Published in final edited form as:

Genet Epidemiol. 2009 ; 33(Suppl 1): S29–S32. doi:10.1002/gepi.20469.

Improving the Signal-to-Noise Ratio in Genome-Wide Association Studies

Lisa J. Martin^{1,2}, Guimin Gao³, Guolian Kang³, Yixin Fang⁴, and Jessica G. Woo^{1,2}

¹Cincinnati Children's Hospital Medical Center, Cincinnati, OH

²University of Cincinnati School of Medicine, Cincinnati, OH

³The University of Alabama at Birmingham, Birmingham, AL

⁴Georgia State University, Atlanta, GA

Abstract

Genome-wide association studies employ hundreds of thousands of statistical tests to determine which regions of the genome may likely harbor disease-causing alleles. Such large-scale testing simultaneously requires stringent control over type I error and maintenance of sufficient power to detect true associations. These contradictory goals have led some researchers beyond Bonferroni correction of p -values to an exploration of methods to improve the detection of a few true effects in the presence of many unassociated loci. This paper reviews how Genetic Analysis Workshop 16 Group 5 investigators proposed to adjust for multiple tests while simultaneously using information about the structure of the genome to improve the detection of true positives.

Keywords

multiple testing; genetics; statistics

INTRODUCTION

The human genome may help explain underlying etiology of many diseases in humans. Technological advances have made it possible to screen hundreds of thousands or even millions of single-nucleotide polymorphisms (SNPs). Armed with this information, researchers hope to identify risk variants to help predict, treat, and prevent disease [Guttmacher and Collins, 2005].

While the ability to collect so much genetic information improves the likelihood of success, it also significantly increases the risk of spurious associations. To control for increased false-positive rate, the de facto requirement for genome-wide association studies (GWAS) has become Bonferroni correction. This method ensures the overall type I error rate is held to a specific threshold. Unfortunately, Bonferroni was developed for relatively small numbers of independent tests and has simply been scaled up for GWAS, resulting in an over-conservative correction that reduces power to detect true associations.

The Genetic Analysis Workshop (GAW16) Group 5 (controlling false positives) included five investigative teams who examined GWAS multiple testing strategies using GAW16 datasets [Amos et al., 2009; Kraja et al., 2009]. This paper reviews how Group 5

investigators proposed to adjust for multiple tests while simultaneously using information about the structure of the genome to improve the detection of true positives. Strategies included empirically estimating the true null distribution, identifying modest effects, and using prior information.

Methods to estimate the null distribution

One of the major concerns for researchers performing any statistical analysis is whether the underlying null distribution is truly reflected by a specified distribution (e.g., χ^2) [Ludbrook, 1994]. In GWAS, the appropriateness of a specified distribution is further complicated by the fact that many of the SNPs exhibit varying levels of correlation due to linkage disequilibrium. Permutation analysis generates an empirical null distribution based on the data at hand, thus is not subject to these problems. Two Group 5 investigators approached the problems of permutation efficiency and evaluating the appropriateness of the null distribution across allele frequencies.

Empirically deriving the null distribution using permutation in the context of GWAS is computationally intensive. Taking the correlation structure among test statistics for different markers into account may improve efficiency. For example, Lin [2005] proposed a Monte Carlo (MC) method based on approximating the joint distribution of test statistics under the null. Computing p -values adjusted for correlated tests (p_{ACT}) by approximate numerical integration of the asymptotic multivariate normal distribution of the test statistics has also been proposed [Conneely and Boehnke, 2007]. Both the MC and p_{ACT} methods work as step-down procedures, but they use different statistics for a single marker. Using the GAW16 rheumatoid arthritis (RA) GWAS data, Kang et al. [2009] evaluated these two multiple testing methods and the Bonferroni procedure by simulation. Results suggest that the MC method and p_{ACT} method can identify more significant SNPs than the Bonferroni correction methods. Also, results show that the MC method can have slightly higher power than the p_{ACT} method, but that the p_{ACT} method was significantly faster than the MC method. However, because both tests were limited by the number of markers that could be analyzed at one time, the genome was broken into blocks. Further, these blocks were assumed to be independent, an assumption which may not be valid. The results suggest that these approximate methods may be valuable for GWAS, but are hindered by limited block size.

