Towards Automated Hypothesis Testing in Neuroscience

Daniel Garijo*[1], Shobeir Fakhraei*[1], Varun Ratnakar[1], Qifan Yang[3], Hanna Endrias[2], Yibo Ma[1], Regina Wang[1], Michael Bornstein[3], Joanna Bright[3], Yolanda Gil[1,2], and Neda Jahanshad[3]

[1] Information Sciences Institute, University of Southern California
{dgarijo,shobeir,varunr,yiboma,gil}@isi.edu
[2] Department of Computer Science, University of Southern California
{endrias}@usc.edu
[3] Imaging Genetics Center, Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC
{qifan.yang,mbornste,joannabr,neda.jahanshad}@usc.edu

Abstract. Scientific data generation in the world is continuous. However, scientific studies once published do not take advantage of new data. In order to leverage this incoming flow of data, we present Neuro-DISK, an end-to-end framework to continuously process neuroscience data and update the assessment of a given hypothesis as new data become available. Our scope is within the ENIGMA consortium, a large international collaboration for neuro-imaging and genetics whose goal is to understand brain structure and function. Neuro-DISK includes an ontology and framework to organize datasets, cohorts, researchers, tools, working groups and organizations participating in multi-site studies, such as those of ENIGMA, and an automated discovery framework to continuously test hypotheses through the execution of scientific workflows. We illustrate the usefulness of our approach with an implemented example.

Keywords: Hypothesis Evaluation, Scientific Workflow, Ontology, Automated Discovery, Neuroscience

1 Introduction

Scientific discoveries are based on hypothesis testing and rigorous data analysis. Such analyses are often time consuming and include steps that are difficult to interpret from scientific publications, and therefore, hard to systemically reproduce. Often, the designed hypothesis is tested only once against the acquired data sample and later archived. Interestingly, in empirical sciences such as the biological sciences, it is not uncommon for a hypothesis to yield contradictory results when evaluated on different data samples. In our data-driven world, data that may be potentially relevant for testing a hypothesis is being continuously

* co-first author

generated but is often not studied to its full potential for hypothesis re-evaluation in combination with other related data. The lack of an integrated system to constantly monitor the hypothesis of interest and update the underlying analysis when new data become available, is one of the challenges for automatic hypothesis re-evaluation. Having a framework that can keep such hypotheses alive requires systematically capturing the knowledge about the data and analytics involved in the hypothesis testing, which is often heterogeneous and compartmentalized.

In this paper, we propose a solution to address the above challenges in the neurosciences based on our previous work for Automated DIscovery of Scientific Knowledge (DISK) [1]. We have extended DISK to explore brain-aging related hypothesis and data by generalizing the ability for the system to connect to external knowledge bases, including projects available within the Enhancing Neuro Imaging Genetics through Meta-Analysis (ENIGMA)[4] consortium [2], a neuroscience collaboration where projects span many contributors from different institutions around the world. In our proposed solution we address challenges of *data*, *analytics*, and *hypothesis* complexity. The *data* shared through imaging initiatives such as the ENIGMA consortium includes multiple levels of heterogeneity, and are regularly expanding in volume. The *analytics* related to such data requires the use of dozens of interconnected tools, each of which may require substantial domain knowledge. The underlying *hypotheses* may depend on a range of possible multi-modal technical, neurological, clinical, demographic, and genetic data which could be collected across multiple datasets.

2 Related Work

Two closely related research areas in machine learning are online algorithms [3] (algorithms that revise their models when new data become available), and data-stream specific models [4] (that deal with challenges of reprocessing portions of prior data to scale to large data streams). A major advantage of our work over these methods is that our analytical steps do more than learning from data. For example, some of our steps may include integrating the relevant cohort properties. Another important difference is that our system can react when new kinds of data become available and invoke new analytic tools or algorithms different from the original ones. In addition, distinctive to active agents such as Robot Scientist [5], our method simply listens and reacts to the data that others collect. Moreover, in contrast to other hypothesis evaluation solutions, such as EXPO [6] and HELO [7], our approach represents supporting evidence for hypotheses as reproducible computational components, records their evolution in reaction to new data, and updates their confidence intervals.

