PRINCIPAL COMPONENTS REGRESSION: MULTIVARIATE, GENE-BASED TESTS IN IMAGING GENOMICS

Derrek P. Hibar[1], Jason L. Stein[1], Omid Kohannim[1], Neda Jahanshad[1],
Clifford R. Jack, Jr.[2], Michael W. Weiner[3,4], Arthur W. Toga[1], Paul M. Thompson[1],
and the Alzheimer's Disease Neuroimaging Initiative

[1]Laboratory of Neuro Imaging, Dept. of Neurology, UCLA School of Medicine, Los Angeles, CA,
[2]Mayo Clinic, Rochester, MN, [3]Depts. of Radiology, Medicine and Psychiatry, UC San Francisco,
San Francisco, CA, [4]Dept. of Veterans Affairs Medical Center, San Francisco, CA

ABSTRACT

In imaging genomics, there have been rapid advances in genome-wide, image-wide searches for genes that influence brain structure. Most efforts focus on univariate tests that treat each genetic variation independently, ignoring the joint effects of multiple variants. Instead, we present a *gene-based* method to detect the joint effect of multiple single nucleotide polymorphisms (SNPs) in 18,044 genes across 31,662 voxels of the whole brain in a tensor-based morphometry analysis of baseline MRI scans from 731 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Our *gene-based* multivariate statistics use principal components regression to test the combined effect of multiple genetic variants on an image, using a single test statistic. In some situations, which we describe, this can boost power by encoding population variations within each gene, reducing the effective number of statistical tests, and reducing the effect dimension of the search space. Multivariate gene-based methods may discover gene effects undetectable with standard, univariate methods, accelerating ongoing imaging genomics efforts worldwide.

Index Terms— principal components regression, multivariate, voxelwise, imaging genomics, GWAS

1. INTRODUCTION

In imaging genomics, the vast amount of information in the images (>100,000 voxels) and across the genome (>12 million known variants) presents computational and statistical challenges when relating genetic variants to the structure and function of the brain. Power issues arise due to the small effect sizes of each genetic variant, and the huge numbers of statistical comparisons. Most techniques use some type of data reduction, limiting the number of genetic variants or imaging features studied, or both. The ultimate goal of these gene-hunting studies is to create a method that discovers which genetic variants affect the brain in a statistically powerful and biologically meaningful way.

In typical GWAS studies, each genetic variant (usually a SNP) is independently tested for its association to the phenotype – a mass univariate method, where no data reduction is used across the genome. For example, one study [1] performed a genome-wide search of around 500,000 SNPs, and found a novel variant in the *GRIN2B* gene that is associated with temporal lobe volume. The gene *GRIN2B* encodes a glutamate receptor that is already the target of drugs (memantine) used to treat Alzheimer's disease. Findings such as these are promising as they have biological relevance, but do not rely on a prior hypothesis about any specific SNP. However, performing mass univariate methods on imaging summary measures (such as temporal lobe volume) or ad hoc regions of interest (ROI), collapses the variation across the brain into a single number.

Several studies now perform genome-wide searches at each voxel across the brain [2]. This approach avoids having to pre-select an *ad hoc* brain region of interest and does not require prior hypotheses about which genetic variants, or which regions of interest, matter. One study [3] performed a genome-wide, brain-wide search, termed a voxelwise genome-wide association study (vGWAS), in 740 subjects from ADNI. However, none of the genetic variants identified was significant after multiple comparisons correction; several variants were promising candidates for further analysis. Future GWAS studies in imaging will likely need to reduce the number of tests and multiple comparisons using Bayesian priors, machine learning, or dimension reduction in the image or the genome. This may prioritize certain regions of the image or the genome, for later meta-analysis across multiple datasets.

Given recent advances in high-throughput genotyping, densely-packed sets of SNPs, or genetic markers, can capture increasing amounts of variation throughout the genome. Methods that consider combinations of SNPs from the same gene should more accurately describe gene effects on images than methods that test the

978-1-4244-4128-0/11/$25.00 ©2011 IEEE

independent effect of each SNP [4]. By associating the joint effect of multiple SNPs within a gene, in this study we set out to show that gene-based approaches can be more powerful, in some situations, than traditional univariate approaches. For example, if a gene contains multiple causal variants with small individual effects, univariate methods would miss these associations if a very stringent significance threshold is used (as in GWAS).

