Genome-Wide Association Study of Brain Connectivity Changes for Alzheimer's Disease

Samar S. M. Elsheikh[§], Emile R. Chimusa[¶], Nicola Mulder[§], Alessandro Crimi[†,‡], and Alzheimer's Disease Neuroimaging Initiative[⋆]

June 8, 2018

§ Computational Biology Group, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, South Africa

¶ Division of Human Genetics, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, South Africa

† University Hospital of Zürich, Switzerland

‡ African Institute for Mathematical Sciences, Ghana

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

1 abstract

Variations in the human genome have been found in the literature to be an essential factor that affects the susceptibility of Alzheimer's disease (AD). Genome-wide association studies (GWAS) have identified genetic loci that significantly contribute to the risk of AD. The availability of the genetic data, coupled with brain images technologies leaving rooms for more opportunities and allows further study designs. Although methods have been proposed towards integrating imaging and genetic information, the statistical power remains a primary challenge preventing the production of replicable associations. Moreover, for such kind of studies, the measurement of disease is often taken at a single time point, therefore, not allowing the disease progression to be taken into consideration. In longitudinal settings, we analysed a DTI and Single Nucleotide Polymorphisms (SNPs) datasets obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI), by conducting a GWAS on the global network metrics absolute change, after which we used the GWAS summary statistics to compute the gene scores. We observed significant associations between the change in the AD brain and genes which reported to manipulate AD susceptibility, brain structure and function.

2 Introduction

Coming ...

3 Materials and Methods

3.1 Datasets and Data Processing

Our analysis was conducted on Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset publicly available at (adni.loni.usc.edu)[1].

[1]The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner,MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be com-

We used two types of datasets; 1) the cohort measurements of Diffusion Tensor Images (DTI) along with T1-Weighted to allow co-registration, taken at two time points, at the baseline and follow-up after 12 months, and; 2) the PLINK binary files (BED/BIM/FAM) genotypic data for AD, controls and Early MCI (Early Mild Cognitive Impairment).

3.2 Brain Connectivity Metrics

To assess longitudinal changes, we evaluate those metrics at different time points, and we compute their absolute differences. To be in line with previous works on AD and connectomics [PNTT13, BTB+11, JRH+13], , in this work, we focus on specific network segregation and integration features.

Segregation represents the ability of a network to form communities/clustering which are well organized [DT+15].

Integration represents the network's ability to propagate information efficiently[DT+15].

Louvain modularity and transitivity were computed to quantify segregation, while global efficiency and characteristic path length were computed as integration measurements.

The *Louvain modularity* is a community detection method that partitions the network using a two-step optimization[RS10]. First, the method groups individual nodes into communities by optimizing modularity locally, then it constructs a new network whose nodes are the newly formed communities. This is repeated until a maximum of modularity is attained and a hierarchy of communities is produced. For weighted graphs, modularity is defined as

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{1}$$

where A_{ij} is the weight of the edge connecting between nodes i and j from the adjacency matrix $\mathbf{A}$, k_i and k_j are the sums of weights of the edges connected to node i and j respectively, $m = 1/(2A_{ij})$, c_i and c_j are the communities of nodes i and j, and δ is a simple delta function.

The *transitivity* is a variation of the clustering coefficient computed directly on the global network [RS10] , and it reflects the prevalence of clustered connectivity around its nodes. It can be defined as

$$T^W = \frac{\sum_{i \in N} 2t_i^W}{\sum_{i \in N} k_i(k_i - 1)}, \tag{2}$$

with t_i^W being the weighted geometric mean of triangles around node i and k_i its degree.

Weighted global efficiency represents how efficiently information is exchanged over the network . It is computed as the inverse of the average weighted shortest path length in the network d_{ij}^W, and is also inversely related to the characteristic path length as

$$E^W = \frac{1}{n(n-1)} \sum_{i \in N} \sum_{j \in N, i \neq j} (d_{ij}^W)^{-1}, \tag{3}$$

where n is the number of nodes.