3 Background

In this section we describe our domain of focus and the sub-components that we leverage to develop our solution.

[4] http://enigma.usc.edu

3.1 The ENIGMA Consortium

The ENIGMA consortium [8] is an international network connecting researchers in imaging genomics, neurology and psychiatry, in order to understand brain structure and function, based on multi-modal imaging and genetic data collected from various patient populations. One of the major ambitions of the consortium is to combine various datasets made available via its international partners into larger samples necessary to detect minute gene effects on complex traits that are otherwise not confidently identifiable with smaller isolated samples. Major goals of ENIGMA network include: creating a network of scholars with similar interests in brain imaging, genetics, neuro-psychiatry, and ensuring reproducibility of major findings through member collaborations, while facilitating information, algorithms and data sharing.

Members of the consortium constantly share new datasets and/or results, and run experiments and analysis across all available related data. The challenges involved in this global and dynamic collaborative platform, highlights a need to systematically organize its heterogeneous resources to facilitate identification and retrieval of entities of interest. The ENIGMA network would also benefit from a solution to capture the hypotheses under investigation by its members and their related analysis workflows to make them reproducible, especially if such solution could automatically find the related data and dynamically update the analysis results when new data become available. In this paper we layout the overall architecture and components of such solution for the ENIGMA consortium and report on our developed prototype.

3.2 The Organic Data Science Platform

We use the Organic Data Science framework (ODS) [9] and managing information about ENIGMA (ENIGMA-ODS). ODS is built on Semantic MediaWiki, which uses W3C standards such as RDF and SPARQL to represent its contents in a structured manner. Each wiki page represents a different resource (e.g., a researcher, a project, an organization, etc.) and shows the most relevant properties of that resource's class. For example, the wiki page of an organization will have *name* and *address* properties. Wiki pages can be filled out by users, who may contribute to the population and curation of the ENIGMA-ODS knowledge base.

ENIGMA-ODS is structured based on the ENIGMA Ontology[5] [10], which extends standard vocabularies such as Schema.org[6] and includes a representation for datasets, cohorts, persons, organizations, protocols, instruments, software and working groups together with their more common relationships. However, users may extend the ontology with their own properties and categories whenever necessary. Each dataset has a set of metadata assertions, defined in triples of the form $< subject, property, value >$, where the *subject* identifies the resource being described (e.g., a dataset), the *property* refers to the aspect of the subject

[5] https://w3id.org/enigma
[6] http://schema.org/

we want to describe (e.g., creation date) and the *value* identifies the value of the property for a resource (e.g., creation date is 2-2-2020). The data catalog supports W3C SPARQL queries to specify the desired metadata properties of datasets.

3.3 The DISK Framework

DISK [1, 11] is a framework designed to test and revise hypotheses via automatic analysis of dynamic scientific data. DISK evaluates and revises an input hypothesis via continuously examining related data as they become available. It also triggers new kinds of analyses and workflows with the availability of new kinds of data, tracking the provenance of revised hypothesis and its related details. DISK operates based on the description of available ODS metadata, expressed using domain ontologies with the W3C OWL and RDF Semantic Web standards.

A user defines the hypothesis of interest through the DISK GUI. To evaluate a hypothesis, DISK relies on a library of *Lines of Inquiry* (LOI). A Line of Inquiry includes a hypothesis pattern, a relevant data query pattern, a set of workflows to process that data and one or more meta-workflows to combine workflow results and generate revised confidence values or hypotheses. If a user hypothesis matches the hypothesis of a Line of Inquiry, the system will use the LOI query pattern to search for appropriate data to pass to the LOI workflows for execution.

Workflows are executed via WINGS [12], a semantic workflow system for designing scientific computational experiments that specifies the steps and configuration of data processing by software components. The execution results and their corresponding provenance trace are then stored in a Linked Data repository. Finally, the associated meta-workflows explore this repository and revise the original hypothesis, if necessary. DISK was demonstrated for canceromics, and this paper introduces new extensions for neuroscience [1, 11].