Another concern with GWAS is that the p -values from single SNP-tests are often ranked to identify the most promising SNPs. GWAS SNP arrays include SNPs with a wide range of minor allele frequencies (MAFs). However, low MAF SNPs are often removed from analysis due to concerns about high false-positive rates for these SNPs. Previous work has demonstrated that power to detect associations is influenced by MAF [Ardlie et al. 2002]; thus, the underlying null distribution may also differ by MAF. Tabangin et al. [2009] examined the effects of MAF on the type I error rate based on permutations generated from simulated data. Five to ten true negative SNPs were selected from each of five MAF bins. For each SNP, one million permutations were performed. Using nominal type I error rates ($\alpha = 10^{-4}$, 10^{-5} , and 10^{-6}), they compared the observed and expected number of false-positive findings at each SNP. Using SNPs with higher MAFs (e.g., 25% or 50%) resulted in significantly fewer false positives than expected by chance. Further, SNPs with smaller MAFs exhibited more variability in false positives, but were not inflated relative to expected false positives. These results bring into question the need to remove SNPs with low MAF due to concerns about inflation of the false-positive rate, but suggest that estimation of the empirical null through permutation should be performed within MAF bins rather than genome-wide. Moreover, the finding that there were significantly fewer false positives than expected for common variants suggests that the multiple testing correction of 10^{-8} may be

too conservative even in the best case scenario of independent tests. Thus, identifying appropriate thresholds for genome-wide error rate may result in increased power.

Both of these papers demonstrate that accounting for the appropriate null distribution can improve power. However, given the differences in the approaches, future studies should examine how approximate methods that account for MAF could be utilized to efficiently adjust for multiple testing across the genome. Indeed, these papers together suggest that approximate methods may be appropriate for SNPs with MAF above 10%, but SNPs below this MAF level should be estimated separately as their underlying null distribution may differ.

Identification of modest effects

Often, individual variants identified using GWAS explain only a small proportion of the variability in traits [Wray et al., 2008]. This may be due to the low power to identify genetic variants with modest effects using traditional single-SNP tests. In Group 5, two investigators attempted to use multi-locus or global test statistics to identify modest effects.

A major challenge for GWAS investigators is determining whether there is statistical evidence that additional, modest, true associations exist once initial association testing has identified major loci. To address this issue, Parkhomenko et al. [2009] adapted the higher criticism (HC) test statistic [Donoho and Jin, 2004] to genetics and GWAS. HC is a global statistic based on a mixture model that assumes a proportion of the tests follows a non-null distribution. To evaluate the performance of this method, they used the RA dataset, eliminating previously identified regions of association [Plenge et al., 2007b]. The HC test statistic was applied to identify remaining modest associations. When using a genome-wide threshold $\alpha = 0.05$, the null hypothesis that none of the remaining SNPs are associated with the disease was rejected using the asymptotic threshold. Parkhomenko [2009] also estimated empirical thresholds corresponding to 90th, 95th, and 99th percentile of HC statistic based on 1,000 permuted replicates. However, only with a liberal 90th percentile did the results indicate the presence of modest effects. Thus, this is a novel method that may provide evidence for the presence of additional remaining SNPs modestly associated with the trait.

Another challenge for researchers is determining whether a cluster of SNPs exhibiting promising but not significant association should be pursued further. The challenge is that these clusters of SNPs may exhibit differential levels of linkage disequilibrium, thus providing both redundant and non-redundant information within a cluster. To address this question, Pankratz [2009] developed a novel method called the non-redundant summary (NRS) test statistic. This method not only considers the cumulative evidence of multiple SNPs in a region (by summing the negative $\log_{10} p$ -values) but also down-weights the contribution of additional markers in proportion to one minus the pairwise linkage disequilibrium of that SNP with markers already included in the summary statistic. Three significant regions were identified by the NRS statistic, which were also identified using a single-SNP test. **Pankratz** also applied the NRS statistic to a Parkinson disease dataset (personal communication), and identified a region using the NRS statistic that was not detectable using single SNP analysis. Thus, this appears to be an objective method of examining clusters of SNPs to determine whether there is an accumulation of statistical evidence of association within a region.

These two papers demonstrate that multilocus and global statistics are important tools that can be used in combination with single-SNP tests to detect genetic regions of interest, and can successfully identify the presence of additional SNPs of modest effect once major effects have been identified.