Characteristic path length measures the integrity of the network and how it is easy for the information to flow within the network. Let d_{ij} be the number of links (connections) which represent the shortest path between node i and j. Characteristic path length of the network is the average of all the distances between every pair in the network.

$$L^W = \frac{1}{n(n-1)} \sum_{i,j \in N, i \neq j} d_{ij}^W \tag{4}$$

3.3 Integration of the two datasets

To quantify the longitudinal change between the baseline and follow-up visits from the imaging dataset, we calculated the absolute differences for each brain connectivity measure between the two time points. We then merged the absolute differences with the PLINK fam file, using the subject ID. In such settings, subjects who were not present in both datasets were removed and not considered in the rest of this analysis. In total, we were able to merge 57 subjects.

bined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org

3.4 Data analysis

We performed four quantitative GWAS, separately, using PLINK software[PNTB+07] (http://pngu.mgh.harvard.edu/purcell/plink/), a GWAS for each network connectivity measure with a total of 12809667 variants (genotyping rate was 0.84943). We used the absolute difference for each one of the measures as our quantitative trait (phenotype).

After we obtained the GWAS association results, we used them as input for the Pascal (Pathway scoring algorithm) software[LMR+16] to aggregate SNPs at a gene level, and hence, compute gene scores for the four network measures. Along with the obtained association statistics; Pascal uses a reference population from 1000 Genomes Project[C+12] to correct for linkage disequilibrium (LD) between SNPs. We sat Pascal to compute the gene scores according to the sum of chi-square statistics, and get a p-value for each gene, provided that there were SNPs presents for that gene. Finally, we used python to plot the Manhattan plot, and R studio[RSt15] to plot the QQ-plots. Figure 1 shows the pipeline of this analysis.