4 The Neuro-DISK Framework

We have extended DISK for neuroscience data exploration, analysis execution, and hypothesis testing. The framework integrates the ENIGMA-ODS platform for data search, which has access to all available information from datasets, cohorts, protocols and working groups. Our extension enables newly added and curated datasets to ENIGMA-ODS to be used in assessing existing or new Lines of Inquiry. We validate our framework by testing the hypothesis: *"Is the effect size of the number of APOE4 alleles on Hippocampus volume associated with the age of the cohort?"*. This hypothesis is important in Alzheimer's disease (AD) studies, which is the most common neuro-degenerative disorder and severely impacts patients' daily behaviors, thinking, and memory over a wide range of ages [13]. The hippocampus, the brain's memory hub, has been shown to be particularly vulnerable to Alzheimer's disease pathology, and is already atrophied by the time clinical symptoms of AD first appear [14]. The e4 haplotype (set of two alleles)

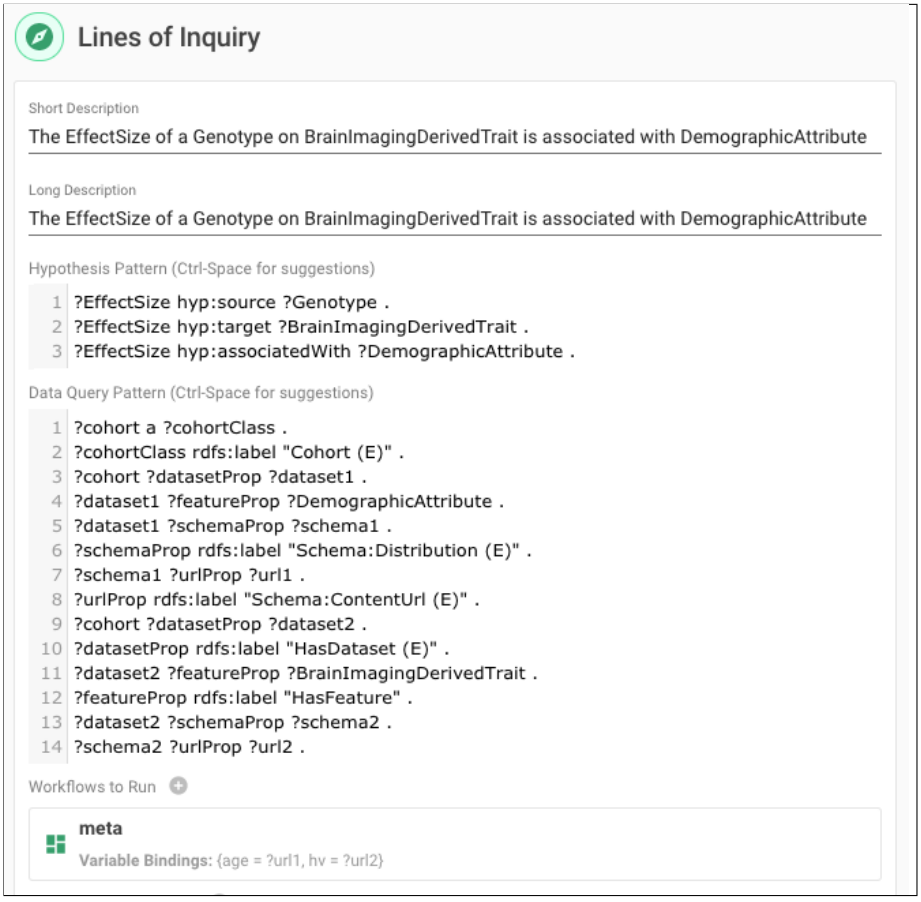

Fig. 1. An example of a line of inquiry for assessing the association between the effect size of a genotype on a brain-imaging derived trait for a particular cohort (or study population), with a meta-level demographic attribute such as age.

of the *APOE* (apolipoprotein E) gene, is the most significant single genetic risk factor for late-onset Alzheimer's disease [15]. At each of two positions in the genome, a possible e4 allele contributes to this genetic risk. However, there have been inconsistent findings in determining whether the e4-risk factor contributes to differences in brain structure, particularly that of hippocampal volume. Several imaging-genetic studies have found a significant correlation between this major genetic risk factor for Alzheimer's disease, and higher rates of hippocampal volume loss [16], while others have found no correlation with volume [17]. Here, by using a meta-regression design, we investigate whether findings attempting to relate APOE4 genotype and hippocampal volume, specifically the effect sizes associated with studies, may be due to the age of the cohorts being studied, and a function of the study sample-sizes.

