene on various brain morphological properties across the brain in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. Benchmark tests were performed to evaluate the system's performance using data from UK Biobank. The Neuroimaging PheWAS system is freely available. It simplifies the execution of PheWAS on neuroimaging data and provides an opportunity for imaging genetics studies to elucidate routes at play for specific genetic variants on diseases in the context of detailed imaging phenotypic data.

Keywords Discovery science · Magnetic resonance imaging · Genetics · High-performance computing · Web-based system

Introduction

In the past decade, numerous genome-wide association studies (GWASs) that associate specific traits with genetic variants across the genome have been performed using disease-specific definitions to identify novel genetic influences on

many diseases (Hindorff et al. 2009; Horwitz et al. 2019; Visscher et al. 2017). The findings improve the understanding of risks for the diseases, and may guide diagnosis and therapy on a patient-specific basis (Van Cauwenberghe et al. 2016). However, the path from GWAS to biology is not straightforward because an association between a genetic variant at a genomic locus and a trait is not directly informative with respect to the target gene or the mechanism whereby the variant is associated with phenotypic differences (Visscher et al. 2017). The effects of neurological and psychiatric disorders such as Alzheimer's disease (AD), Parkinson's disease (PD), schizophrenia, bipolar disorder and autism on brain structure and function can be seen in neuroimaging data in vivo (Toga 2015). Neuroimaging can therefore provide intermediate endophenotypes, and joint analysis of the genetic and neuroimaging datasets provides a chance for uncovering the genetic architecture of such disorders. Thus, it is of interest to assess

Lu Zhao and Ishaan Batta contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s12021-020-09486-4>) contains supplementary material, which is available to authorized users.

✉ Arthur W. Toga
toga@loni.usc.edu

¹ Laboratory of Neuro Imaging, USC Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA, USA

effects on the brain of candidate genes that have been previously identified in GWASs as exploratory follow-ups of the initial discoveries. For example, when a strong genome-wide supported variant or mutation has been found to be highly significant in a neurological disease on the basis of GWAS, then mapping the genetic marker on the brain is a high priority (Medland et al. 2014).

The recent emergence of neuroimaging and genomic databases of large healthy and diseased cohorts (Bycroft et al. 2018; Jack Jr. et al. 2008; Jernigan et al. 2016; Satterthwaite et al. 2014; Thompson et al. 2010) enables scientific discovery through conducting broad surveys that examine true system-level gene-brain relationships. Previous candidate-gene or whole-genome imaging studies, in general, examined only a limited number of imaging variables of specific brain regions (Glahn et al. 2007). Recent so-called “whole-brain” studies attempted to conduct a broader survey across the brain, but still focused on a single type of imaging measurement at a time, e.g. regional/local cortical volume (Medland et al. 2014; Shen et al. 2010). Cortical volume geometrically is a combination of cortical thickness and surface area. It has been demonstrated that cortical thickness and surface area are biologically independent, as they are driven by distinct cellular mechanisms (Pontius et al. 2008; Rakic 1988), and are differentially affected by genetic factors (Panizzon et al. 2009; Winkler et al. 2010). Thus, different neuroimaging phenotypes may play different roles in imaging genetics (Winkler et al. 2010). The narrow focus of the previous analyses neglects the potential power to detect more distributed gene-brain associations obtained using other imaging phenotypes, and variations in methods and studied samples may also induce false negatives or positives in individual studies. Moreover, previous GWASs have shown that significantly associated genetic markers (single nucleotide polymorphisms (SNPs)) commonly have small effect sizes of the phenotypic variance, thus large sample sizes are needed to increase the statistical power (Medland et al. 2014). Here, we propose a neuroimaging-based phenome-wide association study (PheWAS) approach, which, as an inverse to GWAS, systematically associates genes of interest with a wide variety of neuroimaging phenotypes extracted from large cohorts using a unified genotype-to-phenotype strategy (Fig. 1). Compared with existing single-phenotype methods, this approach provides a broad survey of possible gene-brain relationships on whole populations.

The concept of PheWAS was originally developed for analyses on structured phenotypic data, such as International Classification of Disease (ICD) codes, epidemiologic data, quantitative traits, and clinical conditions (Bush et al. 2016; Denny et al. 2013; Denny et al. 2010; Liao et al. 2014; Neuraz et al. 2013; Pendergrass et al. 2013; Pendergrass et al. 2011). Implementing PheWAS on neuroimaging data is complicated and challenging. The first primary challenge is the

complexity, heterogeneity, and volume of the data involved (Dinov et al. 2016). Especially, neuroimaging genomic data archives are frequently comprised of complex elements that are in heterogeneous file structures and formats, and are, in general, poorly structured. The large data volume (thousands of subjects and tens of thousands of files) and complexity requires intensive computation that is difficult to accomplish using conventional methods (Toga et al. 2015). Advanced computing and storage infrastructure are needed. Sophisticated brain image processing is required for image registration, brain segmentation and parcellation, and extracting various brain structural and/or functional measures as phenotypes. Additionally, the association analysis approaches that are commonly used in case-control GWASs and conventional PheWASs cannot address the spatial correlation across the brain. Statistical brain morphometry analysis that can adapt to the spatial smoothness of the neuroimaging data is needed. Finally, the Manhattan plots commonly used for result visualization in previous GWASs/PheWASs (Carroll et al. 2014; Pendergrass et al. 2012) are not suitable for interpreting the neuroimaging data of hundreds of thousands brain voxels/vertices. Brain mapping visualization techniques are required to clearly display the statistics and anatomical locations of detected patterns.

In this work, we developed a free web-based system - Neuroimaging PheWAS (version 1.0), which addresses the challenges that researchers may face when dealing with large-scale, brain-wide imaging association studies. A summary of the main contributions of Neuroimaging PheWAS is highlighted in Fig. 2. The platform is unique in four ways: (1) it provides an easy-to-use graphical user interface (GUI) for data management and study protocol definition; (2) a cloud-based computational infrastructure enables high-performance brain magnetic resonance imaging (MRI) processing and large-scale PheWAS analyses; (3) it provides a variety of methods for statistical association analysis and multiple comparison correction; (4) it includes a web-based viewer for interactive result visualization and manipulation.

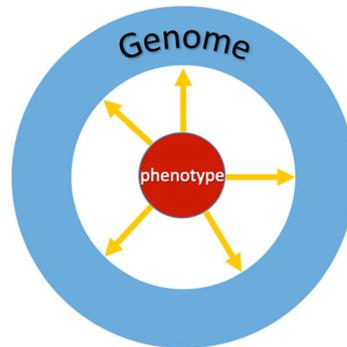
As a case study, we applied Neuroimaging PheWAS on a neuroimaging genomic data set from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database.¹ We analyzed the influences of apolipoprotein E (APOE) genotype, the most established genetic risk factor for AD (Yamazaki et al. 2019), on various surface-/parcellation-based brain morphological properties across the brain. We also performed a series of benchmark tests to evaluate the performance of Neuroimaging PheWAS, using data extracted from the UK Biobank database.² Neuroimaging PheWAS is available online at <http://phewas.loni.usc.edu/phewas>.

¹ <http://adni.loni.usc.edu>

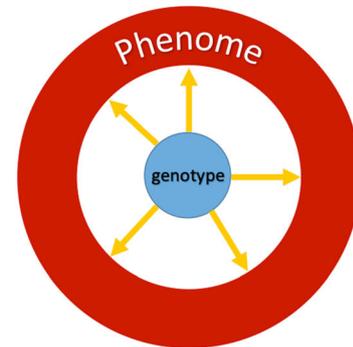
² <https://www.ukbiobank.ac.uk>

Fig. 1 Comparison between GWAS and PheWAS

GWAS: examines associations between specific phenotypes and genetic variants across the genome



PheWAS: examines associations between specific genetic variants and a large number of different phenotypes (phenome)



Materials and Methods

The Neuroimaging PheWAS system (version 1.0) is designed to provide researchers a web-based solution to implement systematic, large-scale association studies on complex parcellation statistics of region-wise neuroanatomical properties, vertex-wise surface-based brain morphometric metrics, and/or other generic phenotypic statistics. The system is comprised of four stages (Fig. 3): 1) initialization, including online account creation and uploading data; 2) defining study protocols and computation of association statistics; 3) real-time, online result visualization and manipulation; 4) downloading results for post hoc analysis and/or publication. The system integrates facilities from multiple platforms to achieve these functionalities (Fig. 3) and requires no specialized hardware or software for end users. It features a user-friendly web GUI designed for different levels of expertise. A User Guide,³ including synthetic, anonymized demo data sets, is accessible on the GUI for the user to learn how to prepare research data files and to use the tool. The web GUI was created using the web scripting languages – PHP⁴ and JavaScript,⁵ along with styling using the front-end framework – Bootstrap.⁶ The modules for various sections on the web tool are implemented such that they communicate with the database management system MySQL⁷ for maintaining records for users, projects and job submissions. The webserver of Neuroimaging PheWAS is set up on a single Intel Xeon CPU (2.3 GHz) with 16 GB of memory, running Debian GNU/Linux 8. It is connected to a high-performance computing (HPC) server⁸ at the Laboratory of Neuro Imaging (LONI), USC Mark and Mary Stevens Neuroimaging and Informatics Institute, University of

Southern California (USC). The LONI HPC server ensures a stable, secure and robust environment with 4096 CPU cores, 38 terabytes of aggregate memory space and 5.3 petabytes of primary storage cluster capacity (see Supplementary Methods for details). It provides Neuroimaging PheWAS the cloud computing (Mauch et al. 2013) resource for storing data and results and for deploying the computational component of the system. The LONI pipeline⁹ (Dinov et al. 2010; Dinov et al. 2009) is used for designing, submitting, executing, and monitoring data analysis workflows on the LONI HPC server. In the following sections, the workflow and technical specifications of the Neuroimaging PheWAS system are described. More details that explain how the GUI interacts with the workflow and system's functionalities are illustrated in Fig. 4.

Initialization

The user first signs up for the system on the Neuroimaging PheWAS website by creating an account with a preferred username and password (Fig. 4a). Once logged in, the user is redirected to the web GUI and is able to upload their research data (Fig. 4d), including tabulated matrices of genotypes, covariates, imaging phenotypes and/or other generic phenotypic statistics. The files are transferred into the user's online data bag on the LONI HPC server automatically and securely (Chard et al. 2016; Czajkowski et al. 2017; Schuler et al. 2016). Neuroimaging PheWAS allows uploading data sets of MR images (NIFTI format) and performing MRI processing using the FreeSurfer software¹⁰ on the LONI HPC grid (Fig. 4c). The distributed computing system parallelizes the FreeSurfer tasks of multiple brain images by assigning the task for each image to a separate dedicated CPU core. After all the FreeSurfer tasks of an imaging data set are completed, the system