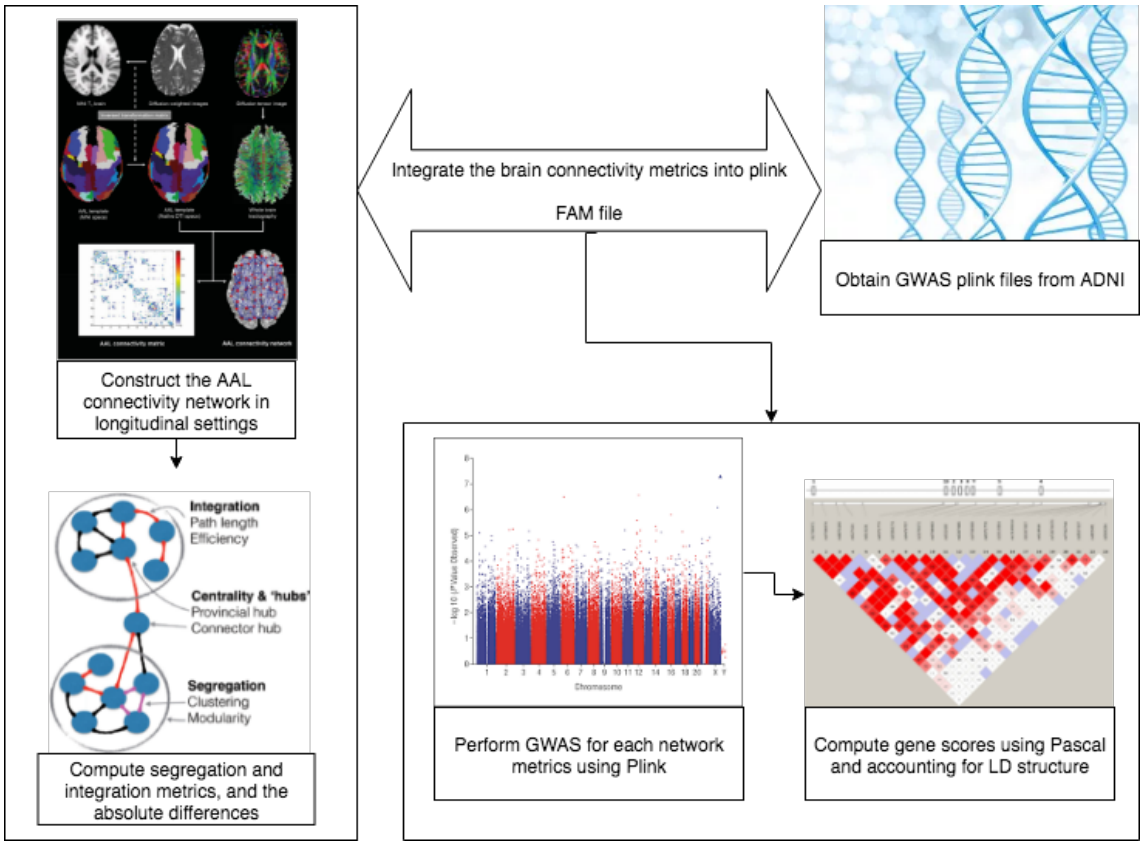

Figure 1: The analysis pipeline. An illustration of the workflow: both genetic and imaging data were obtained from ADNI; the DTI images were processed, along with AAL and T1 weighted images to construct the brain network, after which the global brain network metrics were computed. The latter were merged with the PLINK FAM files for all the subjects present in both datasets. Lastly, GWAS was conducted, and its results were used by PASCAL to calculate the gene scores (accounting for LD patterns using reference dataset)

4 Results

We follow the pipeline in figure ?? to get our results. After processing the two datasets, we merged them and calculated the absolute differences between the baseline and follow-up visits. To have a general idea of how the differences of these connectivity metrics are distributed, we plotted four boxplots, as shown in figure 3. The boxplots correspond to transitivity, global effect, Louvain modularity and characteristics path length, respectively.

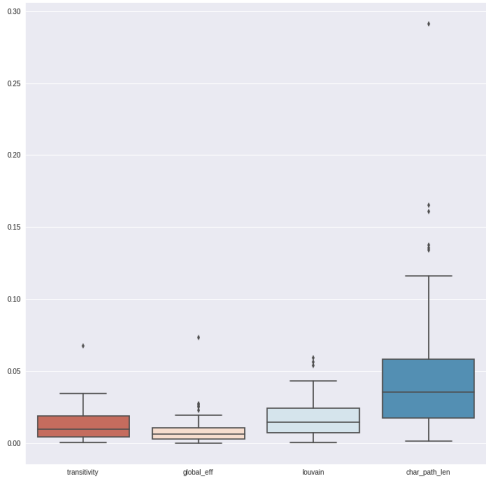

Figure 2: Distribution of global network metrics for controls and AD subjects, jointly. Shortcuts stand for; louvain: Louvain modularity, global_eff: global effect, and char_path_len: characteristic path length

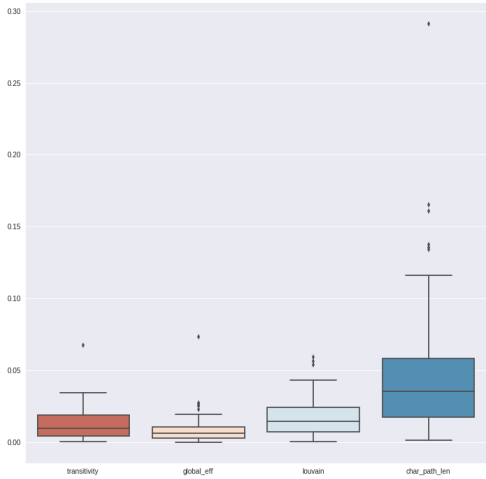

Figure 3: Distribution of global network metrics for controls and AD subjects, jointly. Shortcuts stand for; louvain: Louvain modularity, global_eff: global effect, and char_path_len: characteristic path length

Figure 4 illustrated the longitudinal change of global network metrics - measured as the absolute differences, in more details; it also compares Alzheimer's patients to healthy participants.

