CANCER GENOMICS & PROTEOMICS *4*: 27-34 (2007)

Guideline for Data Analysis of Genomewide Association Studies

HEPING ZHANG, LEI LIU, XUEQIN WANG and JEFFREY R. GRUEN

Yale University School of Medicine, New Haven, CT, U.S.A.

Abstract. *Intensive efforts have been underway to identify common genetic factors that influence health and disease including cancer using genomewide association studies (GWAS). Our experiences have shown that while it is more advantageous to have large detailed data sets, the amount of information generated by GWAS also present major challenges for statistical analyses. While prospects for the oncoming flood of GWAS is exciting, the tools for conducting and evaluating these studies are still in early developmental stages creating some investigator uncertainty and prompting conferences and workshops specifically devoted to these topics. In this review, we summarize important steps for planning the statistical analysis involving genome-wide data from single nucleotide polymorphisms (SNPs). This review is purposely meant to be relatively short and of practical use for the space constraints of typical federal grant proposals.*

The National Institutes of Health (NIH) have issued a number of announcements to advance genome-wide association studies (GWAS) to identify common genetic factors that influence health and disease. Recent successes suggest that the information derived from such studies will help to develop new approaches to reduce disease burden and promote health (37). In a recent request for information (NIH, NOT-OD-06-094), a GWAS was defined as "any study of genetic variation across the entire human genome that is designed to identify genetic associations with observable traits (such as blood pressure or weight), or the presence or absence of a disease or condition". We are fortunate to have had early opportunities to design and participate in several large -scale GWAS, including the NIH National Genomic and Proteomic Network for Preterm Birth Research (RFA-HD-04-002). Our experiences have shown that while it is more advantageous to have large detailed data sets, the amount of information generated by GWAS also present major challenges for statistical analyses. While prospects for the oncoming flood of GWAS are exciting, the tools for conducting and evaluating these studies are still in early developmental stages creating some investigator uncertainty and prompting conferences and workshops specifically devoted to these topics (12, 36, 42). In this review, we summarize important steps for planning the statistical analysis involving genome-wide data from single nucleotide polymorphisms (SNPs). This review is purposely meant to be relatively short and of practical use for the space constraints of typical federal grant proposals.

Correspondence to: Heping Zhang, Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520-8034, U.S.A. e-mail: Heping.Zhang@Yale.EDU

Key Words: Genomewide association, SNP, haplotype, multiple comparisons, study design.

Descriptive Statistics and Data Quality Control

As in any statistical analysis, it is important to scrutinize the data and assess the quality before a formal statistical inference is performed. In addition to genotype data, all GWAS are expected to have detailed demographic and outcome data. Descriptive statistics for the outcome variables and covariates should be prepared and examined. For genotype calls, the associated software system, such as Affymetrix's GTYPE <http://www.affymetrix.com>, usually has some built-in quality assurance functions. Steps should be taken to assess concordance with Hardy-Weinberg equilibrium (HWE) and to scrutinize pedigree data for Mendelian inheritance errors. Potential non-random missing data also warrant attention for high-throughput SNP platforms. Copy number variation can lead to odd patterns of errors, and we need to examine missing data in genomic regions where such variation is present to resolve the errors. Inconsistencies and errors as a result of SNP genotyping should be resolved and corrected prior to starting any analysis. Some widely available software packages such as Merlin estimate error probabilities, adherence to HWE, and other preflight analyses.

Sample Size and Power

Sample size and power estimation are necessary in study planning and design, although investigators may tend to underestimate the uncertainty. In nearly all grant applications, investigators are tempted to assure a high level

of power (at least 80%). The frustration and failures of identifying susceptible genes for complex diseases are widely recognized with some exceptions (14). One of the major reasons is that there is generally little information with regard to the underlying genetic mechanisms. To consider the complexity and assure the power of GWAS, the NIH is promoting data sharing and the establishment of public databases to increase sample size. While investigators must acknowledge the inherent uncertainties and depict a realistic picture of what one study can accomplish, we illustrate here with an example as to how the sample size and power calculation may be presented. Because there is usually little information for genetic mechanisms, we recommend that investigators adopt conservative approaches for power estimation, including the Bonferroni method to correct for genome-wide type I error. In addition, we recommend use of different analytic strategies (described in depth below). The Bonferroni method is conventionally depicted as the most conservative approach to dealing with multiple comparisons, but our simulation based on HapMap data suggests that perhaps it is not as conservative as generally portrayed. The Bonferroni method provides a reasonable approximation even when the number of SNPs is in the range of 5,000 to 250,000 provided that they are reasonably spaced. However, if we use far more than 250,000 SNPs, the Bonferroni method could be over conservative because it appears that 250,000 is near the limit of statistically independent SNPs in the human genome according to our unpublished simulation.

