Power Estimates for Voxel-Based Genetic Association Studies using Diffusion Imaging

Neda Jahanshad[1], Peter Kochunov[2], David C. Glahn[3],
John Blangero[4], Thomas E. Nichols[5,6], Katie L. McMahon[7],
Greig I. de Zubicaray[8], Nicholas G. Martin[9], Margaret J. Wright[9],
Clifford R. Jack, Jr.[10], Matt A. Bernstein[10], Michael W. Weiner[11,12],
Arthur W. Toga[1], and Paul M. Thompson[1] for ADNI

[1] Imaging Genetics Center, Lab of Neuro Imaging, UCLA, Los Angeles, CA USA
[2] Maryland Psychiatric Research Center, U of Maryland, Baltimore, MD USA
[3] Olin Neuropsychiatry Research Center, Yale School of Medicine, New Haven, CT
[4] Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX
[5] Department of Statistics & Warwick Manufacturing Group, U of Warwick, Coventry, UK
[6] Oxford Centre for Functional MRI of the Brain (FMRIB), Oxford University, UK
[7] University of Queensland, Centre for Advanced Imaging, Brisbane, Australia
[8] University of Queensland, School of Psychology, Brisbane, Australia
[9] Queensland Institute of Medical Research, Brisbane, Australia
[10] Department of Radiology, Mayo Clinic, Rochester, Minnesota, USA
[11] Department of Radiology, Medicine, and Psychiatry, UC San Francisco, CA, USA
[12] Department of Veterans Affairs Medical Center, San Francisco, CA, USA

Abstract: The quest to discover genetic variants that affect the human brain will be accelerated by screening brain images from large populations. Even so, the wealth of information in medical images is often reduced to a single numeric summary, such as a regional volume or an average signal, which is then analyzed in a genome wide association study (GWAS). The high cost and penalty for multiple comparisons often constrains us from searching over the entire image space. Here, we developed a method to compute and boost power to detect genetic associations in brain images. We computed voxel-wise heritability estimates for fractional anisotropy in over 1100 DTI scans, and used the results to threshold FA images from new studies. We describe voxel selection criteria to optimally boost power, as a function of the sample size and allele frequency cut-off. We illustrate our methods by analyzing publicly-available data from the ADNI2 project.

Keywords: Neuroimaging genetics, heritability, GWAS, DTI, multiple comparisons correction.

1 Introduction

Imaging genetics is an emerging field in which variations in the human genome are related to brain differences. Genome-wide association studies (GWAS), for example, test for statistical associations between brain measures and up to a million single nucleotide polymorphisms (SNPs[1]) in the genome. To simplify the screening effort, studies often focus on a single measure extracted from brain scans, such as the overall volume of the hippocampus [1]. Despite successful findings with simple summary measures, the image space contains many more features, e.g., at individual voxels, which can be used for genetic testing. Testing all these features, however, imposes a high cost of multiple comparisons. GWAS tests at each voxel are computationally feasible but underpowered, due to the large number of tests performed [2]. Efforts to boost voxelwise power include set-based multivariate tests (multilocus genetic tests), ICA [3], and sparse regression methods in the image and/or genome [4-6]. If we are able to analytically reduce the image space *a priori* to only those regions where we expect sufficient power for association, we could maximize our search space while minimizing the penalty for multiple comparisons.

Power in GWAS is a topic of great interest, as any algorithms to boost power would make genetic studies more efficient, faster, or more cost-effective. Power calculations can inform data collection, i.e., how many subjects to recruit. Power can be computed for a GWAS analysis of a single measure, or (conversely) to find how many subjects are required to detect a pre-defined effect size on a brain measure, for one single nucleotide polymorphism (SNP) with a given minor allele frequency (MAF). One commonly used tool for calculating power is the Genetic Power Calculator [7]. Power and required sample sizes can be computed for association tests with binary outcomes (e.g., diagnosis in a case-control study) [8], or quantitative traits such as brain measures. Another approach to optimize power is to select subjects non-randomly, to optimize power to detect effects of rare variants [9]. However, in large neuroimaging studies, often data have already been collected, sometimes without genetic discovery in mind, and the challenge is to compute image features to maximize power to detect genetic associations. A key goal is to prioritize the list of >100,000 possible voxelwise phenotypes[2] to only those that meet a desired expected power threshold, of say 80%, for a given sample size, N.