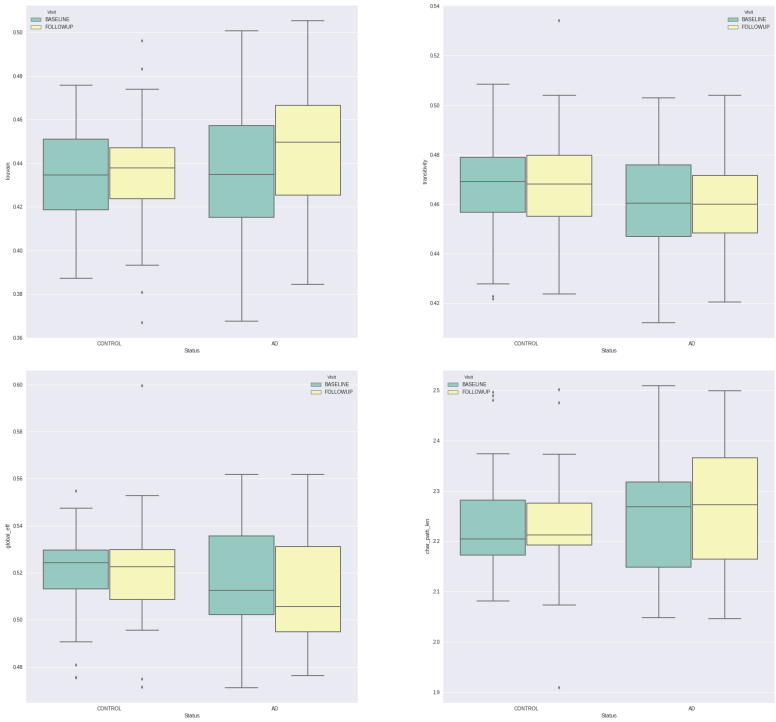

Figure 4: Boxplots for global network metrics to compare AD and controls in the baseline (green) and follow-up (yellow). The metrics are, Louvain modularity (top left), transitivity (top right), global effect (bottom left) and characteristic path length (bottom right)

The left plot in figure 5 compares AD patients with controls regarding how the differences in brain connectivity metrics are distributed (histograms) and correlated with each other (boxplots). Note that the diagonal shows a stacked histogram of the differences for each measure, therefore, the AD histogram is plotted on top of the control histogram. While the right plot in figure 5 show the distribution of the actual brain connectivity measure, and it compares their distribution in the baseline and follow-up visits. We then performed a standard GWAS analysis using PLINK, with

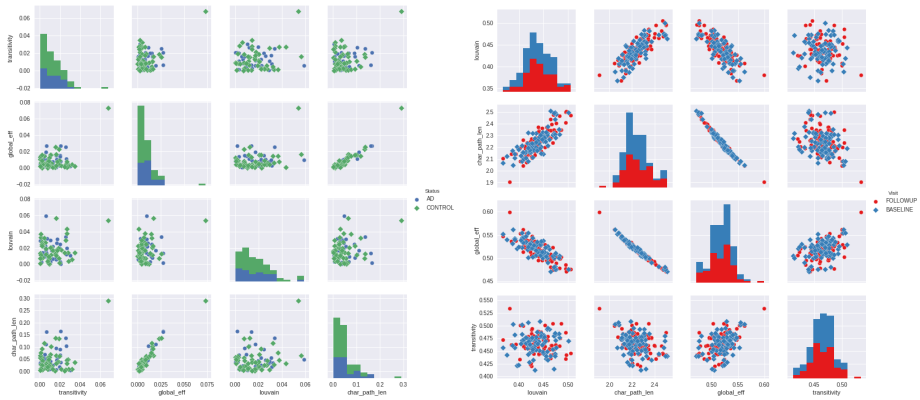

Figure 5: Global network metrics scatter plots. The left plot shows the distribution of the absolute difference in network metrics; it compares AD to controls. While right plot shows the distribution of the actual metrics in the baseline and follow-up, for all participants

the phenotypes as a quantitative measure, representing the change in connectivity metrics in this analysis. Figure 6 shows the GWAS results for brain segregation measures, and figure 7 shows the GWAS results for brain integration measures. In those figures, the Manhattan plot appears to the

left. The x-axis of the Manhattan plot represents the locations along the genome, while the y-axis is $(-\log 10(p - value))$ and not the actual p-values, each dot represents a single SNP. The right plot is the qqplot for the same results, and the diagonal line represents the expected (under the null hypothesis) distribution of p-values. Again, each dot in the qqplot represents a single SNPs, and the actual p-values transformed using the same expression; $(-\log 10(p - value))$. Using the