Reduction of the number of tests. The obvious challenge with performing a test for each SNP is the large number of tests. The other problem is the potentially limited information in a SNP. Thus, it is natural and necessary to reduce the number of tests by either (a) selecting a subset of SNPs, and (b) considering haplotypes. For the purpose of our discussion, we use for comparison a GWAS for a complex disease (caused by an unknown number of environmental and genetic factors), recruitment of 1,000 cases and 1,000 controls, and a marker panel of 500,000 SNPs (currently available as the 500,000 randomly selected SNPs on the Affymetrix GeneChip®) (20).

One approach for reducing the total number of tests is to select a subset of SNPs using a two-stage strategy (29). In stage 1, half the cohort, 500 cases and 500 controls, are genotyped with the entire 500K SNP panel. After an interim analysis, in stage 2 the remaining 1,000 subjects are then genotyped and analyzed, presumably with a small subset of the first panel to validate candidates identified in stage 1. However, even with optimal control of type I error and power this strategy does not necessarily lead to substantial increased power over a single stage strategy. This is the first myth of a two-stage study design. The second myth is cost effectiveness. With the standardized large array SNP genotyping platforms now available, producing customized gene chips against specific targets is no longer a cost saving strategy unless the number of candidate SNPs is very small, *e.g.*, in the tens. Therefore, there is no cost saving incentive to examine a smaller set of SNPs, unless the investigators have the ability to design and produce a chip that is denser than the standard chip in regions of interest. With the rapidly evolving technology, the density and cost of SNPs may not be a major factor until very late in the process. For most studies that have little information to begin with, it is too early to consider a formal multiple stage design in the planning, although a follow-up study should be in place when highly promising SNPs emerge.

Selection of TagSNPs is another approach to reducing the number of SNPs for analysis. Using strong linkage disequilibrium in the dense SNPs may dramatically reduce the number of SNPs and hence the number of significance tests (4, 8, 25, 43). We will discuss below some of the existing approaches to selecting TagSNPs.

There is a general consensus that, if done appropriately, haplotype based analysis is more effective than single SNP based analysis. Thus it is important to conduct haplotype based association analysis in addition to SNP based analysis. It is reasonable to expect that the number of haplotypes to be examined would be far smaller than the number of SNPs, and the same study is likely to have enhanced power to assess the haplotypes under the assumption that haplotype relative risks and frequencies are comparable with those of a SNP. In reality, we do not know the exact situation. Most likely, haplotype relative risks and frequencies are higher than those for a SNP, leading to further improvement of the power and underscoring the rationale for haplotype based association analysis.

Sample size estimates. First, we examine the power to assess a SNP with 1000 cases and 1000 controls. For clarity, we can conceptually collapse the possible three genotypes (AA, AB, or BB) into two levels (high and low risk). All computations are based on the "Power" program (9, 18). Figure 1 displays the required sample size as a function of the risk genotype frequency for testing the association between the SNP and the disease. In this example, a sample size of 2,000 (1,000 cases and 1,000 controls) has sufficient power (>80% at alpha 0.0001) to identify a single polymorphism with frequency ranging .05-.12, and conferring a moderate OR of 1.7-2.0. This is comparable power to detect the ApoE4 allele conferring risk for age at onset of Parkinson disease (frequency = .05-.35, OR= 1.8) (16).