Power calculators depend on distributional assumptions, and may require certain parameters to be known, such as the means and variances of the trait in each of the genotype groups. Often, we may not know these means and standard deviations [10]. One approach [11] suggests the use of heritability to avoid the need to specify the mean and variance required for the non-centrality parameter. Other common assumptions are that allele frequencies and genotype counts are known. However, these are often not known beforehand, and ignoring uncertainties can lead to overly optimistic power estimates [12]. Even with a fixed population sample with allele frequencies in Hardy-Weinberg equilibrium (HWE), the genotype distribution of the actual sample can vary. It becomes essential to compute the expected power by taking into account the expected genotype distribution, as has been shown for single alleles [12, 13]. While these problems have been addressed in the context of single variants, with known or unknown allele frequencies, p, power analysis for GWAS studies is still challenging. In GWAS, over one million alleles may be tested, all with different

---

[1] At each location on the genome, a person has a specific "letter" or nucleotide; SNPs are common variants in the genetic code, carried by at least 1% of the population.

[2] A *phenotype* is a biological measure that is subjected to genetic analysis, such as the size of a brain region, or a diffusion imaging measure in a specific region.

minor allele frequencies (MAFs). For gene discovery, including GWAS studies of brain imaging phenotypes, all these alleles will have unknown and varying genotype counts, although a limit can be set as to how much the SNP deviates from HWE, where HWE represents the expected genotype frequencies for each allele given the MAF in the population of interest. If, for example, the allele MAF=0.3 for a given population, then with a sample size of N=1000, $N*p^2$=90 people in the cohort would be expected to have 2 copies of the minor allele (n_2), $2*N*p*(1-p)$=420 people should have 1 copy, (n_1), and $N*(1-p)^2$ = 490 people should have 0 (n_0). Deviations from this count are assessed using a χ^2 test with one degree of freedom (dof); samples with genotype counts that differ significantly, are not in HWE.

Studies using DTI have reported moderate correlations between DTI-based measures of fractional anisotropy (FA) and common variants in specific candidate genes (e.g., *CLU*, *HFE*, *NTRK1*, and *BDNF*) known to associate with neurological diseases and disorders. This motivates the use of FA as phenotypes for GWAS, as they may implicate genes that affect disease risk. Recently, we [14] computed the first large-scale heritability study on voxelwise FA maps, using meta-analysis methods on two cohorts comprising a total of over 1100 subjects. Here, we use these existing voxelwise measures of heritability to estimate the power of GWAS at each voxel of a new dataset. We limit our multiple comparisons correction problem by filtering SNPs based on a user-defined threshold for the MAF. We compute the expected power at each voxel, with appropriate limits on the number of subjects in the genotype groups, based on the HWE. This allows us to focus on voxels of the FA map where we would expect a certain level of power for detecting a genetic effect, given a specific N.

2 Methods

2.1 Heritability and Power Estimates

We can use linear regression to test the effect of a genetic variant on a quantitative trait, such as a brain imaging measure. In this way, the expected value for the trait is modeled as $\mu_i = \mu_0 + \beta * i + \beta_x * x + \varepsilon$, where i =0,1,2 represents the number of copies of the minor allele, and β_X and x represent the nuisance variables such as age and sex. Here the F-ratio can be used to test the equality of the means (H_0: β=0, H_A: $\beta \neq 0$). The F-ratio follows a non-central F-distribution with 1 and N-2 dof, with a non-centrality parameter λ: $\lambda = \frac{\left(\sum_i n_i (i-\bar{x})(\mu_i - \mu)\right)^2}{\sum_i n_i (i-\bar{x})^2 \sigma^2}$,where $\bar{x} = \frac{\sum_i n_i * i}{\sum_i n_i}, \mu = \frac{\sum_i n_i \mu_i}{\sum_i n_i}$ and σ is the within-group SD. The power is $\pi = \Pr\left(F_{1,N-2,\lambda} > F_c\right)$, where F_c is the (1-α)*100 percentile of the central F-distribution, and α is the significance threshold (e.g., 0.05). Note that if we add d covariates to the model, our distribution would have 1 and N-2-d dof. However, the means and variances within the groups may be unknown - or impractical to calculate - for each of the possible GWAS SNPs.