Our hypothesis triggers the Line of Inquiry shown in Figure 1, which studies the correlation between *an effect on a brain characteristic* and *a demographic attribute*. This hypothesis meets the requirements listed in the *Hypothesis Pattern* section of Figure 1, i.e., *APOE4* being the genotype of interest, *Hippocampal volume* a brain imaging derived trait and *age* a meta-level demographic attribute, describing the average age of the cohort. Once the hypothesis pattern is met, Neuro-DISK will aim to find the appropriate datasets to run the workflows associated with the LOI. DISK uses the information under the *Data Query Pattern* section to issue a SPARQL query to the ENIGMA-ODS platform. The query pattern aims to retrieve the dataset URLs (*schema:contentURL*) belonging to the same cohort that contain the target brain characteristic and demographic value. DISK then uses the resulting data URLs as input to the associated workflow in the LOI (i.e., the "*meta*" workflow in Figure 1). The workflow consists of a sample-sized weighted meta-regression to determine whether the magnitude of the target genetic (APOE4) effect on a phenotype, is driven by the target demographic (age).

The underlying data for this analysis was based on imaging phenotypes and genotypes obtained from publicly available international cohorts, including ADNI-1, ADNI-2, DLBS, and the UK Biobank (application ID 15599). To configure the workflow, we incorporated the data from these independent cohorts with brain imaging and APOE4 genotype information. For each cohort, we ran a fixed-effects linear regression to associate the subjects' number of APOE4 risk-alleles (0, 1, or 2) with the mean bilateral hippocampal volumes derived from Freesurfer v5.3 [18]. Age, sex, and intracranial volume (to control for overall head size) were included as covariates in the regression. The resulting beta-value or un-standardized regression-coefficient and its corresponding standard error, were used to generate a standardized z-score for each cohort; the z-score was then regressed against the mean age of each cohort for the meta-regression, as was done in [19] for genome-wide significant findings. We note that given the sample size of UK Biobank (approximately 10,000 sample points at the time of writing, we split the data according to 5-year age bins). DLBS also had a wide age range from 30 to over 80, so that dataset was split into one younger than 60, and another older than 60 (a roughly even split) for this demonstration.

Figure 2 shows the results of our meta-regression analysis, automatically generated via the Neuro-DISK framework. In this proof of principle analysis with a handful of public datasets, age showed a negative association with the APOE4 effect size on hippocampal volume; should this association hold with more data points, it would suggest that the association between the APOE4 genotype and hippocampal volume may be driven by cohorts of individuals with older mean ages, therefore explaining why some studies may not find a significant effect of the most well known Alzheimer's disease risk genotype, with the most well-accepted brain-MRI derived biomarkers for Alzheimer's disease.