³ http://phewas.loni.usc.edu/phewas/user_guide.php

⁴ <https://www.php.net>

⁵ <https://javascript.info>

⁶ <https://getbootstrap.com>

⁷ <https://www.mysql.com>

⁸ cranium.loni.usc.edu

⁹ <http://pipeline.loni.usc.edu>

¹⁰ <http://surfer.nmr.mgh.harvard.edu>

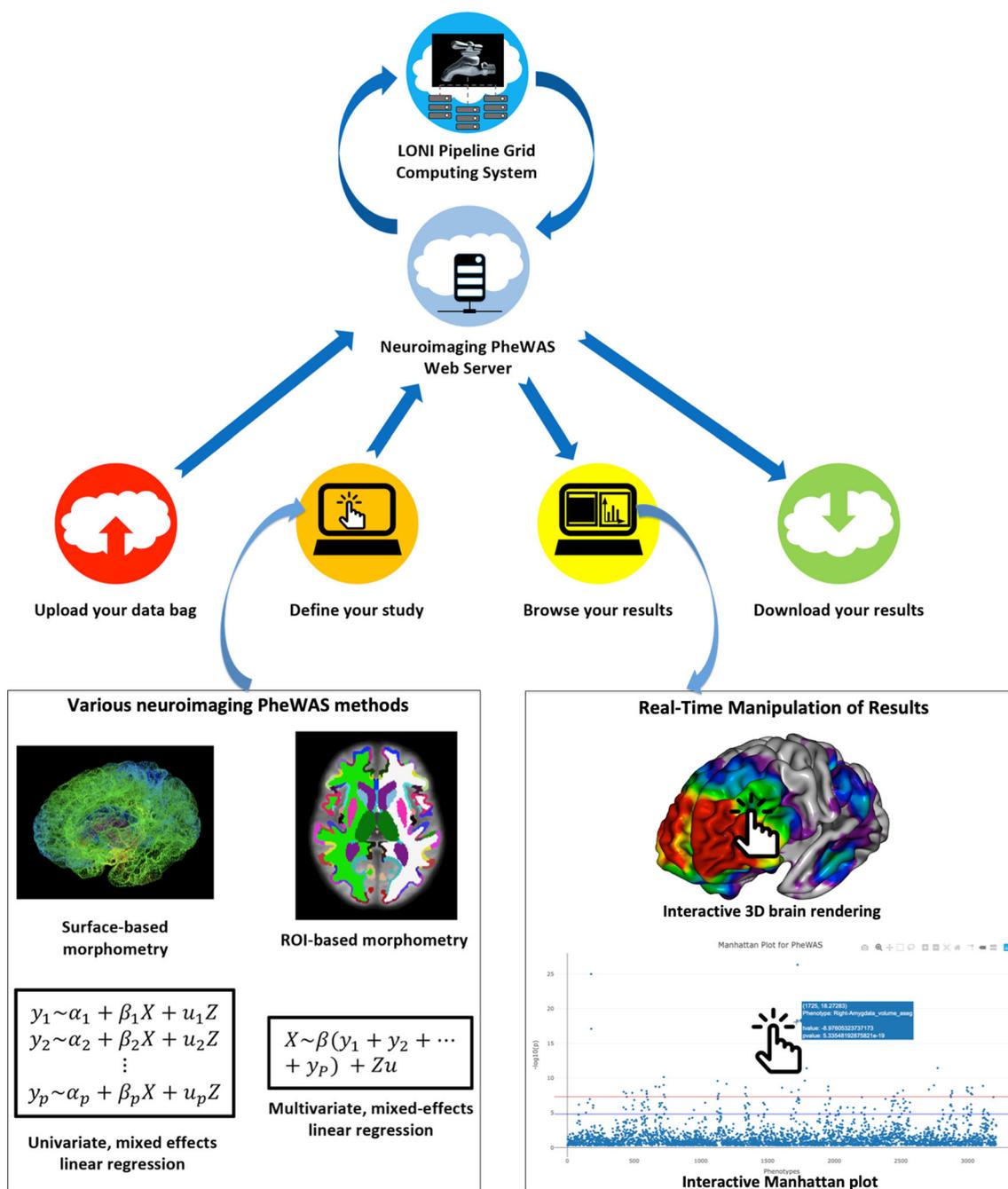


Fig. 2 Summary of the Neuroimaging PheWAS system. Its key features include 1) a user-friendly GUI for anonymous data uploading, study definition and management, 2) a cloud-based computational and storage

infrastructure, 3) multiple methods for statistical association analysis and multiple comparison correction, 4) interactive result visualizations

automatically extracts all parcellation-/surface-based brain morphological measures from the complex FreeSurfer outputs of each image (see Supplementary Methods for details). These statistics are then saved into CSV formatted files for parcellation-based measures and compressed binary files for surface-based measures. These new data files are stored in the user's data bag along with the other uploaded data files and can be viewed in the Data

Manager section of the GUI (Fig. 4d). For the user who performed their own FreeSurfer processing, we provide Java-based and shell scripts (can be downloaded from the User Guide³ on the GUI) for them to locally extract, integrate and save the imaging measures into tabular data files from the complex FreeSurfer outputs. Of note, a number of studies have consistently demonstrated that variation in FreeSurfer version can have a strong effect on brain

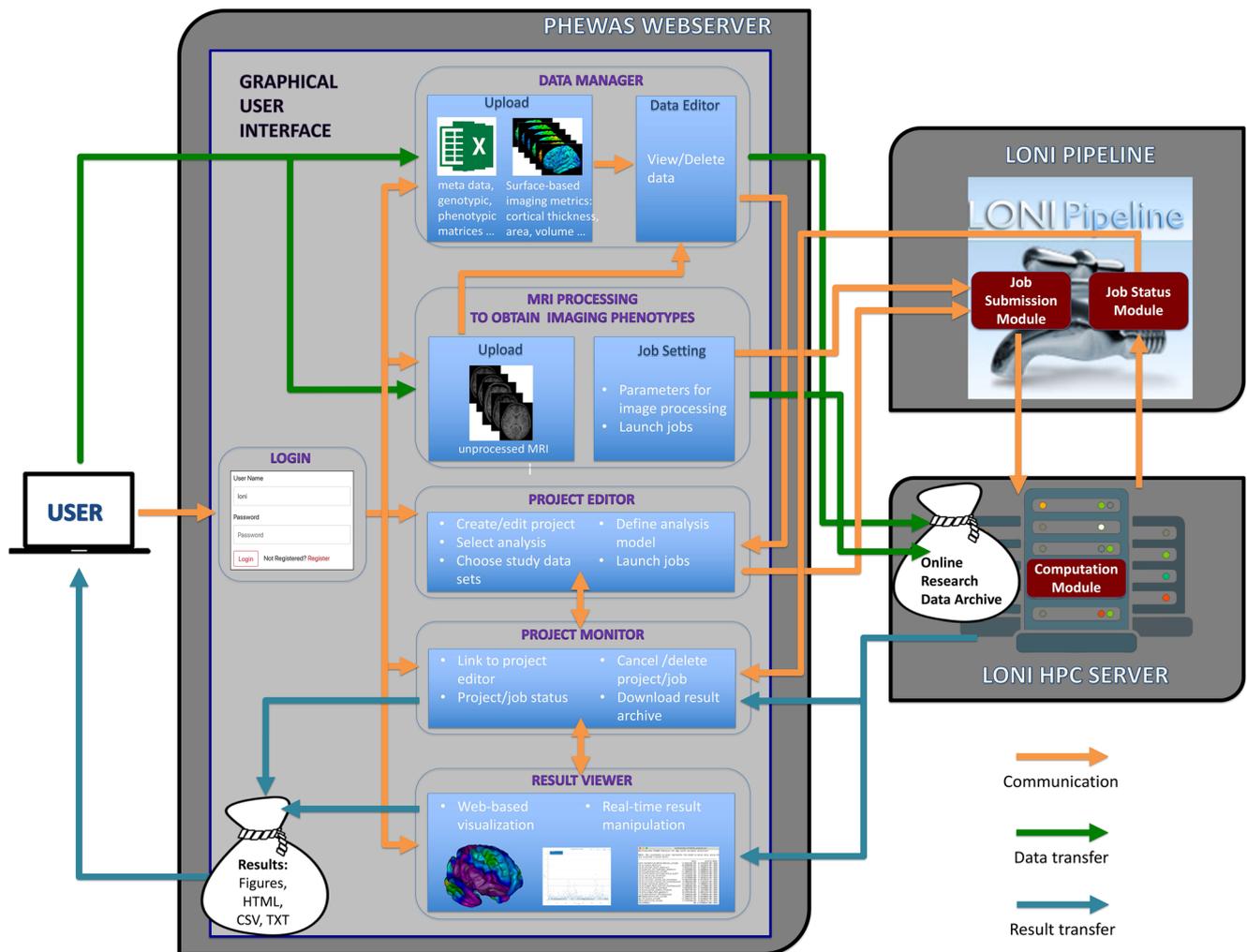


Fig. 3 The infrastructural relations between the integrated platforms and workflows in the Neuroimaging PheWAS system. The user interacts with the GUI hosted at the webserver to manage and manipulate data, studies and results. The webserver is connected to the LONI HPC server where

data and results are stored and computational analyses are deployed. The LONI pipeline submits, executes, and monitors analysis workflows on the LONI HPC grid

structural segmentations and estimates of cortical morphology measures (Bigler et al. 2018; Chepkoech et al. 2016; Gronenschild et al. 2012; Whelan et al. 2016). Therefore, mixing imaging metrics obtained from different versions of FreeSurfer is discouraged as results are expected to differ (Gronenschild et al. 2012). Neuroimaging PheWAS provides different versions of FreeSurfer (from v4.3 to v6.0) to enable the user to select the same version of FreeSurfer that they had used to process new cases in an ongoing study. For new studies, the current FreeSurfer v6.0 is recommended due to its more reliable outputs than the older versions (Whelan et al. 2016).

PheWAS Projects

The user can create and edit projects at the Project Editor section of the GUI (Fig. 4e). A project refers to an association

study specified by a customized project name, the analysis type, the study data files, and the statistical model. The analysis types supported by Neuroimaging PheWAS include surface-based morphometry (SBM) for vertex-wise cortical morphological data and region-of-interest (ROI) based association analysis (univariate or multivariate) for brain parcellation statistics (see section Statistical Analysis and Supplementary Methods for details). When an analysis type is selected, the system identifies the research data files that correspond to the analysis type (e.g. surface data for SBM, parcellation data for ROI-based analysis) from the user's online data bag and lists them in the data fields on the Project Editor panel for the user to select. The system allows selecting one or multiple phenotypic data files for a project. For instance, in a project, the user can test genetic effects on cortical thickness only or on multiple brain morphological measures (thickness, area, volume and etc.) systematically. The Project

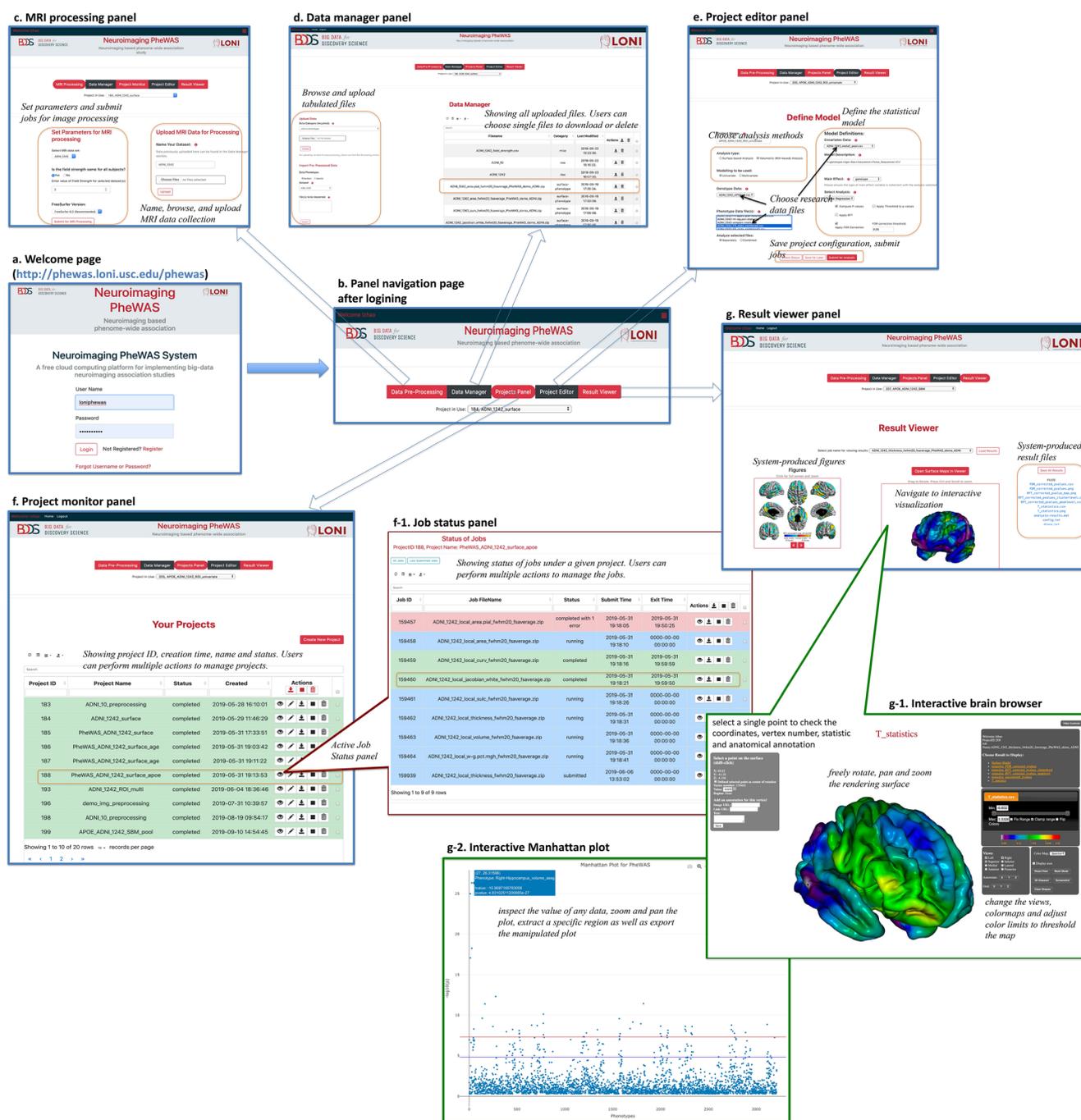


Fig. 4 The web-based GUI for the Neuroimaging PheWAS system. **a** Entering to <http://phewas.loni.usc.edu/phewas> using any web-browser, the user can register and log in to the Neuroimaging PheWAS system. **b** Once logged in, the user can access different panels with the panel navigator. **c** In the MRI Processing panel, the user can browse, select and upload MRI files as a collection and submit jobs for MRI processing. **d** In the Data Manager panel, the user can browse, select and upload tabulated files of research data. The dashboard lists the uploaded data files and allows downloading or deleting selected files. **e** In the Project Editor panel, the user can choose analysis methods and

study data sets, define the statistical model, and submit analysis jobs. **f** The Project Monitor panel displays information and status of projects and allows multiple actions to monitor and manage the projects. The user can activate the Job Status panel (**f-1**) to check the information of jobs under a project, and to terminate/delete jobs. **g** Result Viewer visualizes result statistics as figures and tables, which are downloadable both as individual files and archives. The user also can activate the interactive 3D result rendering on a standard brain surface model (**g-1**) and the interactive Manhattan and Q-Q plots (**g-2**) for real-time result manipulation