In Table I, we use Parkinson disease among 65 years of age or older and age-related macular degeneration (AMD) (14) as examples to examine the required sample sizes in

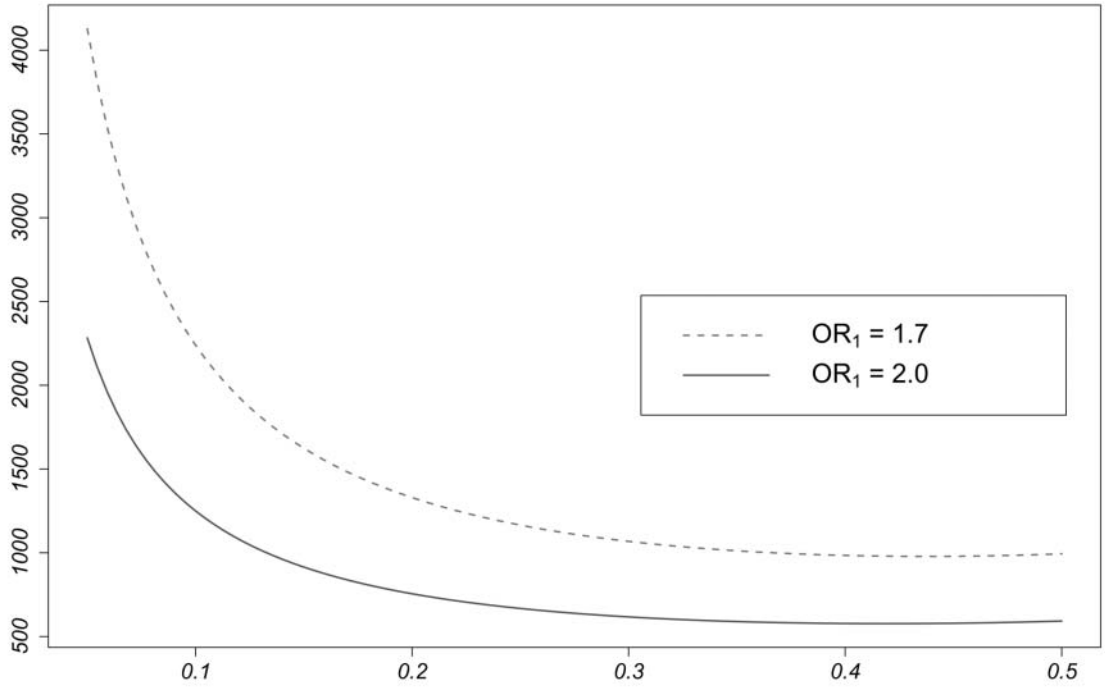

Figure 1. *The required sample size is plotted against the genotype frequency of a single SNP in order to detect a specified OR (odds ratio) (1.7 or 2). The risk in the population is assumed to be 2%. The type I error is controlled at 0.0001 and power at 80%.*

Table I. *Examples of clinically relevant polymorphisms. The type I error is controlled at 0.0001 and power at 80%. The mode of inheritance is assumed to be dominant.*

Disease	Allele	Risk	OR	Frequency	# of required cases	Reference
Parkinson Disease	ApoE4	15%	1.8	0.05	3588	(16)
				0.20	1128	
				0.35	839	
AMD	CFH	50%	7.4	0.36	99	(5, 14)

genetic studies of common diseases. For clarity, we assume the mode of inheritance is dominant. When the prevalence is high and risk allele is relatively common, a relatively small study would have sufficient power. However, in most studies, 1,000 cases and 1,000 controls would be reasonable sample sizes.

We mentioned earlier that haplotype based association tests may be more powerful than individual SNP based association tests. Morris (22) considered a variety of reasonable settings and reported that a sample of 1000 cases and 1000 controls is adequate to identify associations for moderate GRRs (genotype relative risk) in the range of 1.5. Here, GRR is the fold increase in risk for having a disease over the general population due to having a disease allele genotype. We refer to Morris (22) for the details.

SNP-based Analysis

We recommend that the analysis proceed in several steps: first examine a single SNP-based association, next a single gene based association, and finally gene-gene interactions within a biological pathway.

Testing a single SNP-based association. The analysis can be done through the standard logistic regression model by defining two dummy variables for three possible genotypes that produces the estimates for odds ratios (OR) and their confidence levels. This analysis is applicable regardless of the number of SNPs to be tested, although the overall *p*-value must reflect the number of tests.