Using prior information

In the genomics era, much information exists that may provide prior evidence of the importance of specific genes or variants to a given disease or trait. For example, prioritization could be based on previous linkage or association analyses, gene expression data, or biologically relevant pathways [Loza et al., 2007]. Given these types of prior information, the number of tests performed could be dramatically reduced, thus improving the power to detect associations [Li et al., 2008]. Prior information can also aid investigators in identifying potentially interesting associations. In the last few years, a number of investigators have catalogued changes in gene expression profiles by genotypes across many SNPs [Duan et al., 2008; Pant et al., 2006; Stranger et al., 2007], which may help cull many false-positive associations.

In Group 5, Fang et al. [2009] also attempted to use prior information to identify the most promising associations when studying rare diseases. Rather than assuming that either allele would be equally likely to associate with disease (two-sided hypothesis), they assumed that rare diseases would be more likely to associate with the minor allele of a SNP (one-sided hypothesis). To address this issue, they used the North American Rheumatoid Arthritis Consortium dataset in which a variant of *TRAF1-C5* had been identified [Plenge et al., 2007a]. Fang and colleagues screened the last 10,000 SNPs on chromosome 9, which contain the *TRAF1-C5* locus. To test for association, they used the Armitage trend test [Armitage, 1955], which permits the evaluation of one or two-sided tests. Using the two-sided alternative, 13 SNPs, including 6 SNPs across the *TRAF1-C5* locus, were significantly associated. However, assuming the minor allele was positively associated with disease, only those 6 SNPs across the *TRAF1-C5* locus were significantly associated, consistent with the previous results by Plenge et al. [2007a, b], which were based on a combined analysis of two datasets. These results suggest that false-positive associations may be eliminated by comparing the results of one- and two-sided association tests under appropriate assumptions.

SUMMARY

Genome-wide association studies have successfully identified numerous loci at which common variants influence disease risk or variation in quantitative traits. Despite these successes, the variants identified by these studies have generally explained only a small fraction of the heritable component of disease risk, and have not been successful in identifying the causal variant(s) at the associated loci [McCarthy and Hirschhorn, 2008]. While identifying the strongest signals has been the focus of GWAS studies to date, it is now necessary to test and apply strategies to enable the detection of additional signals amidst the noise of GWAS data. If the only method for preventing false positives is to set higher thresholds for statistical significance, we will soon be unable to detect true associations. The GAW16 Group 5 authors explored the problems and tested alternative solutions for improving this signal-to-noise ratio in the context of GWAS. The diversity of approaches used by these investigators speaks to the complexity of the multiple testing problem.

The balance between controlling false positives and improving detection of true results led Group 5 investigators to re-examine properties of GWAS test statistics and evaluate methods to use multiple lines of evidence when examining GWAS results. This basic re-thinking of the issues led to methods that improved the efficiency of estimating the correct null distribution and raised the issue that the upper tail of the distribution may not follow standard statistical assumptions. Using biology to determine when one-sided or two-sided test statistics may be more appropriate or to weight the evidence in a region by the additional information content of multi-SNP analysis may enable a more precise exploration

of the genome, once the preliminary analysis is complete. Use of the HC threshold may identify when such in-depth analyses may be fruitful.

It is important to recognize that the set of p -values arising from a GWAS is not a random list generated from a defined statistical distribution. Rather, it results from known biological processes and structure within the genome. For instance, GWAS p -values are not independent of one another, to a degree that varies by linkage disequilibrium between loci. They are also not equivalent to each other because the underlying empirical underlying distribution may differ by minor allele frequency. The directionality of effects and biologic plausibility should be considered as well. Thus, these Group 5 papers have begun to explore opportunities to capitalize on these and other known relationships among loci to evaluate GWAS findings in a more sophisticated framework. GWAS projects remain extremely resource intensive, and should honor the study participants, precious DNA samples, and funding sources by fully utilizing the information obtained. While the vast majority of loci will never yield true associations, all possible lines of evidence should be brought to bear to help sift the true-positive from false-positive results, to improve the signal-to-noise ratio.

Acknowledgments

We thank all of the participants of Group 5 for the lively discussions. Participants included Thomas Dyer, Guolian Kang, Guimin Gao, Nathan Pankratz, Yixin Fang, Elena Parkhomenko, David Trichtler, Meredith Tabangin, and Qunyuan Zhang.

We thank the GAW organizers for providing this opportunity to explore this vital methodological issue. This research was supported in part by NIH grant R01 GM031575.