We assessed whether it would be feasible to extend to a neuroimaging database, a gene-based association method using principal components regression (PCReg). We applied PCReg across all genes, to a large database of voxelwise imaging data. We call our method a voxelwise "gene-wide" association study (vGeneWAS). By performing association tests on whole genes, we greatly reduce the number of tests (from 437,607 SNPs down to 18,044 genes). Using a voxel-based approach, we also avoid known problems associated with focusing on ROIs or summary measures. In addition, we performed direct power comparisons between gene-based tests using PCReg versus traditional univariate regression methods for GWAS.

2. METHODS

2.1. Imaging Measures

Structural MRI data were obtained following the standard ADNI protocol to ensure multisite consistency. Baseline MRI scans for each subject were analyzed using tensor-based morphometry (TBM) as described previously [5]. After quality control selection there were 731 subjects with genotyping data available (172 AD, 356 MCI, and 203 healthy elderly controls; 301 women/430 men; mean age ± sd = 75.56 ± 6.78 years). We did not split the subjects by diagnosis for this analysis, to exploit the broadest phenotypic continuum and maximize statistical power to detect genetic associations [6].

2.2. Genotypes and gene grouping

For details on how genetic data were processed for the ADNI study, please see [7]. We used several quality control measures to filter our SNPs for our analysis as detailed in [1]. Briefly, SNPs were excluded with call rate <95%, significant deviation from Hardy-Weinberg equilibrium $P < 5.7\text{x}10^{-7}$, and a minor allele frequency <0.10. After all rounds of quality control and preparation, 437,607 SNPs remained. Remaining SNPs were then grouped by gene, where "gene" is defined by the gene transcript region including both introns and exons. SNPs not located in a gene were excluded. After quality control, SNP annotation, and gene grouping, 18,044 genes were left for analysis.

2.3 Multi-SNP genetic associations

To test the joint effect of all SNPs in a gene on the volume difference (calculated from TBM) at each voxel, we employed a multiple partial-*F* test. This first estimates the fit of a "reduced model" of any number of nuisance variables on a given dependent variable and then estimates the fit of a second "full model" with the nuisance variables and any number of independent variables on the same dependent variable. Each association test results in an *F*-statistic, which represents the joint effect of the independent variables on the dependent variable, controlling for nuisance variables already in the model. The multiple partial-*F* statistic was calculated for each gene at each voxel using equation 1 below. Here *k* is df(full)-df(reduced) and RSS is the residual sum of squares:

$$F_{k,df(full)} = \frac{RSS(reduced) - RSS(full)}{df(reduced) - df(full)} / \frac{RSS(full)}{df(full)} \qquad (1)$$

Multiple partial-*F* tests are well suited for testing effects of multiple predictors on a given phenotype, but genetic data sometimes complicates testing because SNPs in the same gene are often correlated due to high "linkage disequilibrium" (LD). When the SNP values in a cohort of subjects are treated as a vector (whose components are the SNP value in each subject coded in an additive manner: 0, 1, or 2), then statistical correlations between adjacent SNPs on the genome can make different subjects' vectors highly collinear. The dependence among these almost collinear SNP vectors in the multiple partial-*F* test model can lead to improper signs of beta coefficient estimates, wildly inaccurate magnitudes of beta coefficients, large standard error estimates, and false inferences.

To avoid the complications of collinearity in the statistical model, we first performed principal component analysis (PCA) on the SNPs within each gene, storing all of the orthonormal basis vectors of the SNP matrix that explained the first 95% of the variance in the set of SNPs. Basis vectors with the highest eigenvalues (higher proportions of explained variance) were included until 95% of the variance in the SNPs was explained. The rest were discarded. These new "eigenSNPs" approximate the information in the observed SNPs, but lack the collinearity that disrupts the multiple partial-*F* test models. By first performing PCA followed by a multiple partial-*F* test, our method may be considered a variant of PCReg and produces *F*-statistics equivalent to those proposed previously for non-imaging data [8]. In this study, the independent variables built into the multiple partial-*F* test full model were the column vector output from PCA performed on each gene with age and sex as covariates. In this way, we tested the joint predictive effect of variation throughout a gene on brain volume variations on a voxel-by-voxel level.

