Exhaustive search of the SNP-SNP interactome identifies replicated epistatic effects on brain volume

Derrek P. Hibar[1], Jason L. Stein[1], Neda Jahanshad[1], Omid Kohannim[1], Arthur W. Toga[1], Katie L. McMahon[2], Greig I. de Zubicaray[3], Grant W. Montgomery[4], Nicholas G. Martin[4], Margaret J. Wright[4], Michael W. Weiner[5,6], Paul M. Thompson[1]

[1]Imaging Genetics Center, Laboratory of Neuro Imaging, UCLA School of Medicine, Los Angeles, CA 90095, USA
[2]Center for Magnetic Resonance, School of Psychology, University of Queensland, Brisbane, Queensland, Australia
[3]Functional Magnetic Resonance Imaging Laboratory, School of Psychology, University of Queensland, Brisbane, Queensland, Australia
[4]Genetic Epidemiology Laboratory, Queensland Institute of Medical Research, Brisbane, Australia
[5]Departments of Radiology, Medicine, Psychiatry, UC San Francisco, San Francisco, CA, USA
[6]Department of Veterans Affairs Medical Center, San Francisco, CA, USA

Abstract. The SNP-SNP* interactome** has rarely been explored in the context of neuroimaging genetics (or quantitative genetics, in general) mainly due to the complexity of conducting ~10^{11} pairwise statistical tests. However, recent advances in machine learning, specifically the iterative sure independence screening (SIS) method, have enabled the analysis of datasets where the number of predictors is much larger than the number of observations. The SIS method ranks the predictors in a set based on their cumulative marginal effect on some dependent variable. In this way, SIS can identify a subset of predictors that explain the maximum amount of variance in a given dependent variable. Using an implementation of the SIS algorithm (called EPISIS), we used exhaustive search of the genome-wide, SNP-SNP interactome to identify and prioritize SNPs for interaction analysis. We identified a significant SNP pair, rs1345203 and rs1213205, associated with temporal lobe volume. We further examined the full-brain, voxelwise effects of the SNP-SNP interaction in the ADNI dataset and separately in an independent dataset of young healthy twins (QTIM). We found that each additional loading in the epistatic effect was associated with ~5% greater brain regional brain volume (a protective effect) in both the ADNI and QTIM samples.

Keywords: epistasis, interaction, genome, sure independence, tensor-based morphometry

*__SNP__ (=single nucleotide polymorphism): a single-letter variant in the genome; these variations are common, even in healthy human populations, and their effects on brain measures can be assessed using association testing, at one SNP or up to a million genotyped SNPs.
**__Interactome__: The study of interactions between genetic variants or sets of variants in terms of their effects on traits such as brain measures.

1 Introduction

Traditional univariate methods can test the association of common genetic variants with complex quantitative traits, but they only consider the marginal effect of a single locus and potentially miss variance explained by synergistic or interacting effects of pairs or sets of SNPs [Marchini et al., 2005]. For many complex traits, the similarity of family members drops faster than would be expected as relatedness decreases [Wray et al., 2010]. This implies that there are non-additive (epistatic) interactions involved in the etiology of many complex traits. Statistical interactions have been demonstrated to be plausible representations of the complex interactions of genes in biological pathways [Moore et al., 2009; Stich et al., 2007].

Some prior studies have examined second-order interactive effects of SNPs on brain structure [Pezawas et al., 2008; Wang et al., 2009; Tan et al., 2007]. However, none of these studies has considered genome-wide genotype data; the closest conceptually related study tested for SNP effects on diffusion imaging measures, and aggregated all SNPs with correlated effects into a network [Chiang et al., 2012]. The concept here is different, and aims to assess gene pairs that influence each other's effects on the brain. Prior studies tested interaction effects only for a limited number of popular candidate genes. Any approach based on pre-selecting a pair of genes will overlook a vast search space of potential interactions among SNPs in the genome that have no obvious prior connection. Also, a large main effect is not necessary to be able to detect significant second-order interactions [Marchini et al., 2005]. Given this, prior hypotheses focusing on SNPs with large individual effects may also overlook large second-order effects. Importantly, power estimates for detecting interactive effects are comparable to those for single SNP tests [Marchini et al., 2005]. In simulation studies, the inclusion of interaction terms can boost the power to detect main effects, at least for certain genetic tests [Cordell et al., 2001]. Here we examined the genome-wide, SNP-SNP interactome to test genetic associations with a quantitative biomarker of Alzheimer's disease (temporal lobe volume) in the public Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. We further examine the whole-brain effects of interaction pairs in statistical parametric maps generated with tensor-based morphometry (TBM); we also replicate our tests in an independent, non-overlapping dataset of young healthy twins from the Queensland Twin Imaging (QTIM) study [de Zubicaray et al., 2008].