Editor panel also allows the user to select the files for the target genotype and covariates. The system recognizes the

variables contained in the selected table of covariates by the column names and lists them in an interactive menu. Using the

interactive menu, the user can select and add covariate(s), polynomial (quadratic or cubic) and/or interaction terms into the statistical model. When a project is defined, Neuroimaging PheWAS converts the configuration of the project into shell-based workflows for each of the selected phenotypic data files, and then submits these workflows to execute as parallel jobs on the grid computing system through the LONI Pipeline. A job refers to the computational instance of an analysis defined in the project.

In each job, prior to performing the analysis, Neuroimaging PheWAS harmonizes the matrices of phenotypes, genotype and covariates by matching the samples included in these matrices based on the 'subject IDs'. These matrices are further cleansed controlling for missing values. Individuals that are mismatched across these matrices and/or have missing values are discarded from the analysis. The number of samples deleted in the data harmonization and cleansing is reported in the output log file. Furthermore, the table of ROI-based phenotypes may contain duplicated variables (the elements of two or more variables are identical to each other), which could be produced when combining multiple tables of FreeSurfer parcellation statistics. The system identifies these duplicates and drops them keeping only the first occurrence. The variable names of the identified duplicates are reported in the output log file as well.

Statistical Analysis

Neuroimaging PheWAS employs state-of-the-art algorithms to implement multitudinous association analyses. SBM (Dahnke and Gaser 2018) implements the statistical approach of parametric mapping to examine influences of a genetic variant on diverse brain morphological properties measured from geometric models of the cortical surface at each vertex, such as cortical thickness (Fischl and Dale 2000), surface area (Winkler et al. 2012), volume (Zhao et al. 2013), and surface curvature (Fischl et al. 1999). For region-wise parcellation statistics, Neuroimaging PheWAS supports performing either univariate analysis for individual genotype-to-phenotype relationships or multivariate analysis for a joint effect of multiple phenotypes. Neuroimaging PheWAS provides a variety of statistical tests, including linear regression, t-test, F-test and analysis of variance (ANOVA), to handle different variable types (continuous, categorical, interaction) in SBM and the univariate ROI-based analysis. The multivariate ROI-based analysis is implemented using the MultiPhen¹¹ method (O'Reilly et al. 2012), which regresses a genotype on a collection of phenotypes of any measurement nature (e.g. continuous, categorical, ordinal) and then calculates a p value for the combination of phenotypes. To prevent overfitting in the multivariate analysis, Neuroimaging PheWAS uses the least

absolute shrinkage and selection operator (LASSO) (Tibshirani 1996) to preselect the phenotypes that have the best joint power to predict the genotype as the inputs for the MultiPhen test. Furthermore, Neuroimaging PheWAS integrates diverse methods to correct for multiple comparisons. Statistical results of SBM at all vertices are adjusted for the family-wise error rate (FWER) using the random field theory (RFT) method (Worsley et al. 1992; Worsley et al. 1996) that adapts to spatial smoothness of the neuroimaging data. False discovery rate (FDR) or a customized critical threshold (e.g. GWAS significance level of $p < 5e-8$ or Bonferroni correction level), which are also available for SBM, can be applied to identify univariate genetic effects at different significance levels. More details about the statistical analyses are given in the Supplementary Methods.

Project Monitor

The Project Monitor section (Fig. 4f) of the GUI is a dashboard presenting the information related to each project the user created, including a unique project ID, creation time, name and status. In this panel, the user can choose to take multiple actions to monitor and manage their projects. First, the user can activate the Job Status panel (Fig. 4f-1) to check the information of jobs under a project, including the job IDs, starting and ending time, and execution status (submitted, backlogged, queued, running, completed or completed with errors). The real-time statuses of projects and jobs are automatically updated by communicating with the LONI pipeline in the background, and the status information shown on the GUI is updated with each page load through an AJAX¹² call to a PHP script. Second, the Project Monitor also provides navigation to the Project Editor for the user to create a new project or edit an existing project. Third, the user can choose to terminate or delete a project/job, or to download an archive of results if the project/job is completed successfully.

Result Viewer

When an association analysis is completed, the results are returned to the webserver from the LONI HPC server and visualized on the web-based Result Viewer (Fig. 4g). For an analysis in a selected project, the figures produced by Neuroimaging PheWAS are repeatedly displayed in a slideshow window. The figures include 3D brain maps in pre-defined views of genetic effects, uncorrected p -values, RFT and/or FDR corrected p -values for SBM analysis, Manhattan and Quantile-quantile (Q-Q) plots for univariate ROI-based analysis, and plots of LASSO paths and cross-validation and the summary table for multivariate ROI-based analysis. These displays can be easily enlarged with a mouse

¹¹ <https://rdrr.io/cran/MultiPhen>

¹² <https://developer.mozilla.org/en-US/docs/Web/Guide/AJAX>

click on the window. These resultant figures, maps and matrices are downloadable both as individual files, zipped job/project-folders from the Result Viewer or from the Project Manager.

Except for the system-produced figures and tables, a remarkable feature of the Result Viewer is the web-based, interactive viewer, which enables the user to manipulate result visualizations in real-time so as to help understanding and interpreting the complex, high-dimensional results. For SBM analyses, the system employs the JavaScript library – BrainBrowser¹³ v2.5.5 to visualize 3D surface data, which works in any modern web browser without requiring any browser plugins (Sherif et al. 2014). In an activated BrainBrowser viewer (Fig. 4g-1), the user can select any of the result vectors and render it on a 3D standard brain surface model with an adjustable colormap. The user can also adjust the color limits to threshold the map, freely rotate, pan and zoom the rendering surface, and take screenshots. Another key function of this viewer is that it allows the user to select any single point on the surface model and show its coordinates, vertex number, statistic and anatomical location (if an annotation file is provided). For univariate ROI-based analysis, the R library – manhattanly¹⁴ v0.2.0 is used to produce interactive Manhattan and Q-Q plots for the resultant statistics. In these interactive plots (Fig. 4g-2), the user can inspect the value of a specific data point (e.g. phenotype name and *p* value), zoom and pan the plot, extract a specific region as well as export the manipulated plot.

System Management and Data Security

Neuroimaging PheWAS, together with other LONI web services, is monitored by a central monitoring service,¹⁵ which notifies administrators if the system goes offline or crashes. Neuroimaging PheWAS and other LONI websites are also tested on individual servers that are load balanced using the monitoring software tool – Zabbix.¹⁶ The activities on the Neuroimaging PheWAS webserver are recorded into a server log, which provides administrators the reference to track and manage errors/bugs.

The cloud computing resource is managed by the LONI HPC grid engine and the LONI Pipeline in the background. The LONI grid engine has a first come, first served system in terms of resource allocation, with user limits in place on a variety of levels. By default, Neuroimaging PheWAS is allowed to execute, at the maximum, 768 jobs from all users at a time. On top of the grid engine, the LONI Pipeline allows more than 768 jobs to be submitted through the Neuroimaging

PheWAS system; the jobs exceeding the limit will be queued by the grid engine. Especially, to provide a fair share resource to users, Neuroimaging PheWAS sets another limit for the number of concurrent jobs that each user can execute (current limit is 48). Jobs submitted after the user's limit is reached will be backlogged and will be queued until any of the user's running jobs completes. These limits will be increased with the increase in the demand and usage of Neuroimaging PheWAS in the future. In particular, the user who needs a larger number of CPU cores (e.g. 500) to implement FreeSurfer processing for a big MRI data set (e.g. more than 1000 scans) may contact the system administrators¹⁷ to request this resource. Moreover, considering the runtimes needed for common analyses (see the results of benchmark testing below), any job running for over 24 h will be considered to be stuck or failed and will be terminated. To avoid an overload of the storage space, files of uploaded data and results that are stored for more than 30 days will be deleted.

A set of security measures is implemented to prevent data breaches. The user with a Neuroimaging PheWAS account is limited to the functionalities of the tool only and cannot access the LONI computational and storage infrastructure in the backend. Each user can only access (download or delete) the data that they have themselves uploaded and the results of their own projects through the web GUI and is not able to access other users' data and results. All user communications on the web GUI are secured using the HTTPS protocol. SSH login to the webserver, the computing environment and the data storage at LONI is restricted to system administrators only. Direct root login via SSH is only allowed from the LONI intranet that is secured by two Cisco Adaptive Security Appliances and is not allowed from the public Internet.

Case Study on APOE-Brain Associations

To demonstrate the usability of Neuroimaging PheWAS, we analyzed associations of APOE, the most established genetic risk factor for AD (Yamazaki et al. 2019), with various brain morphological properties across the brain. We queried the ADNI database and identified 1242 subjects with both MRI and genomic data from the ADNI study phases ADNI 1, ADNI GO and ADNI 2 (sample characteristics are summarized in Supplementary Table S1). The MRI acquisition and APOE genotyping protocols can be found on the ADNI website¹ and have been previously described elsewhere, e.g. (Roussotte et al. 2014; Wyman et al. 2013). Baseline MRI scans of the 1242 ADNI subjects, CSV tables of corresponding APOE ϵ 4 dosage and metadata were uploaded to the Neuroimaging PheWAS system as research data. The