Although we suggested the most conservative and simple approach – Bonferroni correction – in the power estimation,

potentially more powerful statistical strategies should be considered in the analysis to take advantage of advanced statistical techniques. For example, we can determine the statistical significance empirically through permutation test. The permutation can be done by randomly and repeatedly permuting the disease status in the cases and controls so that the disease status is expected not to be associated with any SNPs, except by chance. In other words, on the basis of our observed data, we can generate the data under the null hypothesis, which allow us to empirically obtain the distribution of the testing statistic under the null hypothesis and hence produce a genome-wide measure of significance. Another analytic approach for dealing with a large number of significance tests is the use of the false discovery rate (FDR). Instead of controlling for the genome-wide type I error, FDR focuses on the proportion of false positive results in the set of rejected null hypotheses (3). The same permutation procedure can be used to estimate the *p*-values for all SNPs. After sorting all p-values, a cut-off for the *p*-values is determined to achieve a given level of FDR such as 5%.

Another way to increase the power of an association test is to use and examine haplotype blocks by identifying TagSNPs. This will reduce the number of SNPs to be tested and retain as much information as possible. To this end, we can employ existing methods to identify haplotype blocks and TagSNPs. Since many dense SNPs are in strong linkage disequilibrium (4, 7, 25), haplotype blocks can be constructed based on marker-marker linkage disequilibrium estimates (1) and TagSNPs can be selected from the haplotype blocks to represent the polymorphisms within a block. It is noteworthy that the number of tagSNPs depends on LD patterns in a particular sample.

Haplotype-based association. Because SNPs are di-allelic, the information in an individual SNP may be limited. The informativeness of the markers, and hence the power of association tests, can be increased by using haplotypes of several SNPs. Furthermore, because each new allele is associated with its own chromosomal history, haplotype-based analyses are warranted to detect unique chromosomal segments that harbor disease-predisposing alleles (7). We should examine differences in estimated haplotype frequencies among cases and controls. Before the differences can be compared, we will need to construct the haplotypes and estimate their frequencies because haplotypes are not directly observable. While this is still an active research area including the development of molecular haplotyping methods, there are many published approaches for determining haplotypes and haplotype blocks as reviewed by Niu (24). A general approach for choosing haplotype is to identify TagSNPs that are in strong LD.

For a small number of specified SNPs (4 to 10), a population-based approach can be used to find the haplotypes that maximize the likelihood function through the expectation-maximization (EM) algorithm under the assumption of HWE (6). How do we choose a particular set of SNPs for haplotype analysis? One approach is to focus on certain genomic regions or candidate genes and choose, say, 8 TagSNPs from each region or candidate gene. Another approach is to use moving windows. For example, Jawaheer and colleagues used a three-SNP moving window to detect the association between human leukocyte antigens and rheumatoid arthritis (13).

As described above, haplotypes can be constructed with these well-chosen TagSNPs or through moving windows of a few adjacent SNPs. Once the haplotypes and their frequencies are estimated using methods and software, such as PHASE (35), they can be treated as predictors in a logistic regression model (31, 44,). Other methods are also available (7) to compare the distribution profiles of haplotypes between cases and controls. Without considering covariates, the data become a 2 x *k* table, where *k* is the number of estimated haplotypes. Note that the TagSNPs and their haplotypes (and consequently the frequencies of the haplotypes) can only be determined after the data are collected, and hence it is premature to consider specific individual haplotypes. Broadly speaking, however, due to the haplotype uncertainties, the power calculation depends on many unknown quantities. Even though the number of haplotypes to be considered is relatively large, as discussed above, the use of haplotypes may still increase the power over the use of individual SNPs. A common sense approach to gauging the power of the haplotype analysis is by collapsing haplotypes into two major types (a wild type/a risk variant), and then the locus of interest would essentially become a di-allelic locus. In other words, it can be treated no differently from an individual SNP as presented above.

After the SNP-based and haplotype-based associations are performed, it is important to achieve a certain synergy in the two types of analysis. The common sense wisdom is that no association testing in epidemiological studies alone can distinguish between the true-positive and false-positive signals obtained in a multistage genome-wide scan. Approaches that have been suggested include comparative sequence analysis (2, 33), linkage analysis of expression data (21), or computational approaches to predicting function (17, 23, 39, 45), before launching into labor-intensive tests (36).

We should note some caveats in haplotype analyses. The possibilities and construction of haplotypes can lead to higher degrees of freedom in tests that tend to reduce power. Some study strategies such as the use of moving windows may lead to effectively the same number of tests as for single SNP analysis, but they may constrain the information to be utilized.