Alternatively, the heritability of a trait (h^2) is defined as the proportion of the observed variance in a trait (here a brain measure) that is attributable to genetic variation; $h^2 = V_g/V_T$ – or, the model sum-of-squares, divided by the total sum-of-squares (SSM/SST= 1-(SSE/SST)). As suggested in [11], the F-statistic can be directly calculated as a function of h^2 and N. This is because according to an ANOVA table for a regression (**Table 1**), the F-statistic=(SSM*DFE)/(SSE*DFM); as SSE= (1-h^2)*SST and SSM= h^2*SST, then $F = \frac{h^2 * DFE}{(1-h^2) * DFM}$, and in the case of one explanatory variable, $d=1$, $F = \frac{h^2 * (N-2)}{1-h^2}$. We can therefore use the trait's h^2 to calculate power, π.

Table 1 ANOVA table for a multiple regression model, with d explanatory variables.

ANOVA	dof	Sum of squares	Mean squares	F
Model	DFM=1	SSM= $\sum_j (\hat{y}_j - \bar{y})^2$	MSM=SSM/DFM	MSM/MSE
Error	DFE=N-d-1	SSE= $\sum_j (y_j - \hat{y}_j)^2$	MSE=SSE/DFE	
Total	DFT=N-1	SST=SSM+SSE	MST=SST/DFT	

Recently, we [14] meta-analyzed voxelwise measures of heritability from FA maps in two family-based cohorts totaling over 1100 subjects. One cohort was a sample of Caucasians imaged at 4T with 94 directional gradients and voxels of size 1.7x1.7x2mm, and the other was a Mexican-American sample imaged at 3T with 64 gradients and 1.8x1.8x3mm voxels. Given differences in cohort demographics and scanning protocols, and the overall large sample size (1100+), we treat these meta-analyzed values as the best available heritability estimates for the voxelwise phenotypes in any studies of FA mapped to the same space. We carry these heritability measures forward, to estimate power in a new GWAS study of unrelated individuals. We express the power as a function of N, calculating the probability under the F-distribution (1 and N-2 dof) for the heritability at every voxel.

2.2 HWE, MAF, and multiple comparisons correction

Tests of HWE assess whether the sample genotype frequency is aligned with the frequency expected based on reference data compiled from human populations (such as *HapMap*). Including SNPs with significant deviations from HWE can be problematic [16], but the threshold for determining significant deviations is often different between studies [17]. Thresholding this deviation at more stringent values eliminates group formations that could artificially appear to enhance power. A direct filter on the SNPs involves thresholding the MAF, to remove rare variants and SNPs with allele frequencies too low to be adequately sampled. Given the relatively lower N in imaging studies compared to other genetic studies, filtering out low frequency SNPs can reduce the number of tests, alleviating the multiple comparisons penalties, without compromising power. We filter SNPs based on MAF, and adjust possible outcomes with HWE deviations. Once SNPs have been filtered, the significance threshold can be based on the number of SNPs (N_{SNPs}). The false discovery rate (FDR) q-value which ensures that the false positive rate across all *voxels,* is controlled at $q_{FDR} = 0.05/N_{SNPs}$. We count the number of voxels (N_{VOX}) that pass the threshold. Now power may be calculated as $\pi = \Pr(F_h > F_c)$, with F_c determined by $\alpha = 0.05/N_{SNPs}/N_{VOX}$.