REFERENCES

- Amos CI, Chen WV, Seldin MF, Remmers E, Taylor KE, Criswell LA, Lee AT, Plenge RM, Kastner DL, Gregersen PK. Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. *BMC Proc.* 2009; 3 Suppl 7:S2. [PubMed: 20018009]
- Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. *Am J Hum Genet.* 2002; 71:304–311. [PubMed: 12096349]
- Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics.* 1955; 11:375–386.
- Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of p values for multiple correlated tests. *Am J Hum Genet.* 2007; 81:1158–1168. [PubMed: 17966093]
- Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. *Ann Stat.* 2004; 32:962–994.
- Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME. Genetic architecture of transcript-level variation in humans. *Am J Hum Genet.* 2008; 82:1101–1113. [PubMed: 18439551]
- Fang Y, Wang Y, Sha N. Armitage's trend test for genome-wide association analysis: One-sided or two-sided? *BMC Proc.* 2009; 3 Suppl 7:S37. [PubMed: 20018028]
- Guttmacher AE, Collins FS. Realizing the promise of genomics in biomedical research. *J Am Med Assoc.* 2005; 294:1399–1402.
- Kang G, Childers DK, Liu N, Zhang K, Gao G. Genome-wide association studies of rheumatoid arthritis data via multiple hypothesis testing methods for correlated tests. *BMC Proc.* 2009; 3 Suppl 7:S38. [PubMed: 20018029]
- Kraja AT, Culverhouse R, Daw EW, Wu J, Van Brunt A, Province MA, Borecki IB. The Genetic Analysis Workshop 16 Problem 3: Simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study. *BMC Proc.* 2009; 3 suppl 7:S4. [PubMed: 20018031]
- Li C, Li M, Lange EM, Watanabe RM. Prioritized subset analysis: Improving power in genome-wide association studies. *Hum Hered.* 2008; 65:129–141. [PubMed: 17934316]

- Lin DY. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*. 2005; 21:781–787. [PubMed: 15454414]
- Loza MJ, McCall CE, Li L, Isaacs WB, Xu J, Chang BL. Assembly of inflammation-related genes for pathway-focused genetic analysis. *PLoS ONE*. 2007; 2:e1035. [PubMed: 17940599]
- Ludbrook J. Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clin Exp Pharmacol Physiol*. 1994; 21:673–686. [PubMed: 7820947]
- McCarthy MI, Hirschhorn JN. Genome-wide association studies: Potential next steps on a genetic journey. *Hum Mol Genet*. 2008; 17:R156–R165. [PubMed: 18852205]
- Pankratz ND. Non-redundant summary scores applied to the North American Rheumatoid Arthritis Consortium dataset. *BMC Proc*. 2009; 3 Suppl 7:S39. [PubMed: 20018030]
- Pant PV, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. Analysis of allelic differential expression in human white blood cells. *Genome Res*. 2006; 16:331–339. [PubMed: 16467561]
- Parkhomenko E, Trichtler D, Lemire M, Hu P, Beyene J. Using a higher criticism statistic to detect modest effects in a genome-wide study of rheumatoid arthritis. *BMC Proc*. 2009; 3 Suppl 7:S40. [PubMed: 20018032]
- Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, Maller J, Pe'er I, Burt NP, Blumenstiel B, DeFelice M, Parkin M, Barry R, Winslow W, Healy C, Graham RR, Neale BM, Izmailova E, Roubenoff R, Parker AN, Glass R, Karlson EW, Maher N, Hafler DA, Lee DM, Seldin MF, Remmers EF, Lee AT, Padyukov L, Alfredsson L, Coby J, Weinblatt ME, Gabriel SB, Purcell S, Klareskog L, Gregersen PK, Shadick NA, Daly MJ, Altshuler D. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet*. 2007a; 39:1477–1482. [PubMed: 17982456]
- Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *N Engl J Med*. 2007b; 357:1199–1209. [PubMed: 17804836]
- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007; 315:848–853. [PubMed: 17289997]
- Tabangin ME, Woo JG, Martin LJ. The effect of minor allele frequency on likelihood of obtaining false positives. *BMC Proc*. 2009; 3 Suppl 7:S41. [PubMed: 20018033]
- Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk of complex disease. *Curr Opin Genet Dev*. 2008; 18:257–263. [PubMed: 18682292]