The total number of tests of association for vGeneWAS is very high (18,044 genes x 31,662 voxels). Because of the massive processing requirement, we coded a "threaded" version of the PCA and multiple partial-*F* test steps of PCReg to split processing over multiple cores in a single CPU. Processing was further parallelized over a cluster of 10 high-performance 8-core CPU nodes. As a data reduction step, we only saved data on the gene with the lowest *P*-value at each voxel (the "top gene" at each voxel). The total time required to complete an analysis was approximately 13 days.

2.4 Effective number of test for statistical thresholds

As we noted previously [3], the minimum *P*-value at each voxel, in the null case with *n* independent tests, approximately follows a probability density function (PDF) such that:

$$f_{\min}(x) = n(1-x)^{n-1} \qquad (2)$$

The PDF derived from equation 2 is known as a Beta distribution with parameters α=1 and β=*n*. At each voxel, selecting the minimum *P*-value for the top gene then follows a Beta(1, *n*) distribution, where *n* is the effective number of independent tests.

However, genetic loci are inherited in contiguous segments, and some genes co-segregate in blocks. The allele frequencies and structure of genes that co-segregate are more similar than would be expected by chance if all variants were assumed to be independent. Because of this, the effective number of independent tests (M_{eff}) is less than the total number of tests performed (M). By determining M_{eff}, we can more accurately estimate the total number of independent tests performed, given the LD structure of our genotype data.

In our sample, we estimated M_{eff} by performing 5000 permutation tests at three randomly selected, uncorrelated voxels in the brain. We regressed each of the 18,044 genes on the permuted residuals of the reduced model after including the age and sex covariates at each run, and stored the minimum *P*-value. As only the minimum *P*-value is retained (for the best fitting gene), one can build up a reference distribution for the minimum *P*-values, to help gauge the level of surprise in seeing associations in the data. Storing the minimum *P*-values of the permutation tests yields the expected null Beta distribution given our data. We used a maximum-likelihood function to estimate the best fit for the null Beta distribution by varying the β parameter of Beta(1,β). The value of β approximates the effective number of independent tests (M_{eff}) performed on our data.

2.5 Estimation of expected values in simulated maps

A certain amount of spatial smoothness is expected among voxels in an image. This is most likely explained by the non-independence of volume difference measures at adjacent voxels. We examined whether the size of voxel clusters associated with the same gene from our vGeneWAS analysis differed from the cluster sizes expected under the null hypothesis of no association at all, given the non-independence of signals at adjacent voxels in our images. In addition, we wanted to determine whether the number of unique, top genes from across the brain significantly differed from the number of top genes expected by chance. We generated 100 3D simulated cluster maps based on a linking algorithm that forms connections between voxels across the brain based on correlation. The probability of any voxel being linked to another voxel was directly related to how correlated they are to each other. By considering the correlation of a given voxel to all other voxels in the image, as opposed to using a single summary measure of smoothness throughout an image, we were able to model the expected 3D clustering among adjacent voxels and non-independent, spatially separated clusters.