Testing gene-gene interactions. Statistical analysis of GWAS has an undeniable aspect of "exploratory" nature. This is especially true for investigating gene-gene interactions and

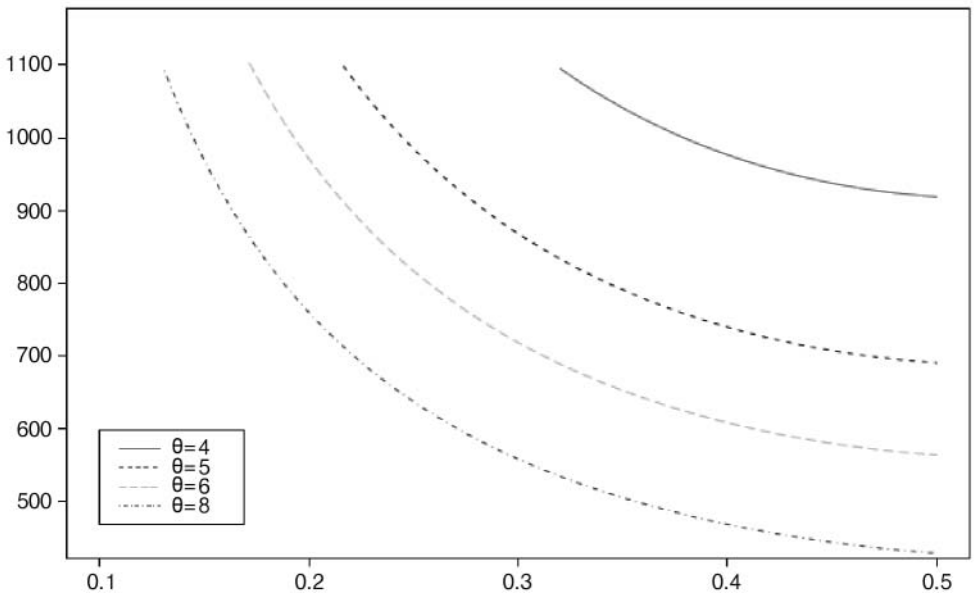

Figure 2. *The required sample size is plotted against the genotype frequency of the 1st SNP in order to detect a specified level of interaction between two SNP genotypes. ORs of 1st and 2nd SNP, as the main effect terms, are 1.2 and 1.5, respectively, and the 2nd SNP genotype frequency is set at 0.2. The type I error is controlled at 0.0001 and power at 80%.*

pathways. Unless we have solid prior evidence, any power calculation of gene-gene interactions in a GWAS is likely off target and not useful. However, it is important to take this exploratory step and generate sound hypotheses. After gene-gene interactions are examined, pathways can be identified according to the genes that interact with each other, and then perhaps expression arrays may further explore the functions of those genes that may underlie the cause for the disease.

Here, let us examine the power to assess the interaction of two SNPs with 1,000 cases and 1,000 controls. Figure 2 displays the required sample size as a function of the risk genotype frequency of the 1st SNP for testing the first hypothesis. A variety of the interaction effects,

$$\theta = \frac{OR(1,1)}{OR(1,0)OR(\mathrm{O},1)}$$

are considered, where the 2nd SNP genotype frequency is set at 0.2. Figure 3 is similar to Figure 2, except that a few possibilities of the genotype frequencies of the second SNP are considered and θ=8.

Based on Figure 2, a sample size of 500 cases and 500 controls would have sufficient power (>80%) to detect gene-gene interactions between two polymorphisms with moderate OR (1.2 and 1.5) and moderate interaction (θ=6).

It can be seen from figure 3 that a sample size of 500 cases and 500 controls is sufficient to detect a strong interaction effect (θ=8) between any two polymorphisms, where each polymorphism has a moderate OR of 1.5 and the frequency of the 2nd polymorphism ranges from 0.4-0.5.

In addition to Figure 2 and Figure 3, the work of Gauderman (10) assures that a study of 1,000 cases and 1,000 controls would have sufficient power to identify gene-gene interactions of a reasonable scale. According to their Table 1, a sample of 500 cases and controls has 80% power at a significance level of 0.05 to detect gene-gene interactions with OR's of about 2.2, for measuring departure from a purely multiplicative model when they considered two asthma genes, *GSTM1* and *GSTT1*, as examples.