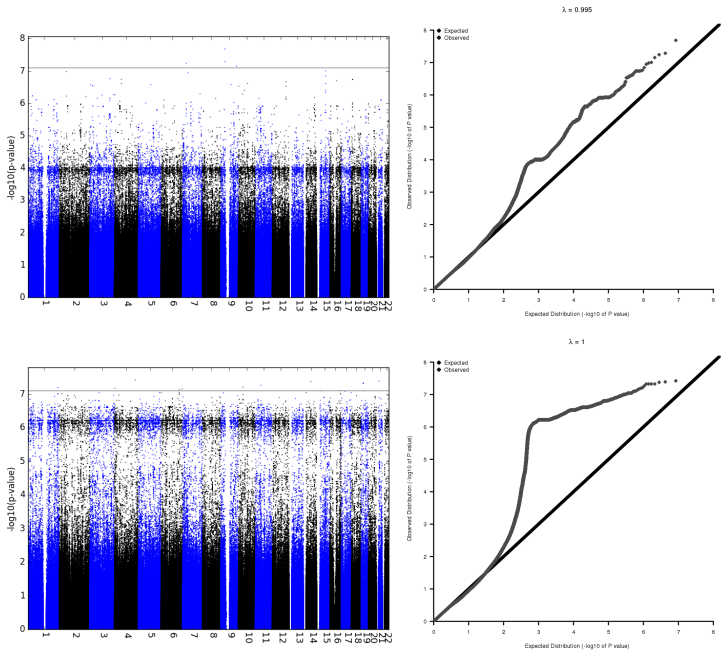

Figure 6: GWAS results for the change in Louvain modularity (top) and transitivity (bottom) segregation connectivity metrics. To the left is the Manhattan plot, while the qqplot is to the right

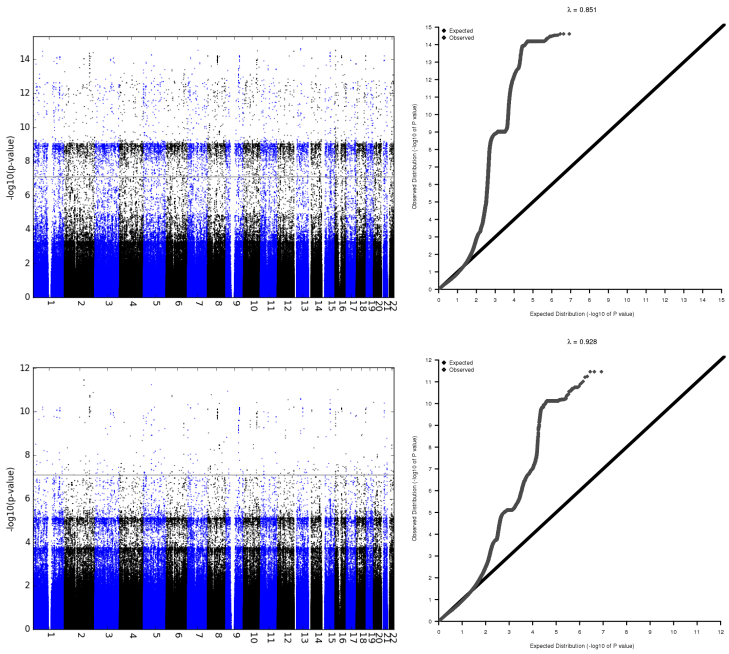

Figure 7: GWAS results for the change in global effect (top) and characteristics path length (bottom) integration connectivity metrics. To the left is the Manhattan plot, while the qqplot is to the right

GWAS association results, namely the p-values, we computed the genes score using Pascal. Figure

8 and figure 9 are the Pascal results for brain segregation and integration measure, respectively. Here, we could not use the standard GWAS cutoff point (7.9×10^{-8}) because the number of genes is much smaller than the SNPs. We then used the number of genes ($21,407$) to calculate a new cutoff point, $\frac{7.9}{21,407}$, hence, we used a cutoff point of the value 0.000369038.