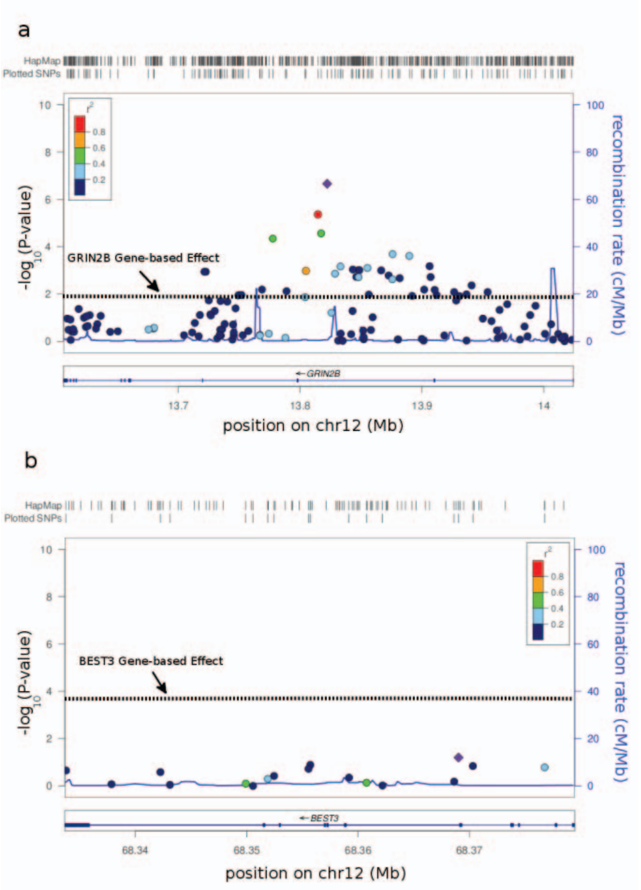
3. RESULTS

3.1 Comparison of methods

To examine differences between gene-based and standard univariate association methods, we compared the results of PCReg to linear regression using the temporal lobe volume (TLV) data from a previous study [1] as the phenotype. We first chose to focus on the top gene or SNP identified by each method, in order to examine performance when the variant chosen is deliberately selected to favor one of the two methods. *GRIN2B* was identified as the gene with the SNP variant that was most significantly associated with TLV using a standard univariate GWAS analysis (*P*=$4.03\text{x}10^{-7}$). We plotted the –$\log_{10}$(*P*-value) of the univariate test for each of the SNPs in the *GRIN2B* gene, in **Figure 1a.** The PCReg gene test results are overlaid (*black dotted line*). Clearly, the main effect detected with linear regression is much greater in this case, and the p-values are much smaller (i.e., –$\log_{10}$(*P*-value) is higher). Notably, we tested each of the 129 SNPs within the *GRIN2B* gene, which would require any significant *P*-values identified to be corrected for multiple comparisons before further study. In comparison, the gene-based test of *GRIN2B* using PCReg was a single test not requiring correction for multiple comparisons and maintained a nominal significance value (*P*=0.012). Also, we compared *BEST3* - the gene identified to be most significantly associated with TLV via PCReg - with the linear regression output of each SNP within the gene (**Figure 1b**). The main effect of the gene-based test was much stronger (*P*=$2.9\text{x}10^{-4}$) than the best linear regression result (*P*=0.063). This demonstrates a case where variance components from individual markers are not significant via linear regression, but may be combined into a single significant test statistic.

Figure 1. Genetic association plots for univariate linear regression versus multi-locus PCReg. The -$\log_{10}$(*P*-value) of each SNP in *GRIN2B* (a) and *BEST3* (b) is plotted against

its position in the gene. Each of the points is color coded by level of LD (compared to the top SNP, *the purple diamond dot*) as measured by r^2.The -$\log_{10}$(P-value) of the gene-based PCReg test for each gene is overlaid on the plot for comparison (*dotted black line*). Plots were generated using the LocusZoom software package.