¹³ <https://github.com/aces/brainbrowser>

¹⁴ <https://cran.r-project.org/web/packages/manhattanly/index.html>

¹⁵ <https://www.site24x7.com>

¹⁶ <https://www.zabbix.com/>

¹⁷ <http://phewas.loni.usc.edu/phewas/contact.php>

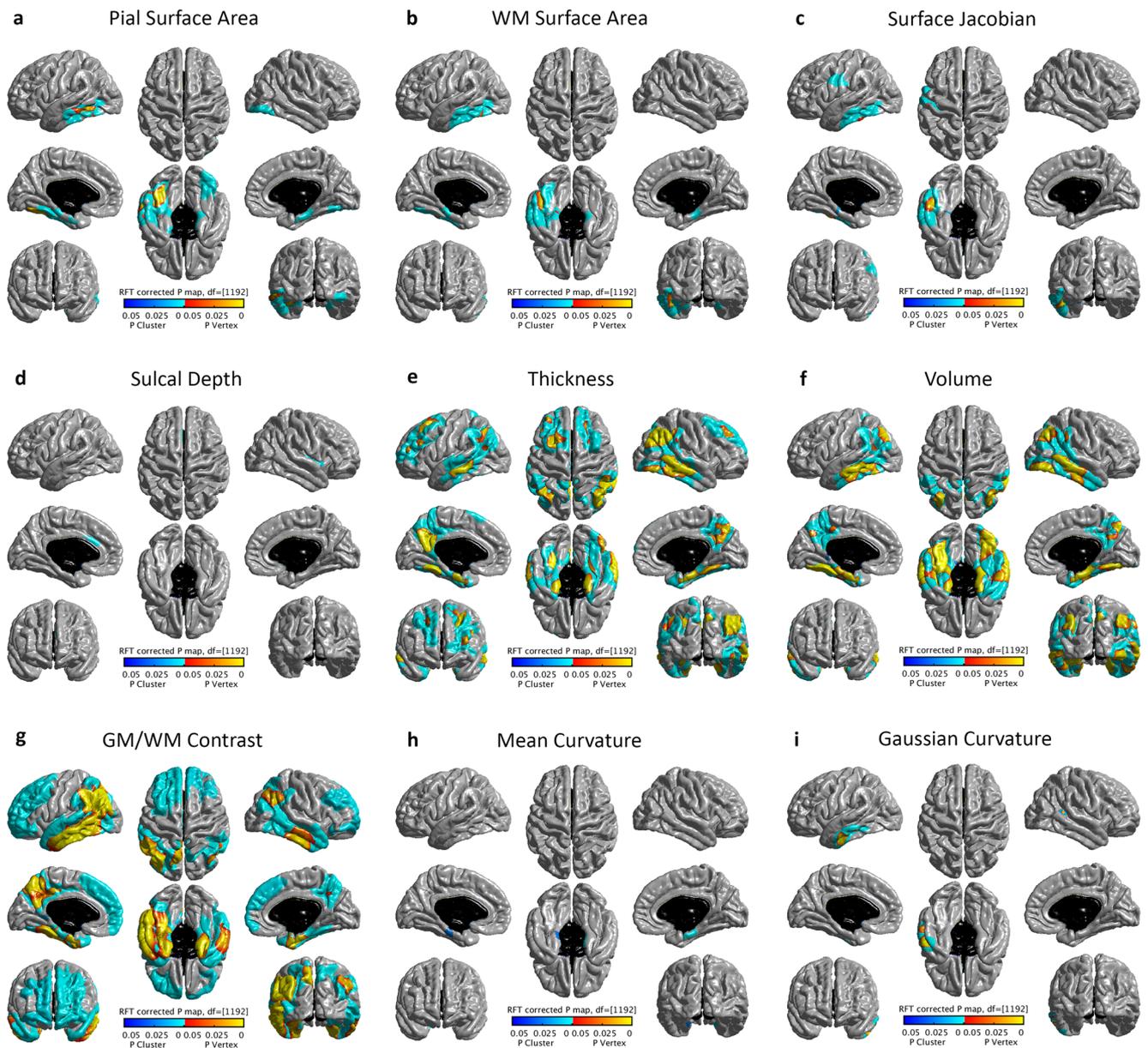


Fig. 5 Maps of APOE-brain associations that survived in RFT correction created by Neuroimaging PheWAS. Areas in blue-cyan represent cluster-level patterns, areas in red-yellow represent peak-level patterns

MRI scans were processed using FreeSurfer v6.0, and then APOE-brain associations were assessed on the surface-/parcellation-based brain morphological measures using SBM, univariate and multivariate ROI-based analysis controlling for age, sex, education, scanner and intracranial volume (ICV) as confounding factors.

Benchmark Testing

We conducted a series of benchmark tests to evaluate the performance of Neuroimaging PheWAS using a data set

extracted from the UK Biobank database² under the approved project 25,641. Details of this data set are available in (Zhao et al. 2019). Runtimes to implement univariate and multivariate ROI-Based analyses in the Neuroimaging PheWAS system were assessed using varying numbers of samples (from 100 to 8000) and phenotypes (from 100 to 3000). For SBM, the experiments were performed with varying number of samples (from 100 to 8000) only, due to the constraint of the surface model (327,684 vertices). Moreover, to evaluate the system's functions of data harmonization and cleansing, we artificially created data mismatching and missingness in

T_statistics

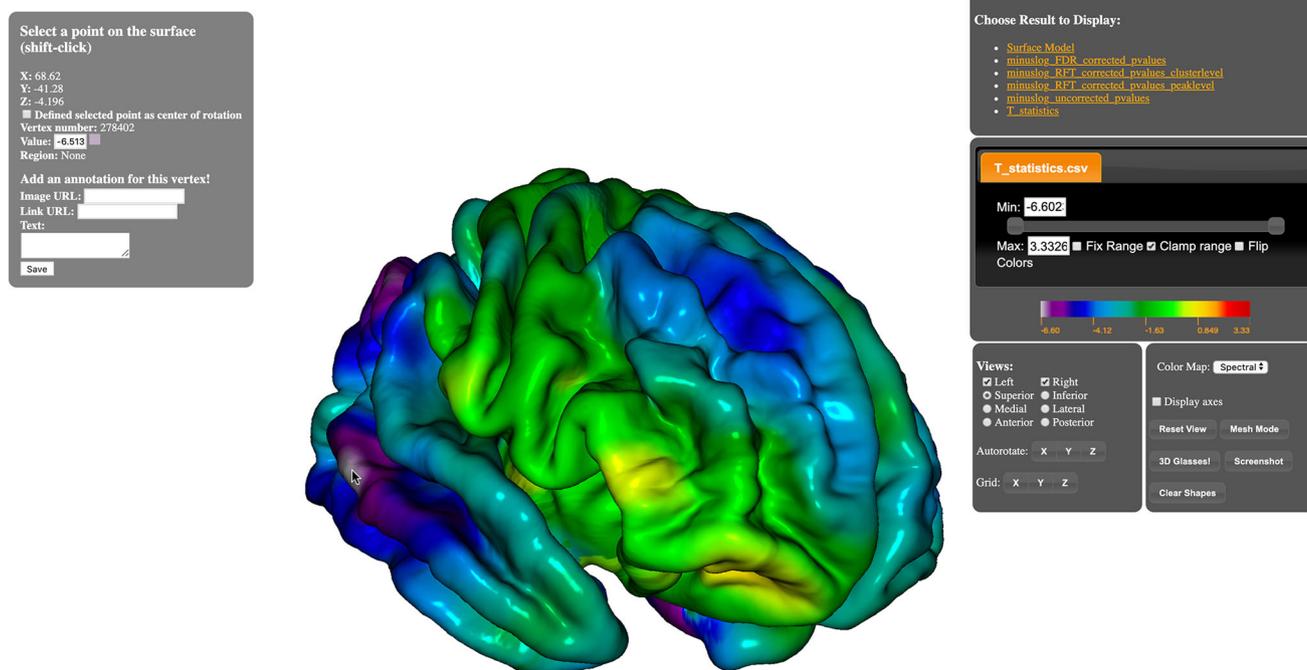


Fig. 6 Screenshot of the interactive viewer showing the T map of APOE effects on vertex-wise cortical thickness. The rendered surface can be freely rotated, panned and zoomed with mouse operations. The right panel

lists available result vectors, and allows changing colormap, color limits and display views. Any single point on the surface model can be selected, and the information of the point is shown on the left panel

harmonized, complete testing matrices by shuffling the samples, unbalancing the sample sizes, and creating missing values and duplicated variables. Experiments using such manipulated testing matrices as inputs were executed, and the processed matrices were examined to determine if the system is able to detect and control the known data flaws correctly. In addition, all the analyses in the case study on APOE-brain associations were repeated on different computational nodes of the LONI HPC Grid to examine the consistency of the outputs.

Results

MRI Processing

For the case study associating APOE genotype with the brain using the ADNI data set, 1235 FreeSurfer processing jobs were completed successfully and 7 jobs failed (runtime for a single job = 9.34 ± 6.53 h). Neuroimaging PheWAS automatically discarded the failed jobs and extracted vertex-wise measures of pial surface area, white matter (WM) surface area, mean curvature, Gaussian curvature, surface Jacobian, sulcal depth, cortical thickness, volume, gray matter (GM)/WM contrast from the output folders of the 1235 successful jobs, and

integrated these metrics across the subjects and saved them into 9 compressed binary files for each morphological property. The system also extracted ROI-based statistics of surface area, volume, thickness, standard deviation of thickness, mean curvature, Gaussian curvature, folding index, curvature index and GM/WM contrast from the 7 different parcellation atlases used in FreeSurfer (aseg, wmparc, aparc, aparc.a2009s, aparc.DKTatlas, BA_exvivo and BA_exvivo.thresh) and then integrated these across the subjects and saved them into 100 CSV files for separate hemispheres, different measures and atlases (see Supplementary Methods for details). These files were imported into the online data bag.

SBM Analysis

For the SBM analysis, we chose the APOE $\epsilon 4$ dosage as the target genotype, the 9 vertex-wise brain morphometric measures as the phenotypes, and selected age, sex, education, scanner, and ICV as covariates. Linear regression was chosen to test the genetic effects of APOE. RFT and FDR were used to control for the multiple comparisons across the brain (327,684 vertices). Maps of APOE-brain associations that survived in RFT correction are shown in Fig. 5. Cluster-level associations (cluster-level RFT corrected $p < 0.05$) were consistently found in the middle, inferior and medial temporal

Table 1 Table produced by Neuroimaging PheWAS summarizing ROI-based phenotypes associated with APOE identified using univariate analysis at the GWAS significance level of $p < 5e-8$. The phenotypes were sorted by p values. Phenotypes were named as <region>_<measure>_<atlas>. lh: left hemisphere, rh: right hemisphere

Phenotype	t value	p value
Right-Hippocampus_volume_aseg	-10.97	4.83E-27
Left-Hippocampus_volume_aseg	-10.66	1.04E-25
Right-Amygdala_volume_aseg	-8.98	5.34E-19
Left-Amygdala_volume_aseg	-8.65	8.25E-18
lh_perirhinal_exvivo_volume_BA_exvivo.thresh	-7.21	4.88E-13
rh_perirhinal_exvivo_volume_BA_exvivo.thresh	-6.92	3.71E-12
lh_perirhinal_exvivo_volume_BA_exvivo	-6.91	4.07E-12
lh_fusiform_volume_aparc.DKTatlas	-6.46	7.48E-11
lh_perirhinal_exvivo_thickness_BA_exvivo	-6.28	2.42E-10
lh_entorhinal_wg.pct	-6.26	2.71E-10
rh_perirhinal_exvivo_volume_BA_exvivo	-6.10	7.15E-10
lh_entorhinal_volume_aparc.DKTatlas	-6.04	1.03E-09
rh_perirhinal_exvivo_thickness_BA_exvivo.thresh	-6.02	1.17E-09
rh_entorhinal_wg.pct	-5.99	1.40E-09
rh_G_pariet_inf-Angular_thickness_aparc.a2009s	-5.95	1.72E-09
lh_inferiortemporal_gauscurv_aparc.DKTatlas	5.95	1.73E-09
lh_entorhinal_thickness_aparc.DKTatlas	-5.93	1.95E-09
rh_inferiorparietal_thickness_aparc	-5.91	2.26E-09
rh_inferiorparietal_thickness_aparc.DKTatlas	-5.90	2.36E-09
lh_perirhinal_exvivo_thickness_BA_exvivo.thresh	-5.89	2.52E-09
rh_middletemporal_volume_aparc.DKTatlas	-5.89	2.55E-09
rh_perirhinal_exvivo_thickness_BA_exvivo	-5.79	4.58E-09
SubCortGrayVol_volume_aseg	-5.77	5.07E-09
rh_entorhinal_thickness_aparc.DKTatlas	-5.75	5.68E-09
lh_entorhinal_thickness_aparc	-5.74	5.92E-09
rh_G_oc-temp_med-Parahip_thickness_aparc.a2009s	-5.74	6.12E-09
lh_G_oc-temp_lat-fusifor_volume_aparc.a2009s	-5.73	6.25E-09
rh_middletemporal_volume_aparc	-5.70	7.43E-09
lh_entorhinal_volume_aparc	-5.69	8.03E-09
rh_middletemporal_thickness_aparc.DKTatlas	-5.66	9.33E-09
rh_G_temporal_middle_thickness_aparc.a2009s	-5.65	1.01E-08
lh_inferiortemporal_gauscurv_aparc	5.63	1.10E-08
rh_G_oc-temp_med-Parahip_volume_aparc.a2009s	-5.59	1.42E-08
lh_S_front_sup_thickness_aparc.a2009s	-5.57	1.59E-08
rh_entorhinal_volume_aparc.DKTatlas	-5.57	1.59E-08
rh parahippocampal_volume_aparc	-5.57	1.60E-08
rh_S_interm_prim-Jensen_thickness_aparc.a2009s	-5.55	1.74E-08
lh_inferiorparietal_volume_aparc	-5.55	1.74E-08
rh_inferiorparietal_volume_aparc.DKTatlas	-5.54	1.91E-08
lh_middletemporal_wg.pct	-5.53	1.93E-08
rh_entorhinal_thickness_aparc	-5.53	1.97E-08
lh_fusiform_volume_aparc	-5.52	2.09E-08
rh parahippocampal_volume_aparc.DKTatlas	-5.49	2.41E-08
lh_inferiorparietal_volume_aparc.DKTatlas	-5.48	2.58E-08
rh_entorhinal_volume_aparc	-5.44	3.15E-08