Marchini *et al.* (19) examined the power of three strategies for analyzing gene-gene interactions in GWAS: strategy I, locus-by-locus search (requiring at least one locus meeting the significant criterion); strategy II, search over all pairs of loci; and strategy III, a two-stage strategy in which all loci meeting some low threshold in a single-locus search are subsequently examined for a significant full model fit. They considered 300K markers, 2,000 cases and 2,000 controls, and three multilocus disease models. They noted that there are many configurations in which the interaction-based search strategies are more powerful than searching locus-by-locus. It would be useful and interesting to examine those interactions revealed by all strategies. Even though some positive results may be missed, given the explorative nature of these analyses, a conservative approach is appropriate and any number of true positive findings would be a success.

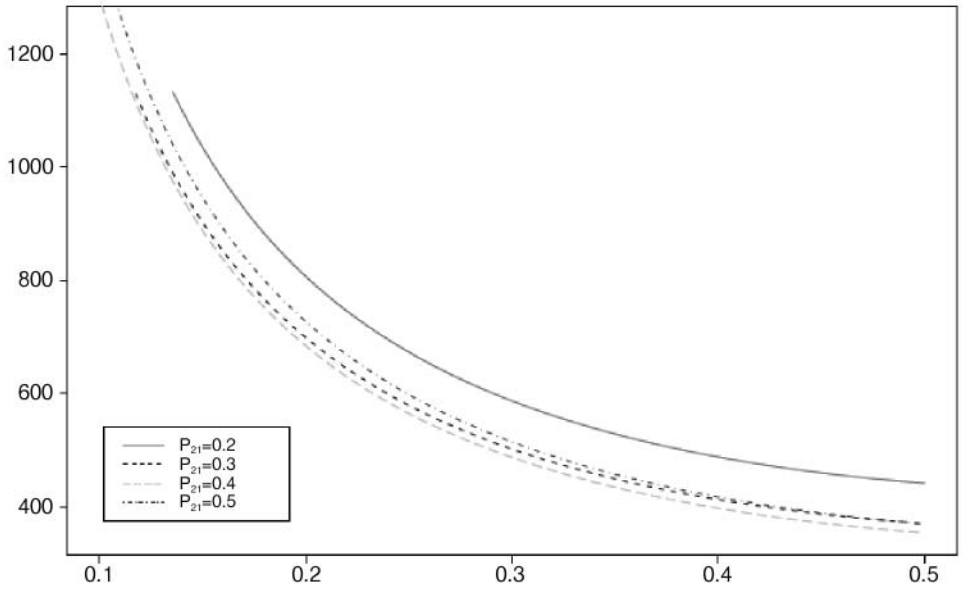

Figure 3. *The required sample size is plotted against the genotype frequency of the 1st SNP in order to detect an interaction effect, θ, of 8 between two SNP genotypes. ORs of 1st and 2nd SNP, as the main effect terms, are 1.5, and the 2nd SNP genotype frequency varies from 0.2 to 0.5. The type I error is controlled at 0.0001 and power at 80%.*

Testing gene-environment interactions. The number of environment covariates to be considered is considerably fewer than the number of genes. We can use a modified strategy based on strategy III described by Marchini *et al.* (19), described above. That is, we begin with all genes meeting some low threshold in a single-gene search, and then pair them with the environment factors to assess their interactions.