2 Methods

2.1 Imaging parameters and study information

We downloaded the full baseline set of 818 high-resolution, T1-weighted structural MRI brain scans from the Alzheimer's Disease Neuroimaging Initiative (ADNI). ADNI is a multi-site, longitudinal study of patients with Alzheimer's disease (AD), mild cognitive impairment (MCI) and healthy elderly controls (HC). Subjects were scanned with a standardized protocol to maximize consistency across sites. We used the baseline 1.5 Tesla MRI scans, i.e., the T1-weighted 3D MP-RAGE scans, with

TR/TE = 2400/1000 ms, flip angle = 8˚, slice thickness = 1.2 mm, and a final voxel resolution = 0.9375 x 0.9375 x 1.2 mm^3. Raw MRI scans were pre-processed to remove signal inhomogeneity, non-brain tissue, and affine registered to the MNI template (using 9 parameters).

Additionally, we obtained 753 high-resolution, T1-weighted structural MRI brain scans from the Queensland Twin Imaging (QTIM) study. QTIM is a longitudinal neuroimaging and genetic study of young, healthy twins and their family members. All structural MRI scans were acquired on a single 4-Tesla scanner (Bruker Medspec): T1-weighted images, inversion recovery rapid gradient echo sequence, TR/TE = 1500/3.35 ms, flip angle = 8˚, slice thickness = 0.9 mm, 256 x 256 acquisition matrix, with a final voxel resolution = $0.9375 \times 0.9375 \times 0.9$ mm^3. Raw MRI scans were pre-processed to remove signal inhomogeneity, non-brain tissue, and affine registered to the ICBM template (using 9 parameters).

2.2 Genotype pre-processing and study demographics

Genome-wide genotyping data were available for the full set of ADNI subjects. We performed standard quality control procedures to ascertain the largest homogenous genetic sub-population in the dataset, using multi-dimensional scaling (MDS) compared to a dataset of subjects of known genetic identity (HapMap III; http://hapmap.ncbi.nlm.nih.gov/). The largest subset contained 737 subjects from the CEU population (Caucasians). We therefore removed the remaining 81 subjects from our analysis to limit the effects of genetic stratification on our statistical analyses [Lander and Schork, 1994]. Additionally, we applied filter rules to the genotype data to remove rare SNPs (minor allele frequency < 0.01), violations of Hardy-Weinberg Equilibrium (HWE $p < 5.7 \times 10^{-7}$), and poor call rate (<95%). Data were further "phased" to impute any missing individual genotypes after filtering using the MaCH program [Abecasis et al., 2010] following the ENIGMA imputation protocol [ENIGMA2 Genetics Support Team, 2012]. After filtering and phasing, 534,033 SNPs remained.

All QTIM subjects were ascertained for genetic similarity, so no subjects were removed before analysis. All 753 subjects in the QTIM dataset clustered with the CEU population, in the MDS analysis. The same genotype filter rules from the ADNI dataset were applied to the QTIM sample's genetic data. After filtering and phasing, 521,232 SNPs remained.