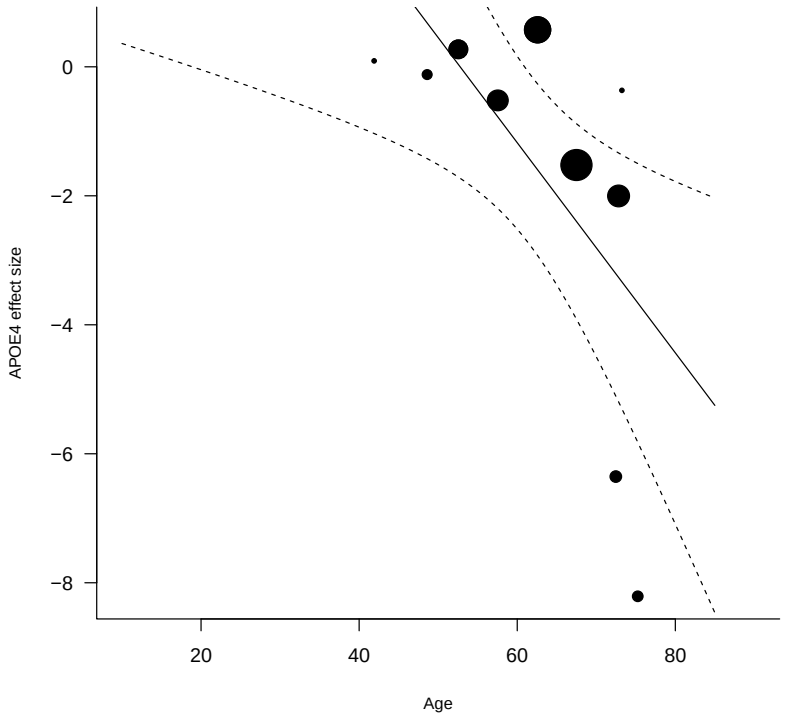

Fig. 2. Meta-regression for age and the effect size of Alzheimer's disease related risk genotype on hippocampal volume (p=0.011). Age is negatively associated with the APOE4 effect size on MRI-derived hippocampal volume. The size of the points are proportional to cohort size, and dashed lines indicate confidence intervals.

5 Conclusions and Future Work

In this paper we described Neuro-DISK, a framework to automatically test hypotheses in the neuroscience domain, specifically in the context of the ENIGMA and international consortium. Our framework integrates the ENIGMA-ODS platform, allowing further testing on previous hypotheses whenever a user contributes new datasets in the system. Note that currently a single hypothesis was tested, and the corresponding variables that were incorporated in the system were selected *a priori*. However, in cases when multiple variables are selected, such as multiple genetic markers, or multiple brain regions, in the same Line of Inquiry, standard multiple comparisons correction techniques including the false discovery rate adjustment are conducted.

Neuro-DISK is still in development, but our current work shows the potential for continuous hypothesis testing in this domain. In this paper, we only used data from four publicly available cohorts. However, as multisite studies are conducted on a larger scale in ENIGMA and other international consortia, upwards of 50 cohorts may be included for evaluating such hypotheses [19]. We are working towards addressing three main challenges: 1) improving synchronization between ENIGMA-ODS and Neuro-DISK to make the system more adaptive to triggering all compatible Lines of Inquiry with the addition of new datasets; 2) designing the query patterns to make them more accessible for users without SPARQL knowledge; and 3) automatically evaluating additional hypotheses based on generated workflow results.

Acknowledgments

We are grateful to the KAVLI foundation for their support of ENIGMA Informatics (PIs: Jahanshad and Gil). We also acknowledge support from the National Science Foundation under awards IIS-1344272 (PI: Gil), ICER-1541029 (Co-PI: Gil), and IIS-1344272 (PI: Gil), and from the National Institutes of Healths Big Data to Knowledge Grant U54EB020403 for support for ENIGMA (PI: Thompson) and High resolution mapping of the genetic risk for disease in the aging brain grant R01AG059874 (PI:Jahanshad). Data used in preparing this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database[7], phases both 1 and 2. As such, many investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators is available online[8]. We also used the DLBS[9] dataset and the UK Biobank in this study. This research was conducted using the UK Biobank Resource under Application Number 11559.

Bibliography

[1] Yolanda Gil, Daniel Garijo, Varun Ratnakar, Rajiv Mayani, Ravali Adusumilli, Hunter Boyce, and Parag Mallick. Automated hypothesis testing with large scientific data repositories. In *Proceedings of the Fourth Annual Conference on Advances in Cognitive Systems (ACS)*, 2016. 2, 4

[2] Paul Thompson, Neda Jahanshad, Christopher R K Ching, Lauren Salminen, Sophia I Thomopoulos, Joanna Bright, Bernhard T Baune, Sara Bertoln, Janita Bralten, Willem B Bruin, and et al. ENIGMA and Global Neuroscience: A Decade of Large-Scale Studies of the Brain in Health and Disease across more than 40 Countries, Jul 2019. 2

[7] adni.loni.usc.edu

[8] http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[9] http://fcon_1000.projects.nitrc.org/indi/retro/dlbs.html