2.3 Accounting for uncertainties in genotype frequency

As mentioned in [12, 13], for a given allele frequency p, the expected power is the weighting of the estimate π (here a function of h^2 and N) by the probabilities of the multinomial counts, $\mathbf{n}=(n_0, n_1, n_2)$. By contrast with prior papers, we limit **n** to only those combinations that fall within HWE_c.

Additionally in this work, as opposed to those mentioned above, we are conducting a GWAS of all SNPs with $p > MAF_c$ (a lower bound on the MAF of the SNPs) rather than a single SNP. Therefore, rather than setting a beta-prior on p, we sum over all possible frequencies, each one with respect to only the possible **n** in HWE, and all possible $p > MAF_c$. Note that, by definition, the $MAF \le 0.5$.

$$E(v)_{GWAS}\left[\pi(h^2,N)\,|\,\mathbf{n}_{HWE}\right]=\sum_{p=MAF_c,\Delta p}^{0.5-\Delta}\frac{\Pr(p,p+\Delta)}{\Pr(MAF_c \le p \le 0.5)}\sum_{\sum n_i=N;n_i\ge 0\forall i}\pi(h^2,N)\frac{N!}{n_0!n_1!n_2!}2^{n_1}p^{2n_0+n_1}(1-p)^{n_1+2n_2}$$

This estimates power for GWAS at the individual voxel, v – and we only include those voxels that pass the FDR correction as mentioned above. The estimated power is a weighted function of: (1) the probability that the p falls within a given interval, Δ, and (2) the probability of the genotype group distribution within HWE. π can now be taken out of both summations, and a single weight, $wt(MAF_c,HWE_c)$, is given to each voxel. $E(v)$ can be written as $\pi(h^2(v),N)*wt$. As an example, we use data from the second phase of the Alzheimer's Disease Neuroimaging Initiative study (ADNI2; publicly available at http://adni.loni.ucla.edu), and show the proportion of SNPs that fall across a range of allele frequencies (**Figure 1**). Using the formulation above, we compute the N_{VOX} that surpass a power threshold while varying the N_{SNPs} examined by varying MAF_c and HWE_c, as well as how this estimate is modified by sample size, N.

2.4 Voxelwise GWAS of the ADNI2 dataset

To date, ADNI2 has genotyped 78 Caucasian subjects scanned with DTI. Scans were processed and aligned to the heritability map detailed in [14]. Limiting our search space to only voxels that pass the 0.8 power threshold with $MAF_c=0.1$ and $HWE_c=1\text{x}10^{-5}$, we ran a GWAS at all remaining voxels using an additive model; this is modeled with a linear regression for each SNP.

3 Results

3.1 Voxels with power > 0.8 as functions of N, MAF_c, HWE_c

The whole-brain white matter skeleton of the DTI FA-template consists of 97158 voxels total. In **Figure 2**, we show how the expected N_{VOX} changes as we threshold the MAF from 0.01 to 0.2 in step sizes of 0.01, and as we become more stringent with the HWE_c. We do not eliminate SNPs due to divergence, but rather filter out the possible combinations of subjects that could result in the given p. Therefore, as we are more strict with HWE_c ($1x10^{-6} < \ldots < 0.01 < 0.05$), more possibilities are deemed divergent and possibly biased, so power is lower, and fewer voxels are selected.

3.2 Voxelwise GWAS in the ADNI2 dataset

In our real data analysis, we first find the number of voxels that will have an expected power $E_{GWAS}(v) > 0.8$ with $MAF_c = 0.1$ ($N_{SNPs} = 522{,}077$ out of 661,478), in the ADNI2 dataset of 78. As this is a dataset including several patients with Alzheimer's disease, we would not necessarily expect the HWE to be upheld, as disease could impact the frequency of some alleles – especially AD risk alleles. We were therefore less strict with HWE_c, selecting $p<10^{-5}$ as our cutoff. We obtained a mask consisting of 1664 voxels (**Figure 3A**), and ran a GWAS at all voxels.