Table 1 (continued)

Phenotype	t value	p value
lh_inferiorparietal_wg.pct	-5.40	3.97E-08
rh_middletemporal_thickness_aparc	-5.39	4.23E-08
lh_G_oc-temp_med-Parahip_thickness_aparc.a2009s	-5.38	4.57E-08
lh_G_precuneus_thickness_aparc.a2009s	-5.37	4.81E-08

cortices for diverse morphological measures except for sulcal depth. Additional cluster-level associations were detected in the left inferior motor cortex (IMC) for the surface Jacobian metric, in the left anterior cingulate cortex (ACC) and the right insula for sulcal depth, in the inferior parietal cortex (IPC) and the precuneus/posterior cingulate cortex (PCC) for thickness, volume and GM/WM contrast, and in the prefrontal cortex (PFC) for thickness and GM/WM contrast. In these patterns, the dosage of APOE $\epsilon 4$ was positively correlated with surface Jacobian and sulcal depth, whereas it was negatively correlated with the other measures (see the T maps shown in Supplementary Fig. S1). Peak-level RFT inference identified more localized APOE effects (peak-level RFT corrected $p < 0.05$) in the inferior and/or middle temporal cortex on pial and WM surface area, surface Jacobian, thickness, volume, GM/WM contrast and Gaussian curvature, in the medial temporal cortex (mTC), IPC, precuneus/PCC on thickness, volume and GM/WM contrast, and in the lateral dorsal prefrontal cortex (LDPFC) on thickness. FDR inference detected nearly all the cluster-level APOE effects with enlarged spatial extents, except for the APOE effect on mean curvature in the bilateral entorhinal cortices (EC) (Supplementary Fig. S2). Besides the system-produced 3D brain maps in pre-defined views as shown in Fig. 5 and Supplementary Fig. S1-S3, SBM results were visualized using the web-based, interactive viewer for manipulating and exploring the complex, high-dimensional results in real-time (see Fig. 6).

Univariate ROI-Based Analysis

For the univariate ROI-based analysis, the 100 CSV files of ROI-based statistics were integrated into a single matrix containing 3228 imaging-derived phenotypes (duplicated variables were automatically detected and discarded by the system). Effects of the APOE $\epsilon 4$ dosage on these phenotypes were assessed using linear regression controlling for age, sex, education, scanner and ICV as covariates. Significant univariate effects were detected using the GWAS ($p < 5e-8$) and Bonferroni ($p < 0.05/3228 = 1.55e-5$) significance levels. The univariate analysis identified 49 phenotypes that were significantly associated with APOE at the GWAS significance

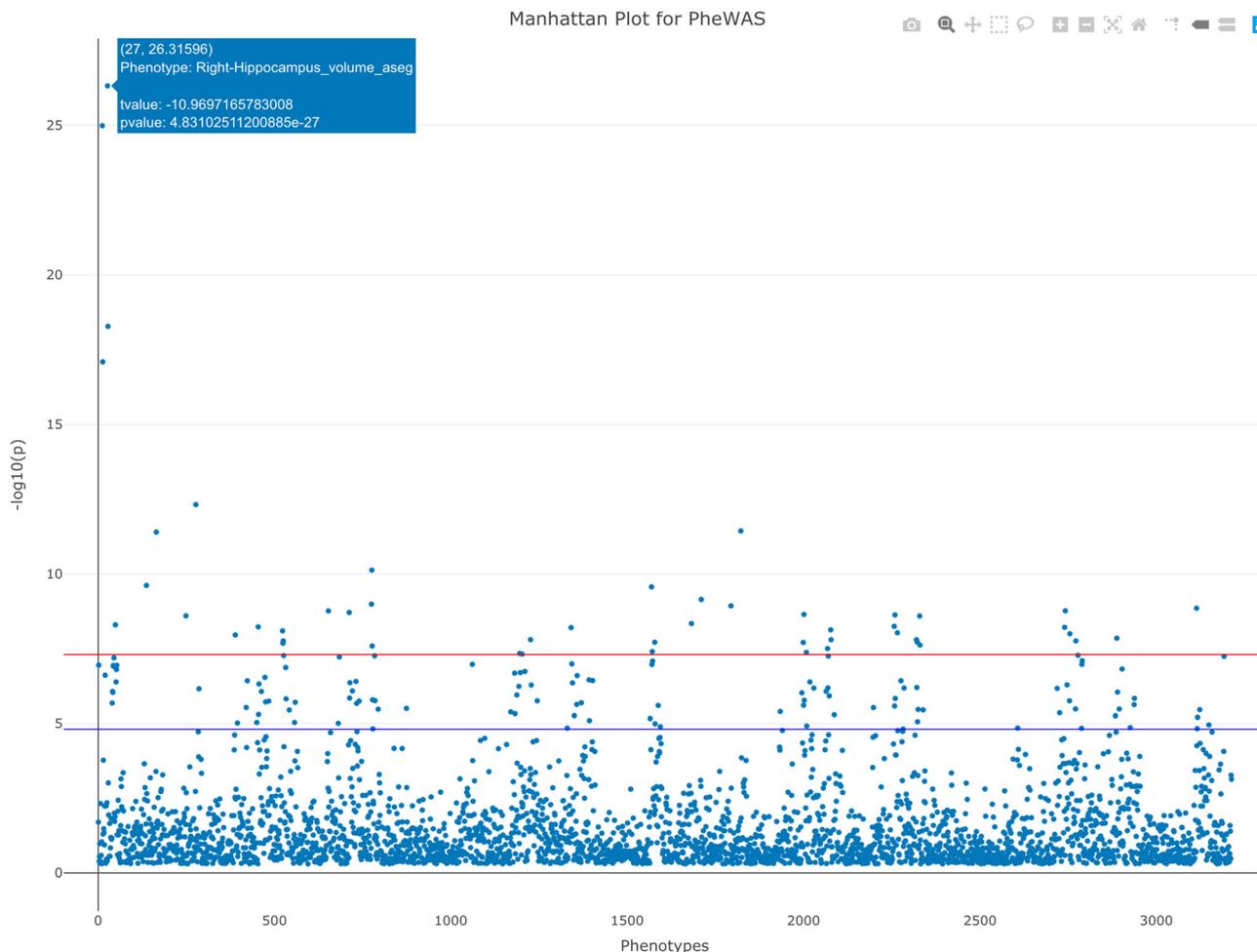


Fig. 7 Screenshot of the interactive Manhattan plot for ROI-based univariate analysis of APOE-brain associations. The red line represents the Bonferroni correction threshold of $p = 1.55e-5$. Annotations appear when hovering the mouse over a point. The blue line represents the GWAS significance threshold of $p = 5e-8$.

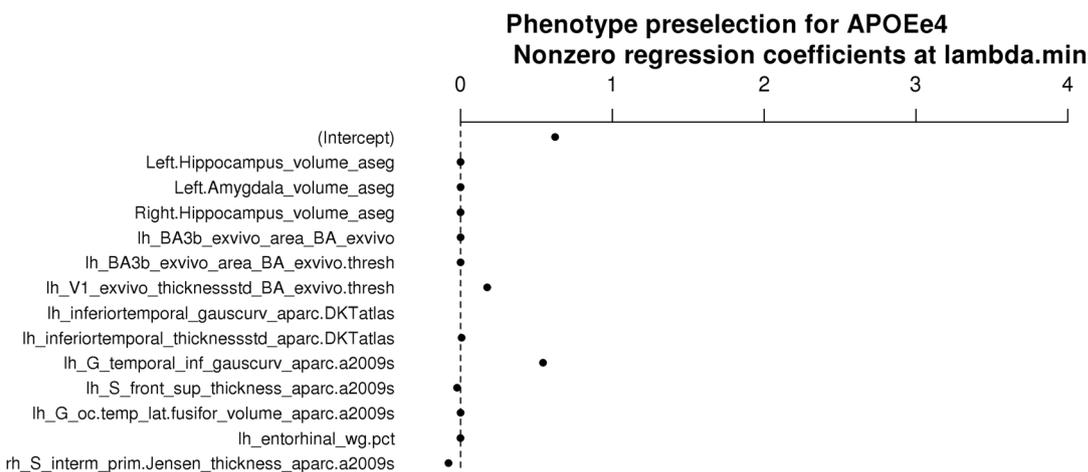


Fig. 8 Coefficient plot produced by Neuroimaging PheWAS summarizing the phenotypes preselected for multivariate association analysis and their LASSO regression coefficients at the tuning parameter Lambda that gave the minimum mean cross-validated error (Lambda.min)

Table 2 Table produced by Neuroimaging PheWAS summarizing the results of the Multivariate analysis for APOE-brain associations using the MultiPhen test. Phenotypes were named as <region>_<measure>_<atlas>. lh: left hemisphere, rh: right hemisphere

Phenotype	Coefficients	<i>p</i> value
Left.Amygdala_volume_aseg	1.11E-05	9.40E-01
Left.Hippocampus_volume_aseg	-2.01E-04	2.43E-01
Right.Hippocampus_volume_aseg	-1.73E-04	2.47E-01
lh_BA3b_exvivo_area_BA_exvivo	3.70E-04	4.53E-01
lh_BA3b_exvivo_area_BA_exvivo.thresh	2.62E-04	6.71E-01
lh_G_oc.temp_lat.fusifor_volume_aparc.a2009s	-7.22E-05	2.36E-01
lh_G_temporal_inf_gauscurv_aparc.a2009s	3.25E+00	5.56E-01
lh_S_front_sup_thickness_aparc.a2009s	-1.14E-01	5.22E-01
lh_V1_exvivo_thicknessstd_BA_exvivo.thresh	1.02E+00	1.66E-01
lh_entorhinal_wg.pct	-6.87E-03	3.67E-01
lh_inferiortemporal_gauscurv_aparc.DKTatlas	5.44E+00	4.52E-01
lh_inferiortemporal_thicknessstd_aparc.DKTatlas	5.59E-01	3.12E-01
rh_S_interm_prim.Jensen_thickness_aparc.a2009s	-2.19E-01	2.25E-01
JointModel	NA	5.25E-38

level of $p < 5e-8$ (Table 1). The Manhattan and Q-Q plots for the univariate analysis were presented in Fig. 7 and Supplementary Fig. S3 respectively. The phenotypes with the most significant p values were the volumes of the bilateral hippocampus ($p < 1e-25$) and amygdala ($p < 1e-17$). Other phenotypes significantly associated with APOE included the volume, thickness, GM/WM contrast and/or Gaussian curvature in the medial (perirhinal cortex, EC, parahippocampus and fusiform), middle and inferior temporal cortices, the thickness, volume and GM/WM contrast in IPC, the thickness in the left superior frontal cortex (SPC) and precuneus, and the total subcortical volume. In addition, 113 additional APOE-brain associations were significant when a Bonferroni correction ($p < 1.55e-5$) was applied, including, for example, the WM volume, mean curvature and surface area in the inferior and medial temporal regions, the volume and GM/WM contrast in the precuneus, and the thickness in the PCC (Fig. 6 and Supplementary Table S2).