After all rounds of genotype pre-processing, the ADNI sample contained 737 subjects (mean age±sd: 75.5±6.8 yrs; 436 males) comprised of 173 patients diagnosed with Alzheimer's disease, 358 subjects with mild cognitive impairment, and 206 healthy elderly controls. The QTIM sample contained 753 subjects (mean age±sd: 23.1±3.0 yrs; 286 males) and consisted of 110 monozygotic twin pairs, 147 dizygotic twin pairs, 3 dizygotic twin trios, 143 singletons, and 87 siblings from 438 families.

2.3 Tensor-based morphometric differences in the full brain

We calculated information on regional brain morphometry using an elastic, nonlinear registration algorithm [3DMI; Leow et al., 2005] applied to the entire brain. Voxel-wise volumetric differences were stored, using the Jacobian value of the deformation matrix obtained by nonlinearly registering a subject's scan to a study-specific minimum deformation template (MDT). Scans from the ADNI and QTIM datasets were processed and analyzed separately (using separate study templates). The MDT for the ADNI sample is a nonlinear average of 40 age-and-sex matched healthy elderly controls [Hua et al., 2012]. The MDT for the QTIM is a nonlinear average of 32 age- and sex- matched, unrelated subjects [Jahanshad et al., 2012]. Nonlinear registration with 3DMI yields a 110 x 110 x 110 voxel statistical parametric map, where the Jacobian value at each voxel represents the expansion required to match the same voxel in the study-specific MDT.

2.4 Genome-wide, gene-gene interaction testing

The EPISIS software is an implementation of the machine-learning algorithm called *sure independence screening* (SIS) developed by Fan and Lv [2008]. The SIS algorithm is a correlation learning method that can be applied to ultra-high dimensional datasets where the number of predictors p is much greater than the number of observations n. Despite the development of robust methods for cases where $p>n$ (e.g., the Dantzig selector of Candes and Tao 2007) the properties of the selector fail when $p>>n$. Fan and Lv [2008] developed the SIS algorithm to reduce the ultra-high dimension of p to a moderately-sized subset, while guaranteeing that the subset still explains the maximum amount of variance explained by the full set of predictors.

We conducted an exhaustive search of association tests of genome-wide SNP-SNP interactions with temporal lobe volume (computed by integrating the Jacobian over an temporal lobe ROI on the MDT; Stein et al., 2010) in the ADNI dataset using the EPISIS software. EPISIS utilizes the massively parallel processing available in GPGPU (General-purpose computing on graphics processing units) framework to test $p(p\text{-}1)/2$ SNP-SNP interactions in the ADNI dataset in a feasible timeframe. We used the SIS algorithm with cell-wise dummy coding [CDC; Ueki and Tamiya, 2012] to reduce the full predictor space into a subset d of $n/\log(n)$ interaction terms [Fan and Lv, 2008]. After screening the full set of possible two-way SNP-SNP interactions, we applied ridge regression [Hoerl, 1962; Kohannim et al., 2011] to the subset of interaction terms (the multiplicative loading of each SNP-SNP pair) and selected significant SNP-SNP interaction terms using the extended Bayesian Information Criterion [EBIC; Chen and Chen, 2008] with $\gamma = 0.5$. The choice of the parameter γ was chosen based on simulations [Ueki and Tamiya, 2012]. A single exhaustive search of the genome-wide, SNP-SNP interactome with EPISIS was completed in 7 hours (using one NVIDIA Tesla C2050 GPU card).

2.5 Voxelwise interaction analysis and replication

We tested the significant SNP-SNP interaction pair selected by ridge regression for association with voxelwise, regional volume differences (V) at each point, i, in the full

brain. The association test at each voxel in the ADNI dataset followed the multiplicative interaction model in multiple linear regression:

$$V_i \sim \beta_0 + \beta_{age}X_{age} + \beta_{sex}X_{sex} + \beta_{snp1}X_{snp1} + \beta_{snp2}X_{snp2} + \beta_{snp1,2}X_{snp1}*X_{snp2} + \varepsilon$$

Additionally, we used QTIM as an independent replication sample of the top SNP-SNP interaction pair identified by ridge regression after EPISIS. The voxelwise association tests assume the multiplicative interaction model, detailed previously. Due to the family design of the QTIM sample, we tested association using mixed-effects modeling as implemented in the R package *kinship* (version 1.3) in order to account for relatedness.