When correcting for all the voxels and the SNPs tested, no SNP had a significance level less than $0.05/(1664*522{,}077) = 5.75x10^{-11}$ in the ADNI2 sample. However, as these voxels were chosen to have the highest power to detect a genetic association, we show that when plotting the minimum *p*-values at each voxel, against what would be expected from a null distribution (here, a Beta distribution with parameters ($1,N_{SNPs}$) [18]) in a Q-Q plot (**Figure 3B**), we show our *p*-values tended to be lower than what would be expected from a null distribution.

4 Discussion

We have presented a method to estimate power for imaging genetics studies that apply genome-wide scanning to multiple phenotypes; in fact, the same method could also be adapted to prioritize targets for genetic analysis. Most power estimates for GWAS studies are computed to plan future data collection, but here we assume imaging data has already been collected (or there is an upper limit on *N*). We then show how to optimize the study to focus the GWAS on the most powerful regions of the high-dimensional image space. Our approach is also flexible - users can select a MAF or HWE cutoff to estimate the best thresholds to optimize the search space.

Limitations of our study include the following: (1) the heritability presented here may differ from that of the population a user is studying. However, we used the best available estimates of heritability of FA to date, given the large sample size, and the use of meta-analysis to mitigate differences in demographics and imaging protocols; (2) Studies have also shown that choosing the inappropriate model for the SNP (e.g., additive when in fact it has a dominant effect) can bias power calculations [13]. Even so, most GWAS studies assume an additive model for each SNP, so we follow this convention. (3) Here, we do not account for the correlation between voxels, or the linkage disequilibrium (LD) structure of the SNPs. If LD is considered, tests are not all independent and we can reduce the number of effective tests, reducing the multiple comparisons penalty. Our method can incorporate this correlation, so our power estimates (assuming SNP independence) are somewhat conservative. Regardless, our methods can estimate the power of voxelwise association tests, providing a starting ground for GWAS of spatially extended phenotypes.

Additionally, it is possible that limiting rarer variants by removing the SNPs with low MAF, may result in removal of SNPs that may have the greatest effect sizes, however with a limited number of scans, it is unlikely that the effect of these rare variants would be accurately obtained.

Here, we show that meta-analyzed heritability estimates are useful for genetic studies on the entire image space. We focus on voxelwise analyses, but so long as heritability estimates have been made on multiple imaging phenotypes, this method can be extended beyond maps of FA values, to mean volume or anisotropy values within regions of interest, or to all of the individual network elements that make up the human connectome.

References

1. Stein, JL, et al., *Identification of common variants associated with human hippocampal and intracranial volumes*. Nature Genetics, 2012. **44**(5): p. 552-61.
2. Hibar, DP, et al., *Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects*. NIMG 2011. **56**(4) 1875-91.
3. Calhoun, VD et al., *A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data*. NIMG 2009. **45** S163-72.
4. Ge, T., et al., *Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures*. NIMG 2012. **63**(2) 858-73.
5. Vounou, M et al., *Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease*. NIMG 2012 **60**(1) 700-16.
6. Silver, M., et al., *Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression*. NIMG 2012. **63**(3)1681-94.
7. Purcell, S et al., *Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits*. Bioinformatics, 2003. **19**(1): p. 149-50.
8. Menashe, I, et al., *PGA: power calculator for case-control genetic association analyses*. BMC Genetics, 2008. **9**: p. 36.
9. Guey, LT, et al., *Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants*. Genet Epid, 2011.
10. Minikel, EV *Power for GWAS and extreme phenotype studies*. 2012 [cited 2013 2/8/2013]:http://www.cureffi.org/2012/12/05/power-for-gwas-and-extreme-phenotype-studies/.
11. Feng, S, et al., *GWAPower: a statistical power calculation software for genome-wide association studies with quantitative traits*. BMC Genet, 2011. **12**: p. 12.
12. Ambrosius, WT, et al, *Power for genetic association studies with random allele frequencies and genotype distributions*. Am J of Hum Genet, 2004. **74**(4): p. 683-93.
13. Kozlitina, J, et al., *Power of genetic association studies with fixed and random genotype frequencies*. Annals of Human Genetics, 2010. **74**(5): p. 429-38.
14. Jahanshad N, et al., *Multi-site genetic analysis of diffusion images and voxelwise heritability analysis: A pilot project of the ENIGMA-DTI working group*. Neuroimage. In Press, 2013
15. Yang, J, et al., *Common SNPs explain a large proportion of the heritability for human height*. Nat Genet, 2010. **42**(7): p. 565-9.
16. Trikalinos, TA et al., *Impact of violations and deviations in Hardy-Weinberg equilibrium on postulated gene-disease associations*. Am J Epid 2006 **163**(4)300-9