Population Substructures

It is well-known that population substructure, sometimes referred to as cryptic substructure, can provide spurious results for case-control association tests. For example, association studies of type II diabetes in Pima Indians (who have high rates of diabetes) were flawed because Pima individuals with a high degree of Caucasian ancestry had lower diabetes susceptibility. Thus, any marker loci that were at higher frequency in the Pima than in Caucasians were "associated" with the disease (10, 11, 15, 27). There have been many discussions on this issue. Some researchers believe that this phenomenon might have been over-stated (34). Further, Risch (28) pointed out that population stratification has been invoked numerous times as the cause for an observed high false-positive rate in association studies using candidate genes, yet it has rarely been demonstrated as the culprit (32). In a relatively recent extensive simulation study, Setakis concluded that explicit allowance for cryptic substructure may often be unnecessary provided that good study design principles have been used so that case and control populations are similar (30). When cases and controls are reasonably matched, two groups should be similar. However, those authors also pointed out that methods that protect against cryptic substructure typically perform well in limiting the number of false positives, and the cost of this protection, in terms of lost power, is often small. Thus, we suggest using appropriate methods to consider cryptic substructure. Several methods of assessing population stratification have been proposed using unlinked markers. A classic statistic for detecting cryptic substructure is Wright's FST (38), which is estimated as a single value that summarizes the average deviation of a collection of populations away from the mean. While there are a number of methods for adjusting associations for substructure and admixture, unlinked markers are used to adjust associations and the methods may be broadly divided into model based and non-model based approaches (30). Genomic control (26) is a non-model based approach that essentially correct asymptotic distribution of the classic Armitage trend test statistic by an over-dispersion factor, which is estimated from the empirical distribution of the trend statistic at a given number of null markers. The alternative methods such as those implemented in the program called STRUCTURE use model-based approaches to determine the underlying population structure (26). According to Setakis, none of these methods is uniformly superior to the others, nor is any one method uniformly inferior in presence of population structure (26). Nonetheless, if population substructures are evident in our data as suggested by

STRUCTURE, we will adjust the population substructure in our association test. The use of genomic control is simple but it can only be applied to SNPs. Generally, we need to use latent variables or mixture models that estimate the number of underlying subpopulations and the probability for an individual marker to be originated from each subpopulation, and then use mixed effects logistic models as described in Satten (30). The advantage of this approach is that it is applicable for both SNPs and haplotypes.

Exploratory Data Analysis

All GWAS are expected to collect rich and important genetic and clinical data. In addition to the hypothesis testing and regression models as described above, many contemporary approaches such as tree-based analysis can be applied to take advantage of all information and to simultaneously examine multiple SNPs, as well as haplotypes (40, 41).

Conclusion

This article is not intended to be an extensive review, but instead as an informative and pertinent guide to the statistical analysis of GWAS. A lot of methodological work has been done and continues to be done, and it is important for statistical geneticists to be familiar with the developments. However, the efforts to recruit families and collect genetic data should not be detoured by the limitations of the current state of the statistical methods. The availability of important data will stimulate exciting developments of statistical and computational methods for mining those data.

Acknowledgements

This research was supported in part by grants K02DA017713, R01DA016750, R01DA12844, T32MH014235, and U01HD050062 from the National Institutes of Health.

References

1 Abecasis GR and Cookson WO: GOLD-graphical overview of linkage disequilibrium. Bioinformatics *16(2)*: 182-183, 2000.
2 Bejerano G, Pheasant M *et al*: Ultraconserved elements in the human genome. Science *304(5675)*: 1321-1325, 2004.
3 Benjamini Y and Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to utiple testing. J Royal Statistical Society, Series B *57*: 289-300, 1995.
4 Daly MJ, Rioux JD *et al*: High-resolution haplotype structure in the human genome. Nat Genet *29(2)*: 229-32, 2001.
5 Despriet DD, Klaver CC *et al*: Complement factor H polymorphism, complement activators, and risk of age-related macular degeneration. Jama *296(3)*: 301-309, 2006.
6 Excoffier L and Slatkin M: Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol *12(5)*: 921-927, 1995.
7 Fallin D, Cohen A *et al*: Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res *11(1)*: 143-151, 2001.
8 Gabriel SB, Schaffner SF *et al*: The structure of haplotype blocks in the human genome. Science *296(5576)*: 2225-2229, 2002.
9 Garcia-Closas M and Lubin JH: Power and sample size calculations in case-control studies of gene-environment interactions: comments on different approaches. Am J Epidemiol *149(8)*: 689-692, 1999.
10 Gauderman WJ: Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol *155(5)*: 478-484, 2002.
11 Hanson RL and Knowler WC: Analytic strategies to detect linkage to a common disorder with genetically determined age of onset: diabetes mellitus in Pima Indians. Genet Epidemiol *15(3)*: 299-315, 1998.
12 Hoh J and Ott J: Mathematical multi-locus approaches to localizing complex human trait genes. Nat Rev Genet *4(9)*: 701-709, 2003.
13 Jawaheer D, Li W *et al*: Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. Am J Hum Genet *71(3)*: 585-594, 2002.
14 Klein RJ, Zeiss C *et al*: Complement factor H polymorphism in age-related macular degeneration. Science *308(5720)*: 385-389, 2005.
15 Lander ES and Schork NJ: Genetic dissection of complex traits. Science *265(5181)*: 2037-2048, 1994.
16 Li YJ, Hauser MA *et al*: Apolipoprotein E controls the risk and age at onset of Parkinson disease. Neurology *62(11)*: 2005-2009, 2004.
17 Livingston RJ, A von Niederhausern *et al*: Pattern of sequence variation across 213 environmental response genes. Genome Res *14(10A)*: 1821-1831, 2004.
18 Lubin JH and Gail MH: On power and sample size for studying features of the relative odds of disease. Am J Epidemiol *131(3)*: 552-566, 1990.
19 Marchini J, P Donnelly *et al*: Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet *37(4)*: 413-417, 2005.
20 Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H *et al*: Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods *1*: 109-111, 2004.
21 Morley M, Molony CM *et al*: Genetic analysis of genome-wide variation in human gene expression. Nature *430(7001)*: 743-747, 2004.
22 Morris AP: Direct analysis of unphased SNP genotype data in population-based association studies *via* Bayesian partition modelling of haplotypes. Genet Epidemiol *29(2)*: 91-107, 2005.
23 Ng PC and Henikoff S: SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res *31(13)*: 3812-3814, 2003.
24 Niu T: Algorithms for inferring haplotypes. Genet Epidemiol *27(4)*: 334-347, 2004.
25 Patil N, Berno AJ *et al*: Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science *294(5547)*: 1719-1723, 2001.