Multivariate ROI-Based Analysis

To perform the multivariate association analysis, the system first removed the effects of confounding factors of age, sex, education, scanner and ICV from the values of the 3228 imaging-derived phenotypes using a linear regression. Next, LASSO was applied to the adjusted phenotypic metrics to preselect the phenotypes that have the best joint power to predict the APOE $\epsilon 4$ dosage (the LASSO path was shown in Supplementary Fig. S4). A 10-fold cross validation was conducted to determine the tuning parameter Lambda that gave the minimum mean cross-validated error (Lambda.min) (Supplementary Fig. S5). 13 phenotypes were preselected by LASSO at Lambda.min (Fig. 8), including the volume of the

left amygdala, the left and right hippocampus in the aseg atlas, the surface area of the left Brodmann area (BA) 3b in the BA_exvivo atlas, the surface area of the left Brodmann area (BA) 3b in the BA_exvivo.thresh atlas, the volume of the left lateral occipito-temporal gyrus, the Gaussian curvature of the left inferior temporal gyrus and the thickness of the left superior frontal sulcus in the aparc.a2009s atlas, the thickness standard deviation of the left V1 (primary visual area) in the BA_exvivo.thresh atlas, the GM/WM contrast of the left EC, the Gaussian curvature and the thickness standard deviation of the left ITC in the aparc.DKTatlas, and the thickness of the right sulcus intermedius primus of Jensen in the aparc.a2009s atlas. Then the MultiPhen test was performed to assess the joint relationship of these preselected phenotypes with the APOE $\epsilon 4$ dosage. Results of the MultiPhen test was summarized in Table 2. The 13 imaging phenotypes showed a predominant joint association with the APOE gene ($p = 5.25e-38$). It is not surprising that all these phenotypes did not show a significant individual association in the joint model ($p > 0.05$), as the significance for each individual variable was assessed considering all the other variables included in the MultiPhen model as covariates.

Benchmark Testing

The results of runtime evaluation show that Neuroimaging PheWAS took less than 2 min to complete the univariate ROI-based analysis for up to 3000 neuroimaging phenotypes and up to 8000 samples (Supplementary Fig. S6). The more complicated multivariate ROI-based analysis was completed within 22 min for up to 3000 phenotypes and up to 8000 samples (Supplementary Fig. S7). The SBM analysis on 327,684 vertices was completed within 4 min for 100 samples

and 24 min for up to 8000 samples (Supplementary Fig. S8). In the tests of data harmonization and cleansing, Neuroimaging PheWAS always correctly identified the artificially created data flaws (data mismatching and missingness, and duplicated variables), and accurately harmonized the samples, discarded the individuals that were not included in all the input matrices and/or had missing values, and dropped the duplicated variables. Furthermore, the outputs of the case study obtained on different computational nodes were consistent, demonstrating the output consistency of the system.

Discussion

We have introduced Neuroimaging PheWAS Version 1.0, a web-based system for implementing large-scale neuroimaging based PheWAS. The system offers the unique features of a user-friendly web GUI, a cloud-based HPC solution for processing large MRI data sets and implementing large-scale statistical gene-brain (or generic) association analyses, and interactive management and manipulation of data and results (Fig. 2).

The constant increase of publicly available genotype and phenotype data (Tryka et al. 2014) creates a demand for tools that enable researchers to handle the intensive computation in large-scale data processing and sophisticated association analyses that facilitate the identification of possible genotype-phenotype relationships at a population-level. For over a decade, GWAS has been the study design and statistical analysis of choice for genetic discovery (Horwitz et al. 2019; Visscher et al. 2017). In this study design, millions of common genetic variants are each tested for an association with a single or a small number of related outcomes or traits. A number of toolboxes and software packages have been developed for different facets of GWAS (Gumpinger et al. 2018). PLINK (Purcell et al. 2007) is, perhaps, the most widely used tool for genetic data processing and GWAS. It allows the user to perform various analyses on SNP data, such as univariate GWAS using two-sample tests and linear regression models, as well as set-based tests and epistasis screenings. In addition to PLINK, there are many other toolboxes that implement different association tests with linear mixed models, such as GCTA (Yang et al. 2011), FaST-LMM (Lippert et al. 2011), EMMAX (Kang et al. 2010), GEMMA (Zhou and Stephens 2012), and with network-based approaches for the joint test of multiple variants, such as SConES (Azencott et al. 2013), dmGWAS (Jia et al. 2011) and DAPPLE (Rossin et al. 2011). Apart from these downloadable software packages, some web-based GWAS tools have been developed, including Matapax (Childs et al. 2012), GWAPP (Seren et al. 2012) and easyGWAS (Grimm et al. 2017). These web applications enable the user to perform GWAS, analyze and annotate the results on a web server.

Previous findings of cross-phenotype associations of single genetic variants observed across candidate gene studies and GWASs were often attributed to the phenomenon of pleiotropy (Solovieff et al. 2013), i.e. a genetic variant or gene affects more than one distinct phenotype. A recent rapid rise in the identification of cross-phenotype associations has spurred interest in systematically identifying associations that form the basis of human pleiotropy (Tyler et al. 2016). PheWAS is the study approach systematically examining the impact of one or some specific genetic variants across a broad range of human phenotypes (Bush et al. 2016; Denny et al. 2016), and is gaining traction in the scientific community. In the mechanics of performing association tests, the PheWAS methodology is similar to the GWAS methodology. However, PheWAS approaches that investigate a wide range of phenotypes do not translate well to conventional GWAS software that is focused on a single phenotype typically. Performing PheWAS using existing GWAS software packages would require scripting many runs of the software. A few efforts have been made to develop PheWAS toolboxes. For example, Carroll and colleagues (Carroll et al. 2014) introduced R PheWAS - an R implementation of the most common functionality needed to perform and visualize ICD data-based PheWAS (Denny et al. 2013). A visualization software has also been developed to assist in presenting and investigating PheWAS results (Pendergrass et al. 2012). To date, PheWAS methods have been deployed primarily using electronic health record (EHR) billing code data (Denny et al. 2016). The recent establishment of neuroimaging genomic databases of large healthy and diseased cohorts (Bycroft et al. 2018; Jack Jr. et al. 2008; Jernigan et al. 2016; Satterthwaite et al. 2014; Thompson et al. 2010) empowers applications of PheWAS to neuroimaging data for new insights into the genetic pathways that shape the brain and the genetic mechanism in the underlying biological etiology of diseases (Hashimoto et al. 2015; Medland et al. 2014). However, the existing PheWAS tools cannot be applied to analyze neuroimaging data directly because of the fundamental differences in the underlying data (EHR vs. imaging). New tools for neuroimaging based PheWAS are required.

To our knowledge, Neuroimaging PheWAS is the first web-based tool that is designed to address the challenges in large-scale, brain-wide imaging PheWAS. The essential foundation for establishing such an advanced system is the unique IT infrastructure at LONI, which is designed and operated to facilitate modern informatics research (see Supplementary Methods for details). The resources provide networking, storage and computational capabilities that ensure a stable, secure and robust environment. These resources have been designed, built and continuously upgraded over the years by the LONI systems administration team. Thus, LONI has the appropriate expertise and operating procedures in place to utilize these resources to their maximum benefit. Based on such robust,

extensive IT infrastructure and skilled IT personnel, Neuroimaging PheWAS provides a cloud-based solution to perform complex, large-scale imaging genetics analyses on the LONI HPC grid. It does so while at the same time reducing the technical expertise required to use these resources. No computer programming skills are required, and it is not necessary to install any software. Second, Neuroimaging PheWAS offers a variety of statistical methods for neuroimaging-based association analyses. SBM explores gene-brain relationships across the entire brain with a high regional specificity (i.e., point by point), without requiring the a priori definition of particular ROIs. Especially, SBM addresses the spatial correlation across the brain using the RFT method (Worsley et al. 1992; Worsley et al. 1996) to correct for FWER. ROI-based analysis is more computationally efficient than SBM and has the benefit of a smaller number of multiple comparisons due to the lower spatial resolution, although it often neglects focal signals and may be biased by variation in the ROI definition. The user of Neuroimaging PheWAS can select the proper analysis based on their research needs and data types. Next, Neuroimaging PheWAS features a user-friendly, web-based GUI, which yields an easy way to manage data, define study protocols and monitor project/job status freeing the user from tedious command line works on their local computing environment. Additionally, the present web solution does not require preinstalled software environments such as R, Matlab and python in order to enable functionalities on a local host computer, whereas many existing GWAS/PheWAS tools require their preinstallation. Another salient feature of Neuroimaging PheWAS is the interactive result viewer that allows inspecting prominent association loci on the 3D brain maps for SBM or on the Manhattan and Q-Q plots for ROI-based analyses. Of note, although the system is essentially designed for associating genetic variants with neuroimaging-derived phenotypes, the user can also use it to implement generic association studies with the employed algorithms' capability to handle multiple variable types and the enriched options in defining the analysis protocol on the GUI. When creating a project for SBM or univariate ROI-based analysis, besides the genotype, the user can choose any variable (e.g. age, gender, or age \times gender and etc.) from the ones included in the statistical model to test its effects on studied phenotypes selecting an appropriate statistical test (i.e. linear regression, t-test, F-test and ANOVA). For the multivariate analysis, the user can use a non-genetic attribute (e.g. a demographic, behavioral, environmental or clinical factor) as the input for 'genotype' to assess the joint relationship of multiple phenotypes with it. Only SBM is a neuroimaging specific approach, whereas the ROI-based analysis can be applied to either neuroimaging-derived or non-imaging phenotypes or a mixture of them.

We demonstrated some of the potential of Neuroimaging PheWAS in a case study analyzing the associations of APOE

with various brain morphological properties across the brain in the ADNI cohort. The SBM and univariate ROI-based analysis consistently associated APOE ϵ 4 dosage with the thickness, volume, area, GM/WM contrast and Gaussian curvature of multiple limbic/paralimbic regions, well in line with previous single-phenotype imaging genetics data (Gutierrez-Galve et al. 2009; Saeed et al. 2018; Stage et al. 2016). The multivariate analysis also revealed a predominant joint relationship of 13 ROI-based phenotypes (mostly in the limbic/paralimbic system) with the genetic variation of APOE ($p < 1e-37$). Conducting such a systematical study without Neuroimaging PheWAS would be a time-consuming and cumbersome process. In addition, Neuroimaging PheWAS has been utilized in our recent study to examine the impacts of potential modifiers of normal aging (demographics, cognitive functions, lifestyle behaviors and specific genetic factors) on age-related brain morphological differences in ~8000 UK Biobank participants (Zhao et al. 2019).

There are several important future considerations that could potentially improve the Neuroimaging PheWAS system. First, the current system supports image processing for structural MRI only, and the neuroimaging-specific analysis (SBM) is limited to brain morphological data. Recently, multimodal neuroimaging methods have become an indispensable tool for neuroscientific research and clinical application (Biessmann et al. 2011). Multimodal data has been increasingly collected by recent large-scale neuroimaging genomic databases (Bycroft et al. 2018; Casey et al. 2018; Jack Jr. et al. 2008). In the future releases, we will embed other well-validated software libraries, such as FreeSurfer/FsFast,¹⁸ FreeSurfer/TRACULA,¹⁹ FreeSurfer/PetSurfer,²⁰ AFNI (Cox 1996) and/or FSL (Jenkinson et al. 2012), so as to expand the functionalities of Neuroimaging PheWAS to process other neuroimaging modalities, e.g. diffusion-weighted imaging (DWI), functional MRI and positron emission tomography (PET) scans, and to facilitate analysis of phenotypic data of brain connectomic and functional properties. Second, quantifying, controlling, and monitoring image quality is an essential prerequisite for ensuring the validity and reproducibility of neuroimaging data analyses. We also plan to link Neuroimaging PheWAS with our web-based brain image quality control (QC) system²¹ (Kim et al. 2019), which features a workflow for the assessment of various modality and contrast brain imaging data. This will enable the user to ensure the quality of their neuroimaging data and the data's validity in the subsequent analyses. Third, the current system requires the user to upload prepared vectors of genetic variants. This may limit the accessibility of Neuroimaging PheWAS to some

¹⁸ <https://surfer.nmr.mgh.harvard.edu/fswiki/FsFast>

¹⁹ <https://surfer.nmr.mgh.harvard.edu/fswiki/Tracula>

²⁰ <https://surfer.nmr.mgh.harvard.edu/fswiki/PetSurfer>

²¹ <https://qc.loni.usc.edu>

novice-level users who have no knowledge of genetic data preprocessing. A future improvement of this study is to integrate PLINK and other genetic data analysis software packages to enable marker-/sample-QC on genetic data (Anderson et al. 2010) and genetic marker extraction. Fourth, the multivariate analysis is currently not applicable to SBM in the system due to the high dimensionality of the vertex-wise data (327,684 vertices \times number of surface-based measures). There exist a few multivariate methods for surface-based data of deformation matrices (Shi et al. 2013; Wang et al. 2010; Worsley et al. 2004) and positional coordinates (Lyttelton et al. 2009), which essentially still are tests for a single kind of measure. Developing a computationally efficient multivariate approach to assess the joint effect across the vertices and various measures is another desirable direction of future work. Finally, comparing and combining findings across studies of different cohorts make it possible to identify credible, reproducible findings and increase statistical power in association mapping through meta-analyses (Thompson et al. 2017). Thus, we plan to improve the Project Editor and the interactive result viewer to feature post-PheWAS meta-analysis to compare statistics such as effect sizes across multiple studies and multimodality brain data.