3 Results

After screening the full set of SNP-SNP interaction pairs for association with temporal lobe volume in the ADNI dataset, we obtained a subset d of SNP-SNP interaction pairs such that $d = n/\log(n)$. The subset is chosen by ranking the marginal correlation coefficients of each interaction pair and selecting the top d SNP-SNP pairs (correlation learning; Fan and Lv, 2008), in this case $d = 111$ pairs). Next, we applied ridge regression to the pruned subset of SNP-SNP interaction pairs. Using the extended BIC [$\gamma = 0.5$; Ueki and Tamiya, 2012] to estimate significance in our ridge regression, we identified a significant interaction between rs1345203 and rs1213205. The distribution of alleles for each SNP and their interaction is given in Table 1.

Study	rs1345203	rs1213205	Interaction
ADNI (n=737)	G/G: 27	A/A: 93	0 loadings: 612
	A/G: 223	G/A: 297	1 loadings: 79
	A/A: 487	G/G: 347	2 loadings: 46
QTIM (n=753)	G/G: 5	A/A: 78	0 loadings: 664
	A/G: 193	G/A: 300	1 loadings: 70
	A/A: 555	G/G: 375	2 loadings: 19

Table 1. The distribution of alleles for the significant SNPs and the number of subjects with each genotype by study. For rs1345203 the minor allele is G and the major allele is A in both studies. The minor allele is A and the major allele is G for rs1213205. The association testing assumes an additive model (each subject is assigned a value 0,1,2 based on the number of minor alleles they have at a given SNP). The interaction column gives the number of subjects in each category after multiplying together the counts of each of the alleles.

We further examined the significant SNP pair, rs1345203 and rs1213205, for whole-brain effects in the statistical parametric maps generated using tensor-based morphometry (TBM). In the ADNI dataset, we found broad effects bilaterally in the temporal and occipital lobes (**Figure 1**) after correcting for multiple tests at a 5% false discovery rate (FDR) using the searchlight FDR method [Langers et al., 2007].

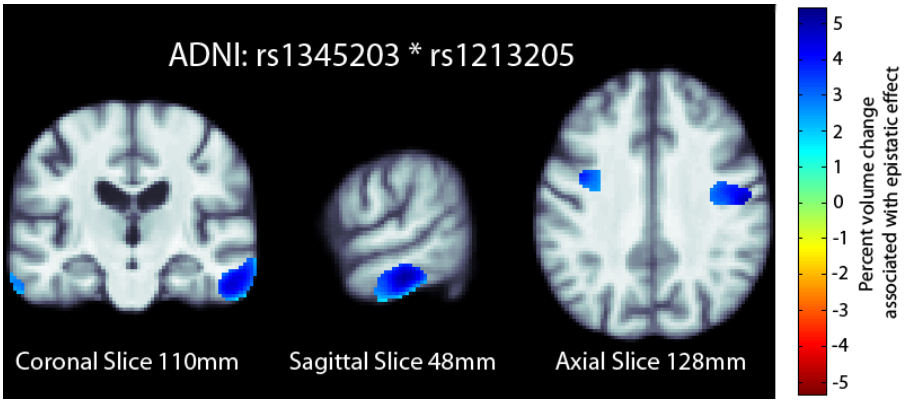

Figure 1. Corrected *p*-maps from the ADNI, overlaid on the study specific template. Only significant regions in the corrected *p*-map are shown after correcting for multiple comparisons with searchlight FDR [Langers et a., 2007] at a 5% false discovery rate. Images follow radiological orientation. The origin is placed at the Posterior-Right-Inferior corner. Cooler colors over the tissue represent tissue expansion (larger regional brain volume) compared to an average template. There is a clear protective effect of the epistatic loadings bilaterally in the temporal and occipital lobes.