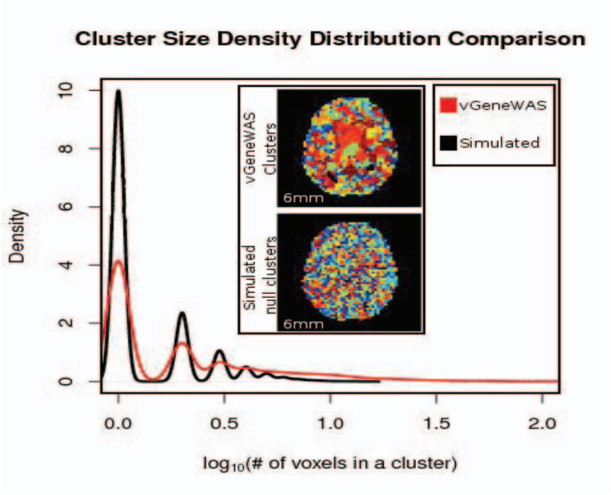


3.2 Voxelwise GeneWAS

By randomly permuting the images, so that they were not assigned to the correct individuals, we compared the distribution of the cluster size values in simulated (null) maps to the cluster sizes obtained from vGeneWAS (**Figure 2**). A large proportion of clusters of voxels associated with the same top gene in vGeneWAS were larger than would be expected based on completely null data. One estimate related to the number of independent voxels is the average number of clusters in simulated maps. This was 11900.8 ± 50.6 (mean ± standard deviation) out of the 31,662 total voxels. We used the number of clusters estimated from the simulation to randomly select (with replacement) from our list of 18,044 genes. We tallied the number of unique genes represented for each simulated cluster map and found the average was 8721.4 ± 44.9 (mean ± standard deviation). We measured the total number of unique genes as 5333 from our run of voxelwise GeneWAS, which is much lower than the number of genes expected based on the null cluster maps. Combined with our cluster size comparisons, this suggests that the top genes identified in our analysis tend to have a much more broadly distributed effect than would be expected if the data were null, even taking into account the intrinsic spatial non-independence of our data.

Among the top genes identified at each voxel across the brain, the *GRB-associated binding protein 2 gene*, *GAB2*, was the most significantly associated gene at any voxel (with P=2.36x10^{-9}) in our analysis and has previously been linked to late-onset Alzheimer's disease (LOAD). One study [9] identified 10 SNPs from the *GAB2* gene that were significantly associated with LOAD and *APOE* allele status in 1411 cases and controls from 20 NIA-sponsored Alzheimer's Disease Centers. *In vivo* testing shows that *GAB2* is over-expressed in certain brain regions such as the hippocampus and posterior cingulate cortex in patients with LOAD [9]. In addition, the *AlzGene* website lists *GAB2* as being in the top 20 genes likely related to AD (October 20, 2010; http://www.alzgene.org/). We identified several other genes highly relevant to brain function; a few are: *LRDD* (P=2.60x10^{-9}), *PRPRB* (P=2.84x10^{-9}), *CHRM5* (P=1.71x10^{-8}), and *S100B* (P=4.75x10^{-8}).

Figure 2. Cluster sizes in vGeneWAS (*red line*) **are compared with a simulated null map** (*black line*). The density of the number of voxels ($\log_{10}$ transformed) in a cluster across the brain are plotted. The simulated null map contains a larger proportion of small cluster sizes than vGeneWAS (higher peaks in the *black line* at values close to the origin on the *x*-axis). The vGeneWAS map contains a larger proportion of large cluster sizes than the average simulated null map (the *red line* is higher at larger values and is more extended). A single slice view of the vGeneWAS and average simulated null cluster maps are pictured for comparison (*inset*). Every unique cluster is assigned its own color. There are more unique clusters than distinct colors making visual inspection difficult, but in general the clusters in the vGeneWAS maps are larger.



3.3 Correction for multiple comparisons

Our Beta-distributed experimental P-values (with M_{eff}=15,636) need to be corrected so that their false discovery rate (FDR) can be assessed [10]. Using the analytic β parameter from the null Beta distribution, we fitted a cumulative distribution function (CDF) to our observed data yielding a new distribution of corrected P-values that deviate from the uniform distribution only when the data are not null.

We found that the false discovery rate for the second most highly associated gene in our results (*LRDD*) could only be controlled at a threshold of q=0.30 (i.e., allowing a 30% false discovery rate) after applying a statistical threshold of P_c=5.36x10^{-4}. In addition, the pFDR q-value threshold [11] was q=0.23 for the most significantly associated gene at any voxel (*GAB2*). In other words, the vGeneWAS results could not be controlled at the conventional false discovery rate, but show promise.