[3] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. 2

[4] Joao Gama. A survey on learning from data streams: current and future trends. *Progress in Artificial Intelligence*, 1(1):45–55, 2012. 2

[5] Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, et al. The automation of science. *Science*, 2009. 2

[6] Larisa N Soldatova and Ross D King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11):795–803, 2006. 2

[7] Larisa N Soldatova, Andrey Rzhetsky, Kurt De Grave, and Ross D King. Representation of probabilistic scientific knowledge. In *Journal of Biomedical Semantics*, volume 4, page S7. BioMed Central, 2013. 2

[8] Paul M Thompson, Jason L Stein, Sarah E Medland, Derrek P Hibar, Alejandro Arias Vasquez, Miguel E Renteria, Roberto Toro, Neda Jahanshad, Gunter Schumann, Barbara Franke, et al. The ENIGMA consortium: large-scale collaborative analyses of neuroimaging and genetic data. *Brain Imaging and Behavior*, 8(2):153–182, 2014. 3

[9] Yolanda Gil and Varun Ratnakarand Paul C. Hanson. Organic data publishing: A novel approach to scientific data sharing. In *Second International Workshop on Linked Science: Tackling Big Data (LISC), held in conjunction with ISWC*, Boston, MA, 2012. 3

[10] M Jang, Tejal Patted, Yolanda Gil, Daniel Garijo, Varun Ratnakar, Jie Ji, Prince Wang, Aggie McMahon, Paul M Thompson, and Neda Jahanshad. Towards automatic generation of portions of scientific papers for large multi-institutional collaborations based on semantic metadata. In *CEUR workshop proceedings*, volume 1931, pages 63–70, 2017. 3

[11] Yolanda Gil, Daniel Garijo, Varun Ratnakar, Rajiv Mayani, Ravali Adusumilli, Hunter Boyce, Arunima Srivastava, and Parag Mallick. Towards continuous scientific data analysis and hypothesis evolution. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 4

[12] Yolanda Gil, Varun Ratnakar, Jihie Kim, Pedro Gonzalez-Calero, Paul Groth, Joshua Moody, and Ewa Deelman. Wings: Intelligent workflow-based design of computational experiments. *IEEE Intelligent Systems*, 26(1), 2010. 4

[13] Alzheimer's Association et al. 2018 Alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 14(3):367–429, 2018. 4

[14] Clifford R Jack, David S Knopman, William J Jagust, Ronald C Petersen, Michael W Weiner, Paul S Aisen, Leslie M Shaw, Prashanthi Vemuri, Heather J Wiste, Stephen D Weigand, Timothy G Lesnick, Vernon S Pankratz, Michael C Donohue, and John Q Trojanowski. Tracking pathophysiological processes in alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, 12(2):207 – 216, 2013. 4

[15] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452, 2013. 5

[16] N Schuff, N Woerner, L Boreta, T Kornfield, LM Shaw, JQ Trojanowski, PM Thompson, CR Jack Jr, MW Weiner, and Alzheimer's Disease Neuroimaging Initiative. MRI of hippocampal volume loss in early alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain*, 132(4):1067–1077, 2009. 5

[17] Donald M. Lyall, Simon R. Cox, Laura M. Lyall, Carlos Celis-Morales, Breda Cullen, Daniel F. Mackay, Joey Ward, Rona J. Strawbridge, Andrew M. McIntosh, Naveed Sattar, Daniel J. Smith, Jonathan Cavanagh, Ian J. Deary, and Jill P. Pell. Is there association between apoe e4 genotype and structural brain ageing phenotypes, and does that association increase in older age in uk biobank? (n = 8,395). *bioRxiv*, 2017. 5

[18] Bruce Fischl. Freesurfer. *NeuroImage*, 62(2):774 – 781, 2012. 6

[19] Derrek P Hibar, Hieab HH Adams, Neda Jahanshad, Ganesh Chauhan, Jason L Stein, Edith Hofer, Miguel E Renteria, Joshua C Bis, Alejandro Arias-Vasquez, M Kamran Ikram, et al. Novel genetic loci associated with hippocampal volume. *Nature Communications*, 8:13624, 2017. 6, 8