We examined the whole-brain effects of the SNP pair on voxelwise, regional brain volume in the statistical parametric maps in an independent dataset (QTIM). The distribution of alleles for each SNP and their interaction in the QTIM sample is given in Table 1. In the QTIM, we identified significant effects in the left temporal lobe and along the border of the left frontal and occipital lobes (**Figure 2**) after correction for multiple tests at 5% false discovery rate (FDR) using the searchlight FDR method.

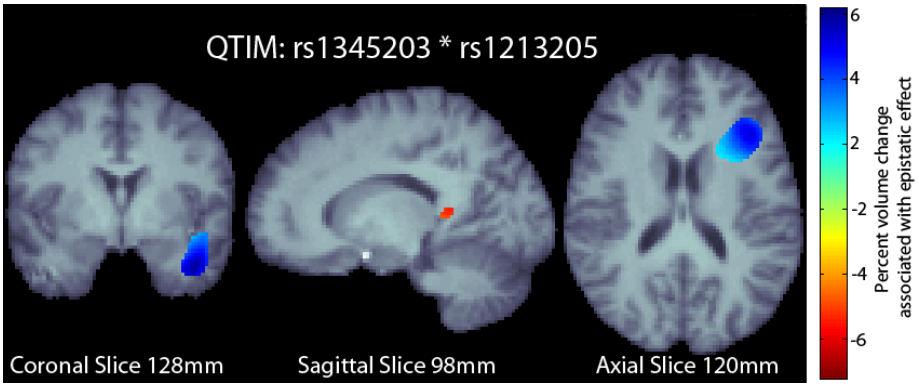

Figure 2. Corrected *p*-maps from the QTIM are overlaid on the study specific template. Only significant regions in the corrected *p*-map are shown, after correction for multiple comparisons with searchlight FDR [Langers et a., 2007] at a 5% false discovery rate. Images follow radiological orientation. The origin is placed at the Posterior-Right-Inferior corner. Cooler colors over the tissue represent tissue expansion (larger regional brain volume) compared to an aver-

age template. There is a clear protective effect of the epistatic loadings in the left temporal lobe and along the boundary of the frontal and occipital lobe.

4 Discussion

The genome is incredibly complex and statistical epistasis has been suggested as an appropriate model for the biological interactions among genes and protein products in related pathways [Moore et al., 2009; Stich et al., 2007]. Following the definition of epistasis given by Fischer [1918], here we examined the multiplicative effect of SNP-SNP pairs on brain volume differences. Significant interaction terms explain additional variance in brain volume beyond what is already explained by the additive SNP terms. In our primary tests of associations with temporal lobe volume in the ADNI dataset, we screened 10^{11} possible SNP-SNP interaction pairs using the GPU acceleration implemented in the EPISIS software. The top 111 interaction pairs were selected after ranking the marginal effect of each SNP-SNP pair on temporal lobe volume, using an implementation of the sure independence screening (SIS) algorithm [Fan and Lv 2008]. We used ridge regression and the extended BIC [Chen and Chen, 2008] to identify a significant interaction between rs1345203 and rs1213205. The functional relevance of the two SNPs is as yet unknown. However, data obtained from the ENCODE dataset (http://genome.ucsc.edu/) show that rs1345203 is located in a transcription factor gene (ELF1/CEBPB) that demonstrates regulatory influence on the DNA structure. The SNP rs1213205 is located in a region of hypersensitivity to cleavage by DNase regulatory elements. Additional work is still required to identify precisely how these two SNPs might affect brain structure, and to further replicate their interaction. Specifically, we need to identify how changes at a given SNP are related to changes in activity in gene transcription or translation into protein products involved in similar biological pathways.