17. Chan, EK et al., *The combined effect of SNP-marker and phenotype attributes in genome-wide association studies*. Animal Genet, 2009. **40**(2) 149-56.

18. Stein, JL et al., *Voxelwise genome-wide association study (vGWAS)*. NIMG 2010. **53**(3) 1160-74.

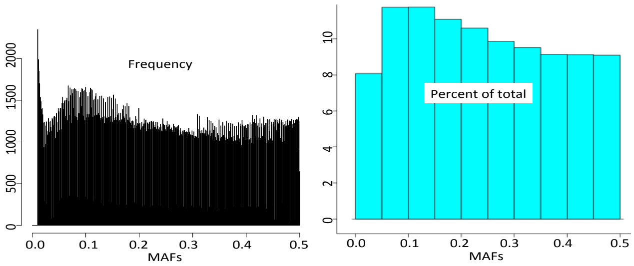

Figure 1 Histograms of frequency and percent (Δ=0.001,0.05) of SNP MAFs in ADNI2.

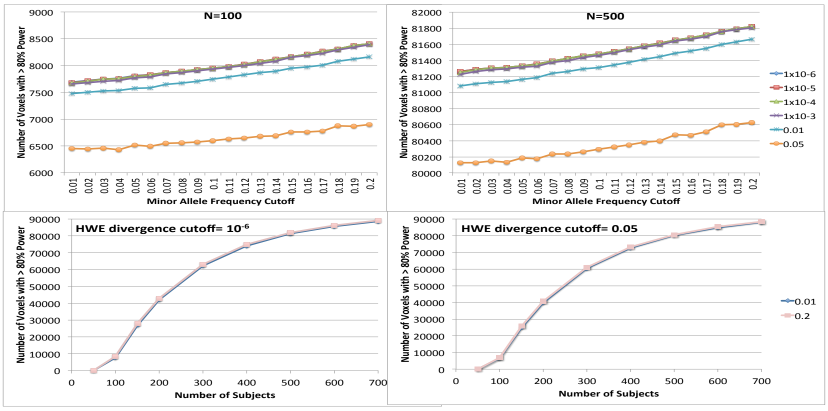

Figure 2: The *top row* shows the number of voxels with sufficient power to detect a genetic effect (>80% power) with respect to the MAF cutoff (MAF_c) and the divergence threshold for HWE (HWE_c), for samples with N=100 and N=500 subjects. As the MAF threshold is increased, fewer SNPs are tested, and this reduces the multiple comparisons correction and increases power for those tested SNPs, if there is a true associated variant in the group. As in prior GWAS power studies, the *bottom row* shows the number of subjects, regardless of MAF_c or HWE_c, and has a far larger influence than either one. The number of voxels with sufficient power levels off here, as N > 500.

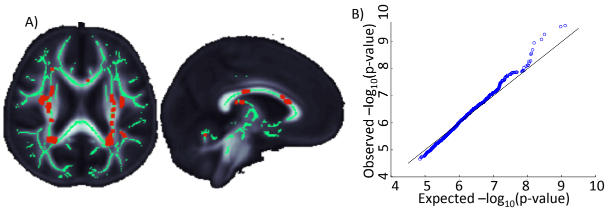

Figure 3: The 1664 voxels found to have sufficient power (> 80%) in the ADNI2 sample (N=78) were used for GWAS, and are *highlighted in red*. Plotting the minimum p-value at each voxel against what would be expected from the null hypothesis shows trends towards significance in this subsample of voxels.