26 Pinkel D, Segraves R *et al*: High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet *20(2)*: 207-211, 1998.
27 Pritchard JK and Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet *65(1)*: 220-228, 1999.
28 Risch N: Searching for genes in complex diseases: lessons from systemic lupus erythematosus. J Clin Invest *105(11)*: 1503-1506, 2000.
29 Satagopan JM, Verbel DA *et al*: Two-stage designs for gene-disease association studies. Biometrics *58(1)*: 163-170, 2002.
30 Satten GA and Epstein MP: Comparison of prospective and retrospective methods for haplotype inference in case-control studies. Genet Epidemiol *27(3)*: 192-201, 2004.
31 Schaid DJ: Evaluating associations of haplotypes with traits. Genet Epidemiol *27(4)*: 348-364, 2004.
32 Setakis E, Stirnadel H *et al*: Logistic regression protects against population structure in genetic association studies. Genome Res *16(2)*: 290-296, 2005.
33 Sidow A: Sequence first. Ask questions later. Cell *111(1)*: 13-16, 2002.
34 Spence MA, Greenberg DA *et al*: The emperor's new methods. Am J Hum Genet *72(5)*: 1084-1087, 2003.
35 Stephens M, Smith NJ *et al*: A new statistical method for haplotype reconstruction from population data. Am J Hum Genet *68(4)*: 978-89, 2001.
36 Thomas DC, Haile RW *et al*: Recent developments in genomewide association scans: a workshop summary and review. Am J Hum Genet *77(3)*: 337-345, 2005.
37 Todd JA: Statistical false positive or true disease pathway? Nat Genet *38*: 731-733, 2006.
38 Wright S: The Genetic Structure of Populations. Ann Eugen *15*: 323-354, 1951.
39 Xi T, IM Jones *et al*: Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. Genomics *83(6)*: 970-979, 2004.
40 Zhang H and G Bonney: Use of classification trees for association studies. Genet Epidemiol *19(4)*: 323-332, 2000.
41 Zhang HP and B Singer: Recursive Partitioning in the Health Sciences. New York, Springer, 1999.
42 Zhang H, Zhong X and Ye Y: Multivariate linkage analysis using the electrophysiological phenotypes in the COGA alcoholism data. BMC Genet *6 Suppl 1*: S118, 2005.
43 Zhang K, Qin Z *et al*: HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. Bioinformatics *21(1)*: 131-134, 2005.
44 Zhao LP, Li SS *et al*: A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet *72(5)*: 1231-1250, 2003.
45 Zhu Y, Spitz MR *et al*: An evolutionary perspective on single-nucleotide polymorphism screening in molecular cancer epidemiology. Cancer Res *64(6)*: 2251-2257, 2004.

Received December 30, 2006
Accepted January 4, 2007