References

1. Abecasis, G. R., Li, Y., Willer, C. J., Ding, J., & Scheet, P. (2010). MaCH: Using Sequence and Genotype Data to Estimate Haplotypes and Unobserved Genotypes. *Genetic Epidemiology*, *34*(8), 816-834.
2. Candes, E., and Tao, T. "The Dantzig selector: Statistical estimation when p is much larger than n." *The Annals of Statistics* 35.6 (2007): 2313-2351.
3. Chen, Jiahua, and Zehua Chen. "Extended Bayesian information criteria for model selection with large model spaces." *Biometrika* 95.3 (2008): 759-771.
4. Chiang, M.-C., et al. "Gene network effects on brain microstructure and intellectual performance identified in 472 twins." *Journal of Neuroscience* 32.25 (2012): 8732-8745.
5. Cordell, Heather J., et al. "Statistical modeling of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes." *Genetics* 158.1 (2001): 357-367.
6. de Zubicaray G. I., et al. (2008) Meeting the challenges of neuroimaging genetics. Brain Imaging Behav 2:258–263.
7. ENIGMA2 Genetics support team. ENIGMA2 1KGP cookbook (v3) [Online]. The
8. Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA) consortium. http://enigma.loni.ucla.edu/wp-content/uploads/2012/07/ENIGMA2_1KGP_cookbook_v3.doc [27 July 2012]

9. Fan, J., and Lv, J. "Sure independence screening for ultrahigh dimensional feature space." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.5 (2008): 849-911.
10. Fisher, Ronald A. "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance." *Transactions of the Royal Society of Edinburgh* 52.02 (1919): 399-433.
11. Hoerl, A. E. "Application of ridge analysis to regression problems." Chemical Engineering Progress 58 (1962): 54-59.
12. Hua, Xue, et al. "Unbiased tensor-based morphometry: Improved robustness and sample size estimates for Alzheimer's disease clinical trials." *Neuroimage* (2012).
13. Jahanshad N, et al. "Brain structure in healthy adults is related to serum transferrin and the H63D polymorphism in the HFE gene." *Proc Natl Acad Sci* (2012): 109(14) E851-9.
14. Kohannim, Omid, et al. "Boosting power to detect genetic associations in imaging using multi-locus, genome-wide scans and ridge regression."*Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*. IEEE, 2011.
15. Lander, E. S., & Schork, N. J. (1994). Genetic dissection of complex traits. *Science, 265*(5181), 2037-2048.
16. Langers, D. R., Jansen, J. F., & Backes, W. H. (2007). Enhanced signal detection in neuroimaging by means of regional control of the global false discovery rate. *NeuroImage, 38*(1), 43-56.
17. Leow, A., Huang, S.C., Geng, A., Becker, J., Davis, S., Toga, A., Thompson, P., 2005. Inverse consistent mapping in 3D deformable image registration: its construction and statistical properties. Inf. Process. Med. Imaging 19, 493–503.
18. Marchini, J., Donnelly, P., & Cardon, L.R. "Genome-wide strategies for detecting multiple loci that influence complex diseases." *Nature Genetics* 37.4 (2005):413-417.
19. Moore, Jason H., and Scott M. Williams. "Epistasis and its implications for personal genetics." *American Journal of Human Genetics* 85.3 (2009): 309.
20. Pezawas, L., et al. "Evidence of biologic epistasis between *BDNF* and *SLC6A4* and implications for depression." *Molecular Psychiatry* 13.7 (2008): 709-716.
21. Stich, Benjamin, et al. "Power to detect higher-order epistatic interactions in a metabolic pathway using a new mapping strategy." *Genetics* 176.1 (2007): 563-570.
22. Stein, Jason L., et al. "Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease." *Neuroimage* 51.2 (2010): 542-554.
23. Tan, H.-Y., et al. "Epistasis between catechol-O-methyltransferase and type II metabotropic glutamate receptor 3 genes on working memory brain function." *PNAS* 104.30 (2007): 12536-12541.
24. Ueki, M., & Tamiya, G. "Ultrahigh-dimensional variable selection method for whole-genome gene-gene interaction analysis." *BMC Bioinformatics* 13.1 (2012): 72.
25. Wang, Y., et al. "Evidence of Epistasis Between the Catechol-*O*-Methyltransferase and Aldehyde Dehydrogenase 3B1 Genes in Paranoid Schizophrenia." *Biological Psychiatry* 65.12 (2009): 1048-1054.
26. Wray, N.R., et al. "Multi-locus models of genetic risk of disease." *Genome Med* 2.10 (2010).