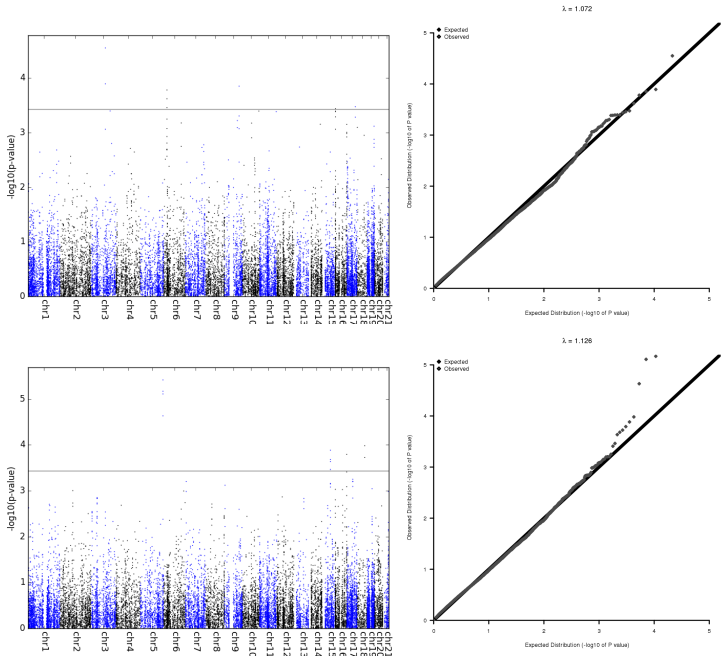

Figure 8: Gene score p-values Manhattan plot (left) and qqplot (right) for Lovuain modularity (top) and transitivity (bottom) segregation connectivity metrics. Results obtained using Pascal software

For each Pascal result (and for both brain segregation and integration measures) we constructed a table that contains the most ten significant SNPs. These tables contain six columns; the gene symbol, the genes start and end locations, respectively, the gene ID, the number of SNPs in that gene and the p-value corresponds to the gene score calculated with Pascal. Tables for Louvain modularity, transitivity, global effect and charactaristic path length are Table 1 Table 2 Table 4 and Table 3, respectively

Table 1: Top 10 significant gene scores obtained using Louvain GWAS results

Gene Symbol	**Results are sorted according to p-value**					
	Chr	Start	End	Gene ID	SNPs	Pvalue
QTRTD1	chr3	113775581	113807268	79691	185	2.811E-5
KIAA1407	chr3	113682983	113775460	57577	267	1.278E-4
FAM225A	chr9	115875175	115882126	286333	171	1.409E-4
BTN1A1	chr6	26501457	26510652	696	259	1.658E-4
LOC285819	chr6	26472171	26482737	285819	289	2.398E-4
PRKAR1A	chr17	66507920	66547457	5573	331	3.356E-4
HCG11	chr6	26521933	26527612	493812	252	3.489E-4
UBE2I	chr16	1359153	1377019	7329	392	3.634E-4
P2RY12	chr3	151055235	151102544	64805	350	4.014E-4
FOXI2	chr10	129535537	129539450	399823	387	4.032E-4

4.1 Acknowledgement

We would like to acknowledge our funders, the Organisation for Women in Science for the Developing World (OWSD), the Swedish International Development Cooperation Agency (Sida) and the University of Cape Town for their continuous support. Most of the computations in this work were performed using facilities provided by the University of Cape Town's ICTS High Performance

Table 2: Top 10 significant gene scores obtained using transitivity GWAS results

Gene Symbol	**Results are sorted according to p-value**					
	Chr	Start	End	Gene ID	SNPs	Pvalue
TSPAN17	chr5	176074387	176086059	26262	213	3.852E-6
EIF4E1B	chr5	176057682	176073642	253314	199	6.783E-6
MIR4281	chr5	176056439	176056501	100422962	180	7.789E-6
SNCB	chr5	176047209	176057557	6620	234	2.338E-5
LOC100996669	chr18	59415408	59421928	100996669	257	1.043E-4
TRIP4	chr15	64680002	64747502	9325	164	1.309E-4
LOC100287036	chr16	89387540	89391518	100287036	187	1.614E-4
RNF152	chr18	59482303	59560304	220441	602	1.889E-4
KIAA0101	chr15	64657210	64673702	9768	108	2.0815E-4
ZNF609	chr15	64791618	64978266	23060	366	2.317E-4