3.4 Power comparisons

To assess the differences in power afforded by vGeneWAS relative to existing univariate methods, we compared the P_c-values from vGWAS obtained in our previous study [3], with the P_c-values resulting from vGeneWAS (**Figure 3**). The proportion of P_c-values greater than a given FDR threshold for each method is directly related to differences in effect sizes. The FDR of the results from vGWAS could only be controlled at a threshold value of q=0.50, whereas the FDR threshold for vGeneWAS is somewhat lower, although not passing the conventional FDR level (q=0.30; **Figure 3**). This suggests that the vGeneWAS method may have more power, in principle, to detect genetic associations, although neither test controlled the false discovery rate at the conventional level.

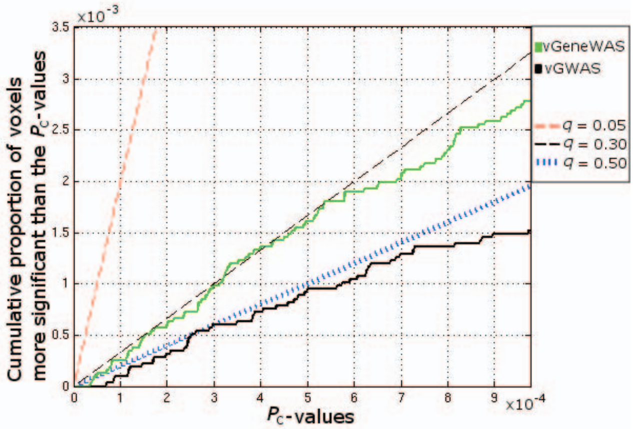


Figure 3. vGeneWAS may control the false discovery rate better than vGWAS. The cumulative distribution function (CDF) of P_c-values from vGeneWAS (*solid green line*) is compared to the CDF of P_c-values from vGWAS [3]. (*solid black line*). Three lines represent different correction thresholds of q=0.05 (*red dashed*), q=0.30 (*black dashed*), and q=0.50 (*blue dotted*).

4. CONCLUSION

We showed that, in certain cases, gene-based methods may offer more power than traditional univariate methods. In addition, our analysis identified a known Alzheimer's risk gene, *GAB2*, lending plausibility to the method. Still, effect sizes may be too small to detect even with multivariate statistics and meta-analytic approaches may prove most useful in the future (e.g., in multi-site efforts such as the ENIGMA consortium [12]).

5. ACKNOWLEDGMENTS

This study was supported by NIH grants EB008281 HD050735, RR013642, AG036535 (to PT).

6. REFERENCES

[1] Stein, J.L., et al., 2010. Genome-wide analysis reveals novel genes influencing temporal lobe structure with relevance to neurodegeneration in Alzheimer's disease. **Neuroimage** 51, 542-554.

[2] Hibar, D., et al., 2010. Voxelwise genome-wide association of Diffusion Tensor Images identifies putative novel variants influencing white matter integrity in 467 related young adults. **Society for Neuroscience**, San Diego, CA.

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

[4] Neale, B.M., Sham, P.C., 2004. The future of association studies: gene-based analysis and replication. **Am J Hum Genet** 75, 353-362.

[5] Hua, X., et al., 2008. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: An MRI study of 676 AD, MCI, and normal subjects. **Neuroimage** 43, 458-469.

[6] Cannon, T.D., Keller, M.C., 2006. Endophenotypes in the genetic analyses of mental disorders. **Annu Rev Clin Psychol** 2, 267-290.

[7] Saykin, A.J., et al., 2010. Alzheimer's Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. **Alzheimers Dement** 6, 265-273.

[8] Wang, K., Abbott, D., 2008. A principal components regression approach to multi-locus genetic association studies. **Genet Epidemiol** 32, 108-118.

[9] Reiman, E.M., et al., 2007. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. **Neuron** 54, 713-720.

[10] Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society Series B-Methodological** 57, 289-300.

[11] Storey, J.D., 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. **Annals of Statistics** 31, 2013-2035.

[12] The ENIGMA Consortium, 2011. Genome-wide association meta-analysis of hippocampal volume: results from the ENIGMA consortium. OHBM, June 2011.