Conclusions

Neuroimaging based PheWAS aims to scan integrative high-throughput imaging data of a wide variety of brain structural and functional properties for exploring relationships between candidate genes and the brain at a system level. Such research may assist the development of precision medicine for better understanding diseases, from genetic determinants to the genetic mechanism in the underlying biological etiology. The unprecedented scale and complexity of the imaging genomic data have introduced computational obstacles requiring new biomedical data science tools. Neuroimaging PheWAS is the first web-based tool to implement brain-wide imaging genetics analyses of large populations. The system provides the distinct features of a user-friendly GUI, a cloud-based computational infrastructure, multiple association analysis methods as well as interactive result visualizations. It meets the needs of end-users and enables researchers to focus on scientific questions both at the biological as well as the computational ends, without them having to possess extensive computational or storage infrastructure and programming expertise.

Information Sharing Statement

The Neuroimaging PheWAS system is freely available online at <http://phewas.loni.usc.edu/phewas/>.

Acknowledgements This research was conducted, using the UK Biobank Resource under approved project 25641. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Funding Information This work was supported by the Big Data for Discovery Science (BDDS) (NIH Grant No. U54EB020406), the Laboratory of Neuro Imaging Resource (LONIR) (NIH Grant No. P41EB015922), and the Genetic Influences on Human Neuroanatomical Shapes (NIH Grant No. R01MH094343).

Compliance with Ethical Standards

Conflict of Interests All the authors declare no biomedical financial interests or potential conflicts of interest regarding the publication of this paper.

References

- Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5, 1564–1573.
- Azencott, C. A., Grimm, D., Sugiyama, M., Kawahara, Y., & Borgwardt, K. M. (2013). Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29, i171–i179.
- Biessmann, F., Plis, S., Meinecke, F. C., Eichele, T., & Muller, K. R. (2011). Analysis of multimodal neuroimaging data. *IEEE Reviews in Biomedical Engineering*, 4, 26–58.
- Bigler, E. D., Skiles, M., Wade, B. S. C., et al. (2018). FreeSurfer 5.3 versus 6.0: Are volumes comparable? A chronic effects of Neurotrauma Consortium study. *Brain Imaging and Behavior*. <https://doi.org/10.1007/s11682-018-9994-x>.
- Bush, W. S., Oetjens, M. T., & Crawford, D. C. (2016). Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nature Reviews Genetics*, 17, 129–145.
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., & Marchini, J. (2018). The UK biobank

- resource with deep phenotyping and genomic data. *Nature*, *562*, 203–209.
- Carroll, R. J., Bastarache, L., & Denny, J. C. (2014). R PheWAS: Data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*, *30*, 2375–2376.
- Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., Charani, B., Mejia, M. H., Hagler Jr., D. J., Daniela Cornejo, M., Sicut, C. S., Harms, M. P., Dosenbach, N. U. F., Rosenberg, M., Earl, E., Bartsch, H., Watts, R., Polimeni, J. R., Kuperman, J. M., Fair, D. A., Dale, A. M., & Workgroup, A. I. A. (2018). The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, *32*, 43–54.
- Chard, K., D'Arcy, M., Heavner, B., Foster, I., Kesselman, C., Madduri, R., Rodriguez, A., Soiland-Reyes, S., Goble, C., Clark, K., Deutsch, E. W., Dinov, I., Price, N., Toga, A. (2016). I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets. 2016 IEEE international conference on big data (big data), 319–328.
- Chepkoech, J. L., Walhovd, K. B., Grydeland, H., Fjell, A. M., & Initiative, A.S.D.N. (2016). Effects of change in FreeSurfer version on classification accuracy of patients with Alzheimer's disease and mild cognitive impairment. *Human Brain Mapping*, *37*, 1831–1841.
- Childs, L. H., Lisek, J., & Walther, D. (2012). Matapax: An online high-throughput genome-wide association study pipeline. *Plant Physiology*, *158*, 1534–1541.
- Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173.
- Czajkowski, K., Kesselman, C., Schuler, R. (2017). ERMREST: A collaborative data catalog with fine grain access control. 2017 IEEE 13th international conference on E-science (E-science), 510–517.
- Dahnke, R., & Gaser, C. (2018). Surface and shape analysis. *Brain Morphometry*, *136*, 51–73.
- Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K., Wang, D., Masys, D. R., Roden, D. M., & Crawford, D. C. (2010). PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, *26*, 1205–1210.
- Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D., Field, J. R., Pulley, J. M., Ramirez, A. H., Bowton, E., Basford, M. A., Carrell, D. S., Peissig, P. L., Kho, A. N., Pacheco, J. A., Rasmussen, L. V., Crosslin, D. R., Crane, P. K., Pathak, J., Bielinski, S. J., Pendergrass, S. A., Xu, H., Hindorf, L. A., Li, R., Manolio, T. A., Chute, C. G., Chisholm, R. L., Larson, E. B., Jarvik, G. P., Brilliant, M. H., McCarty, C. A., Kullo, I. J., Haines, J. L., Crawford, D. C., Masys, D. R., & Roden, D. M. (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, *31*, 1102–1110.
- Denny, J. C., Bastarache, L., & Roden, D. M. (2016). Phenome-wide association studies as a tool to advance precision medicine. *Annual Review of Genomics and Human Genetics*, *17*, 353–373.
- Dinov, I. D., Van Horn, J. D., Lozev, K. M., Magsipoc, R., Petrosyan, P., Liu, Z., Mackenzie-Graham, A., Eggert, P., Parker, D. S., & Toga, A. W. (2009). Efficient, distributed and interactive Neuroimaging data analysis using the LONI pipeline. *Frontiers in Neuroinformatics*, *3*, 22.
- Dinov, I., Lozev, K., Petrosyan, P., Liu, Z., Eggert, P., Pierce, J., Zamanyan, A., Chakrapani, S., Van Horn, J., Parker, D. S., Magsipoc, R., Leung, K., Gutman, B., Woods, R., & Toga, A. (2010). Neuroimaging study designs, computational analyses and data provenance using the LONI pipeline. *PLoS One*, *5*, e13070.
- Dinov, I. D., Heavner, B., Tang, M., Glusman, G., Chard, K., Darcy, M., Madduri, R., Pa, J., Spino, C., Kesselman, C., Foster, I., Deutsch, E. W., Price, N. D., Van Horn, J. D., Ames, J., Clark, K., Hood, L., Hampstead, B. M., Dauer, W., & Toga, A. W. (2016). Predictive big data analytics: A study of Parkinson's Disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS One*, *11*, e0157077.
- Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 11050–11055.
- Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, *9*, 195–207.
- Glahn, D. C., Thompson, P. M., & Blangero, J. (2007). Neuroimaging endophenotypes: Strategies for finding genes influencing brain structure and function. *Human Brain Mapping*, *28*, 488–501.
- Grimm, D. G., Roqueiro, D., Salome, P. A., Kleeberger, S., Greshake, B., Zhu, W., Liu, C., Lippert, C., Stegle, O., Scholkopf, B., Weigel, D., & Borgwardt, K. M. (2017). easyGWAS: A cloud-based platform for comparing the results of genome-wide association studies. *Plant Cell*, *29*, 5–19.
- Gronenschild, E. H., Habets, P., Jacobs, H. I., Mengelers, R., Rozendaal, N., van Os, J., & Marcelis, M. (2012). The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. *PLoS One*, *7*, e38234.
- Gumpinger, A. C., Roqueiro, D., Grimm, D. G., & Borgwardt, K. M. (2018). Methods and tools in genome-wide association studies. *Methods in Molecular Biology*, *1819*, 93–136.
- Gutierrez-Galve, L., Lehmann, M., Hobbs, N. Z., Clarkson, M. J., Ridgway, G. R., Crutch, S., Ourselin, S., Schott, J. M., Fox, N. C., & Barnes, J. (2009). Patterns of cortical thickness according to APOE genotype in Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, *28*, 476–485.
- Hashimoto, R., Ohi, K., Yamamori, H., Yasuda, Y., Fujimoto, M., Umeda-Yano, S., Watanabe, Y., Fukunaga, M., & Takeda, M. (2015). Imaging genetics and psychiatric disorders. *Current Molecular Medicine*, *15*, 168–175.
- Hindorf, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 9362–9367.
- Horwitz, T., Lam, K., Chen, Y., Xia, Y., & Liu, C. (2019). A decade in psychiatric GWAS research. *Molecular Psychiatry*, *24*, 378–389.
- Jack Jr., C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P. J., Whitwell, J. L., Ward, C., Dale, A. M., Felmlee, J. P., Gunter, J. L., Hill, D. L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C. S., Krueger, G., Ward, H. A., Metzger, G. J., Scott, K. T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J. P., Fleisher, A. S., Albert, M., Green, R., Bartokis, G., Glover, G., Mugler, J., & Weiner, M. W. (2008). The Alzheimer's Disease Neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*, 685–691.
- Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). FSL. *Neuroimage*, *62*, 782–790.
- Jernigan, T. L., Brown, T. T., Hagler Jr., D. J., Akshoomoff, N., Bartsch, H., Newman, E., Thompson, W. K., Bloss, C. S., Murray, S. S., Schork, N., Kennedy, D. N., Kuperman, J. M., McCabe, C., Chung, Y., Libiger, O., Maddox, M., Casey, B. J., Chang, L., Ernst, T. M., Frazier, J. A., Gruen, J. R., Sowell, E. R., Kenet, T., Kaufmann, W. E., Mostofsky, S., Amaral, D. G., Dale, A. M., Pediatric Imaging, N., & Genetics, S. (2016). The Pediatric