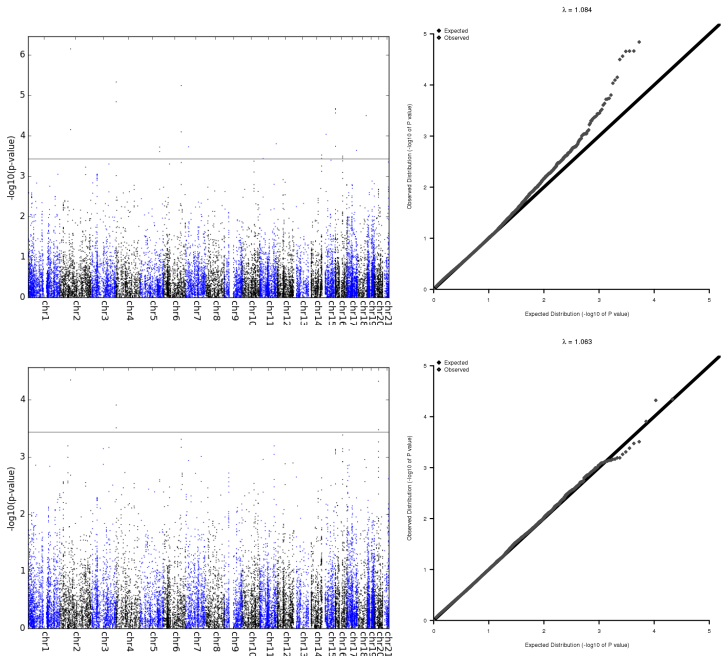

Figure 9: Gene score p-values Manhattan plot (left) and qqplot (right) for global effect (top) and characteristic path length (bottom) integration connectivity metrics. Results obtained using Pascal software

Table 3: Top 10 significant gene scores obtained using characteristic path length GWAS results

Gene Symbol	**Results are sorted according to p-value**					
	Chr	Start	End	Gene ID	SNPs	Pvalue
SUCLG1	chr2	84650646	84686586	8802	135	4.512E-5
DEFB116	chr20	29891014	29896388	245930	199	4.768E-5
TMED11P	chr4	1108984	1116952	100379220	336	1.240E-4
RNF212	chr4	1065265	1107352	285498	479	3.098E-4
DEFB115	chr20	29845466	29847435	245929	195	3.349E-4
CIAPIN1	chr16	57462086	57481369	57019	200	4.1461E-4
KIAA1244	chr6	138483052	138665800	57221	497	4.912E-4
DEFB119	chr20	29964965	29978452	245932	220	5.480E-4
DBIL5P2	chr2	63344985	63346677	100169989	106	6.424E-4
CLDN25	chr11	113650517	113651207	644672	253	6.4243E-4

Table 4: Top 10 significant gene scores obtained using global effect GWAS results

Gene Symbol	**Results are sorted according to p-value**					
	Chr	Start	End	Gene ID	SNPs	Pvalue
SUCLG1	chr2	84650646	84686586	8802	135	7.111E-7
TMED11P	chr4	1108984	1116952	100379220	336	4.681E-6
HEBP2	chr6	138725335	138734582	23593	203	5.727E-6
RNF212	chr4	1065265	1107352	285498	479	1.443E-5
TSR3	chr16	1399240	1401873	115939	279	2.166E-5
UBE2I	chr16	1359153	1377019	7329	392	2.173E-5
GNPTG	chr16	1401899	1413352	84572	292	2.201E-5
BAIAP3	chr16	1383605	1399442	8938	310	2.733E-5
LOC400655	chr18	70818475	70931733	400655	454	3.174E-5
DNAH6	chr2	84743578	85046713	1768	576	7.075E-5

Computing team: http://hpc.uct.ac.za 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.

References

[BTB+11] J.A. Brown, K.H. Terashima, A.C. Burggren, L.M. Ercoli, K.J. Miller, and other. Bain network local interconnectivity loss in aging apoe-4 allele carriers. *Proceedings of the National Academy of Sciences*, 108(51):20760–20765, 2011.

[C+12] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

[DT+15] G. Deco, G. Tononi, et al. Rethinking segregation and integration: contributions of whole-brain modelling. *Nature Reviews Neuroscience*, 16:430–439, 2015.

[JRH+13] N. Jahanshad, P. Rajagopalan, X. Hua, D. Hibar, , et al. Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity. *Proceedings of the National Academy of Sciences*, 110(12):4768–4773, 2013.

[LMR+16] David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS computational biology*, 12(1):e1004714, 2016.

[PNTB+07] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[PNTT13] G. Prasad, T.M. Nir, A.W. Toga, and P.M. Thompson. Tractography density and network measures in Alzheimer's disease. In *Biomedical Imaging, 2013 IEEE 10th International Symposium on*, pages 692–695, 2013.

[RS10] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[RSt15] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015.