- Imaging, Neurocognition, and Genetics (PING) data repository. *Neuroimage*, 124, 1149–1154.
- Jia, P., Zheng, S., Long, J., Zheng, W., & Zhao, Z. (2011). dmGWAS: Dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27, 95–102.
- Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42, 348–354.
- Kim, H., Irimia, A., Hobel, S. M., Pogoyan, M., Tang, H., Petrosyan, P., Blanco, R. E. C., Duffy, B. A., Zhao, L., Crawford, K. L., Liew, S. L., Clark, K., Law, M., Mukherjee, P., Manley, G. T., Van Horn, J. D., & Toga, A. W. (2019). The LONI QC system: A semi-automated, web-based and freely-available environment for the comprehensive quality control of Neuroimaging data. *Frontiers in Neuroinformatics*, 13, 60.
- Liao, K. P., Diogo, D., Cui, J., Cai, T., Okada, Y., Gainer, V. S., Murphy, S. N., Gupta, N., Mirel, D., Ananthakrishnan, A. N., Szolovits, P., Shaw, S. Y., Raychaudhuri, S., Churchill, S., Kohane, I., Karlson, E. W., & Plenge, R. M. (2014). Association between low density lipoprotein and rheumatoid arthritis genetic factors with low density lipoprotein levels in rheumatoid arthritis and non-rheumatoid arthritis controls. *Annals of the Rheumatic Diseases*, 73, 1170–1175.
- Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8, 833–835.
- Lytelton, O. C., Karama, S., Ad-Dab'bagh, Y., Zatorre, R. J., Carbonell, F., Worsley, K., & Evans, A. C. (2009). Positional and surface area asymmetry of the human cerebral cortex. *Neuroimage*, 46, 895–903.
- Mauch, V., Kunze, M., & Hillenbrand, M. (2013). High performance cloud computing. *Future Generation Computer Systems-The International Journal of Science*, 29, 1408–1416.
- Medland, S. E., Jahanshad, N., Neale, B. M., & Thompson, P. M. (2014). Whole-genome analyses of whole-brain data: Working within an expanded search space. *Nature Neuroscience*, 17, 791–800.
- Neuraz, A., Chouchana, L., Malamut, G., Le Beller, C., Roche, D., Beaune, P., Degoulet, P., Burgun, A., Lorient, M. A., & Avillach, P. (2013). Phenome-wide association studies on a quantitative trait: Application to TPMT enzyme activity and thiopurine therapy in pharmacogenomics. *PLoS Computational Biology*, 9, e1003405.
- O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M. R., & Coin, L. J. (2012). MultiPhen: Joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*, 7, e34861.
- Panizzon, M. S., Fennema-Notestine, C., Eyler, L. T., Jernigan, T. L., Prom-Wormley, E., Neale, M., Jacobson, K., Lyons, M. J., Grant, M. D., Franz, C. E., Xian, H., Tsuang, M., Fischl, B., Seidman, L., Dale, A., & Kremen, W. S. (2009). Distinct genetic influences on cortical surface area and cortical thickness. *Cerebral Cortex*, 19, 2728–2735.
- Pendergrass, S. A., Brown-Gentry, K., Dudek, S. M., Torstenson, E. S., Ambite, J. L., Avery, C. L., Buyske, S., Cai, C., Fesinmeyer, M. D., Haiman, C., Heiss, G., Hindorff, L. A., Hsu, C. N., Jackson, R. D., Kooperberg, C., Le Marchand, L., Lin, Y., Matise, T. C., Moreland, L., Monroe, K., Reiner, A. P., Wallace, R., Wilkens, L. R., Crawford, D. C., & Ritchie, M. D. (2011). The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic Epidemiology*, 35, 410–422.
- Pendergrass, S. A., Dudek, S. M., Crawford, D. C., & Ritchie, M. D. (2012). Visually integrating and exploring high throughput Phenome-wide association study (PheWAS) results using PheWAS-view. *BioData Mining*, 5, 5.
- Pendergrass, S. A., Brown-Gentry, K., Dudek, S., Frase, A., Torstenson, E. S., Goodloe, R., Ambite, J. L., Avery, C. L., Buyske, S., Buzkova, P., Deelman, E., Fesinmeyer, M. D., Haiman, C. A., Heiss, G., Hindorff, L. A., Hsu, C. N., Jackson, R. D., Kooperberg, C., Le Marchand, L., Lin, Y., Matise, T. C., Monroe, K. R., Moreland, L., Park, S. L., Reiner, A., Wallace, R., Wilkens, L. R., Crawford, D. C., & Ritchie, M. D. (2013). Phenome-wide association study (PheWAS) for detection of pleiotropy within the population architecture using genomics and epidemiology (PAGE) network. *PLoS Genetics*, 9, e1003087.
- Pontious, A., Kowalczyk, T., Englund, C., & Hevner, R. F. (2008). Role of intermediate progenitor cells in cerebral cortex development. *Developmental Neuroscience*, 30, 24–32.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81, 559–575.
- Rakic, P. (1988). Specification of cerebral cortical areas. *Science*, 241, 170–176.
- Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., & International Inflammatory Bowel Disease Genetics, C., Cotsapas, C., Daly, M. J. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genetics*, 7, e1001273.
- Roussotte, F. F., Gutman, B. A., Madsen, S. K., Colby, J. B., Thompson, P. M., & Alzheimer's Disease Neuroimaging, I. (2014). Combined effects of Alzheimer risk variants in the CLU and ApoE genes on ventricular expansion patterns in the elderly. *The Journal of Neuroscience*, 34, 6537–6545.
- Saeed, U., Mirza, S. S., MacIntosh, B. J., Herrmann, N., Keith, J., Ramirez, J., Nestor, S. M., Yu, Q., Knight, J., Swardfager, W., Potkin, S. G., Rogaeva, E., St George-Hyslop, P., Black, S. E., & Masellis, M. (2018). APOE-epsilon4 associates with hippocampal volume, learning, and memory across the spectrum of Alzheimer's disease and dementia with Lewy bodies. *Alzheimers Dement*, 14, 1137–1147.
- Satterthwaite, T. D., Elliott, M. A., Ruparel, K., Loughhead, J., Prabhakaran, K., Calkins, M. E., Hopson, R., Jackson, C., Keefe, J., Riley, M., Mentch, F. D., Sleiman, P., Verma, R., Davatzikos, C., Hakonarson, H., Gur, R. C., & Gur, R. E. (2014). Neuroimaging of the Philadelphia neurodevelopmental cohort. *Neuroimage*, 86, 544–553.
- Schuler, R.E., Kesselman, C., Czajkowski, K. (2016). Accelerating data-driven discovery with scientific asset management. Proceedings of the 2016 IEEE 12th international conference on E-science (E-science), 31–40.
- Seren, U., Vilhjalmsson, B. J., Horton, M. W., Meng, D., Forai, P., Huang, Y. S., Long, Q., Segura, V., & Nordborg, M. (2012). GWAPP: A web application for genome-wide association mapping in Arabidopsis. *Plant Cell*, 24, 4793–4805.
- Shen, L., Kim, S., Risacher, S. L., Nho, K., Swaminathan, S., West, J. D., Foroud, T., Pankratz, N., Moore, J. H., Sloan, C. D., Huentelman, M. J., Craig, D. W., DeChairo, B. M., Potkin, S. G., Jack Jr., C. R., Weiner, M. W., Saykin, A. J., & Alzheimer's Disease Neuroimaging, I. (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.
- Sherif, T., Kassis, N., Rousseau, M. E., Adalat, R., & Evans, A. C. (2014). BrainBrowser: Distributed, web-based neurological data visualization. *Frontiers in Neuroinformatics*, 8, 89.
- Shi, J., Wang, Y., Ceschin, R., An, X., Lao, Y., Vanderbilt, D., Nelson, M. D., Thompson, P. M., Panigrahy, A., & Lepore, N. (2013). A multivariate surface-based analysis of the putamen in premature newborns: Regional differences within the ventral striatum. *PLoS One*, 8, e66736.

- Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M., & Smoller, J. W. (2013). Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics*, *14*, 483–495.
- Stage, E., Duran, T., Risacher, S. L., Goukasian, N., Do, T. M., West, J. D., Wilhalme, H., Nho, K., Phillips, M., Elashoff, D., Saykin, A. J., & Apostolova, L. G. (2016). The effect of the top 20 Alzheimer disease risk genes on gray-matter density and FDG PET brain metabolism. *Alzheimers Dement (Amst)*, *5*, 53–66.
- Thompson, P. M., Martin, N. G., & Wright, M. J. (2010). Imaging genomics. *Current Opinion in Neurology*, *23*, 368–373.
- Thompson, P. M., Andreassen, O. A., Arias-Vasquez, A., Bearden, C. E., Boedhoe, P. S., Brouwer, R. M., Buckner, R. L., Buitelaar, J. K., Bulayeva, K. B., Cannon, D. M., Cohen, R. A., Conrod, P. J., Dale, A. M., Deary, I. J., Dennis, E. L., de Reus, M. A., Desrivieres, S., Dima, D., Donohoe, G., Fisher, S. E., Fouche, J. P., Francks, C., Frangou, S., Franke, B., Ganjgahi, H., Garavan, H., Glahn, D. C., Grabe, H. J., Guadalupe, T., Gutman, B. A., Hashimoto, R., Hibar, D. P., Holland, D., Hoogman, M., Pol, H. E. H., Hosten, N., Jahanshad, N., Kelly, S., Kochunov, P., Kremen, W. S., Lee, P. H., Mackey, S., Martin, N. G., Mazoyer, B., McDonald, C., Medland, S. E., Morey, R. A., Nichols, T. E., Paus, T., Pausova, Z., Schmaal, L., Schumann, G., Shen, L., Sisodiya, S. M., Smit, D. J. A., Smoller, J. W., Stein, D. J., Stein, J. L., Toro, R., Turner, J. A., van den Heuvel, M. P., van den Heuvel, O. L., van Erp, T. G. M., van Rooij, D., Veltman, D. J., Walter, H., Wang, Y., Wardlaw, J. M., Whelan, C. D., Wright, M. J., Ye, J., & Consortium, E. (2017). ENIGMA and the individual: Predicting factors that affect the brain in 35 countries worldwide. *Neuroimage*, *145*, 389–408.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B: Methodological*, *58*, 267–288.
- Toga, A. W. (Ed.). (2015). *Brain mapping: An encyclopedic reference*. Amsterdam: Academic.
- Toga, A. W., Foster, I., Kesselman, C., Madduri, R., Chard, K., Deutsch, E. W., Price, N. D., Glusman, G., Heavner, B. D., Dinov, I. D., Ames, J., Van Horn, J., Kramer, R., & Hood, L. (2015). Big biomedical data as the key resource for discovery science. *Journal of the American Medical Informatics Association*, *22*, 1126–1131.
- Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M., & Feolo, M. (2014). NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Research*, *42*, D975–D979.
- Tyler, A. L., Crawford, D. C., & Pendergrass, S. A. (2016). The detection and characterization of pleiotropy: Discovery, progress, and promise. *Briefings in Bioinformatics*, *17*, 13–22.
- Van Cauwenberghe, C., Van Broeckhoven, C., & Sleegers, K. (2016). The genetic landscape of Alzheimer disease: Clinical implications and perspectives. *Genetics in Medicine*, *18*, 421–430.
- Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *American Journal of Human Genetics*, *101*, 5–22.
- Wang, Y., Zhang, J., Gutman, B., Chan, T. F., Becker, J. T., Aizenstein, H. J., Lopez, O. L., Tamburo, R. J., Toga, A. W., & Thompson, P. M. (2010). Multivariate tensor-based morphometry on surfaces: Application to mapping ventricular abnormalities in HIV/AIDS. *Neuroimage*, *49*, 2141–2157.
- Whelan, C. D., Hibar, D. P., van Velzen, L. S., Zannas, A. S., Carrillo-Roa, T., McMahon, K., Prasad, G., Kelly, S., Faskowitz, J., deZubiracay, G., Iglesias, J. E., van Erp, T. G. M., Frodl, T., Martin, N. G., Wright, M. J., Jahanshad, N., Schmaal, L., Sämann, P. G., Thompson, P. M., & Initiative, A.s.D.N. (2016). Heritability and reliability of automatically segmented human hippocampal formation subregions. *Neuroimage*, *128*, 125–137.
- Winkler, A. M., Kochunov, P., Blangero, J., Almasy, L., Zilles, K., Fox, P. T., Duggirala, R., & Glahn, D. C. (2010). Cortical thickness or grey matter volume? The importance of selecting the phenotype for imaging genetics studies. *Neuroimage*, *53*, 1135–1146.
- Winkler, A. M., Sabuncu, M. R., Yeo, B. T., Fischl, B., Greve, D. N., Kochunov, P., Nichols, T. E., Blangero, J., & Glahn, D. C. (2012). Measuring and comparing brain cortical surface area and other areal quantities. *Neuroimage*, *61*, 1428–1443.
- Worsley, K. J., Evans, A. C., Marrett, S., & Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, *12*, 900–918.
- Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J., & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, *4*, 58–73.
- Worsley, K. J., Taylor, J. E., Tomaiuolo, F., & Lerch, J. (2004). Unified univariate and multivariate random field theory. *Neuroimage*, *23*(Suppl 1), S189–S195.
- Wyman, B. T., Harvey, D. J., Crawford, K., Bernstein, M. A., Carmichael, O., Cole, P. E., Crane, P. K., DeCarli, C., Fox, N. C., Gunter, J. L., Hill, D., Killiany, R. J., Pachai, C., Schwarz, A. J., Schuff, N., Senjem, M. L., Suhy, J., Thompson, P. M., Weiner, M., Jack Jr., C. R., & Alzheimer's Disease Neuroimaging, I. (2013). Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimers Dement*, *9*, 332–337.
- Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C. C., & Bu, G. (2019). Apolipoprotein E and Alzheimer disease: Pathobiology and targeting strategies. *Nature Reviews. Neurology*, *15*, 501–518.
- Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, *88*, 76–82.
- Zhao, L., Boucher, M., Rosa-Neto, P., & Evans, A. C. (2013). Impact of scale space search on age- and gender-related changes in MRI-based cortical morphometry. *Human Brain Mapping*, *34*, 2113–2128.
- Zhao, L., Matloff, W., Ning, K., Kim, H., Dinov, I. D., & Toga, A. W. (2019). Age-related differences in brain morphology and the modifiers in middle-aged and older adults. *Cerebral Cortex*, *29*, 4169–4193.
- Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*, 821–824.