Published in final edited form as:
Genet Epidemiol. 2011 ; 35(Suppl 1): S56–S60. doi:10.1002/gepi.20650.

Inflated Type I Error Rates When Using Aggregation Methods to Analyze Rare Variants in the 1000 Genomes Project Exon Sequencing Data in Unrelated Individuals: Summary Results from Group 7 at Genetic Analysis Workshop 17

Nathan Tintle[1], **Hugues Aschard**[2], **Inchi Hu**[3], **Nora Nock**[4], **Haitian Wang**[3], and **Elizabeth Pugh**[5]

[1]Department of Mathematics, Statistics, and Computer Science, Dordt College, Sioux Center, IA

[2]Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, MA

[3]Department of Information Systems, Business Statistics, and Operations Management (ISOM), Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

[4]Department of Epidemiology and Biostatistics, Division of Genetic and Molecular Epidemiology, Case Western Reserve University, Cleveland, OH

[5]Center for Inherited Disease Research, School of Medicine, Johns Hopkins University, Baltimore, MD

Abstract

As part of Genetic Analysis Workshop 17 (GAW17), our group considered the application of novel and standard approaches to the analysis of genotype-phenotype association in next-generation sequencing data. Our group identified a major issue in the analysis of the GAW17 next-generation sequencing data: type I error and false-positive report probability rates higher than those expected based on empirical type I error levels (as high as 90%). Two main causes emerged: population stratification and long-range correlation (gametic phase disequilibrium) between rare variants. Population stratification was expected because of the diverse sample. Correlation between rare variants was attributable to both random causes (e.g., nearly 10,000 of 25,000 markers were private variants, and the sample size was small [n = 697]) and nonrandom causes (more correlation was observed than was expected by random chance). Principal components analysis was used to control for population structure and helped to minimize type I errors, but this was at the expense of identifying fewer causal variants. A novel multiple regression approach showed promise to handle correlation between markers. Further work is needed, first, to identify best practices for the control of type I errors in the analysis of sequencing data and then to explore and compare the many promising new aggregating approaches for identifying markers associated with disease phenotypes.

Keywords

population structure; correlated markers; next-generation sequencing

Corresponding Author: Nathan Tintle, Department of Mathematics, Statistics, and Computer Science, Dordt College, 498 4th Ave. NE, Sioux Center, IA 51250, Phone: 616-610-7726, ntintle@dordt.edu.

Introduction

The next-generation sequencing era is upon us. With the advent of this new era we usher in a host of questions about the statistical methods with which we will analyze sequencing data. As we explore state-of-the-art sequencing analysis and major questions in the field, it is helpful to first look back at lessons learned from the analysis of single-nucleotide polymorphism (SNP) microarray data.

The methods used to analyze common SNPs (e.g., those with minor allele frequency [MAF] > 5%) measured using SNP microarray technology have matured, with a generally accepted set of best practices for analysis of SNP microarray data (e.g., checking for Hardy-Weinberg equilibrium, quality control measures, consideration of population stratification, and consideration of linkage disequilibrium between SNP markers) (for a review of best practice protocols, see Laurie et al. [2010]). Despite these widely accepted best practices, however, some common problems have remained. The biggest unresolved problem is arguably that of statistical power. To date, most genome-wide association studies report predominantly small effect sizes (e.g., median odds ratio [OR] of all reported ORs in the National Human Genome Research Institute catalog is 1.3 [Hindorff et al., 2011]). Single-marker association methods can detect association only with the genotyped marker and variants in linkage disequilibrium with it, necessitating the genotyping and testing of hundreds of thousands to millions of SNPs to provide genome-wide coverage. Except for strong associations, it is difficult to have sufficient power to identify true associations as statistically significant because of the severe penalty imposed by multiple-testing correction procedures. For example, to have sufficient power (80%) to find a SNP that is significant at the genome-wide level (e.g., 1×10^{-7}) with low MAF (10%) and low population prevalence (10%), which increases the risk of disease by 30% for each copy of the risk allele (e.g., OR = 1.3; additive effect), one would need to have 4,700 case subjects and 4,700 control subjects if the actual causal variant was typed and potentially even more subjects if the risk variant was not typed but in linkage disequilibrium with the true causal variant [Purcell et al., 2003].

In an attempt to combat these power problems, some investigators have successfully used very large numbers of subjects (tens to hundreds of thousands) to find association [e.g., Lindgren et al., 2009; Speliotes et al., 2010]. However, for many diseases, cohorts of such size do not exist. Another approach is to aggregate true variant-phenotype associations across biologically meaningful sets (e.g., genes, sets of genes, or pathways) in order to both intensify the strength of association and substantially decrease the number of tests conducted. In many ways this field of methods is in its infancy, although the approaches have been considered by earlier Genetic Analysis Workshops [e.g., Tintle et al., 2009] and other investigators (see K. Wang et al. [2010] for a recent review).

Interestingly, this class of aggregation methods for SNP microarray data was developed to combat problems similar to those observed in the analysis of genome-wide next-generation sequencing data: specifically, conducting many tests where each individual test may be for a variant showing a relatively weak signal (a function of risk [effect], allele frequency, and the sample size of the study). To date, a number of methods have been proposed for next-generation sequencing data, all of which have a similar motivation: to aggregate signals across all, or some, of the SNPs within a gene with the idea of intensifying the observed signal while decreasing the number of tests conducted (see Dering et al. [2011] for a review).

Genetic Analysis Workshop 17 (GAW17) provided many participants with their first attempt at analyzing genotypes derived from next-generation sequencing data in the context of a simulated phenotype with known characteristics. In our group (Group 7), 11 of 12

participating work groups considered approaches to aggregating next-generation sequencing variant signals to increase power using a mix of methods, including methods originally developed for common variants, methods exclusively developed for next-generation sequencing data, and novel extensions of these methods. In addition, many work groups also considered the performance of single-SNP analyses on next-generation sequencing data, using both previously proposed and novel methods. In this review paper we summarize the main methods proposed and the individual findings of Group 7 participants and paint a broad picture of the current state of the field, including major lessons learned and open problems that need to be resolved.

Methods

Data

The data consist of 697 unrelated individuals genotyped at 24,487 autosomal SNPs, all located within one of 3,205 different genes. Genotypes were called from whole-exome reads obtained from the 1000 Genomes Project, including individuals of European, Asian, and African ethnicity. No quality information (e.g., coverage depth, quality score) was provided for the genotypes, although Almasy et al. [2011] note the use of imputation for missing genotypes.

The organizers of GAW17 simulated two quantitative phenotypes (Q1, Q2) and a latent liability trait for each individual. These traits were caused by 160 SNPs in 36 genes, most of which had low MAF (<0.01; 89 singletons) and many of which were in the vascular endothelial growth factor (VEGF) pathway. All SNPs increased the likelihood of trait values. The latent liability trait, Q1, Q2, and Q4 (caused by SNPs not included in the data set) all positively increased the likelihood of a disease phenotype (yes/no). Two hundred simulated phenotype replicates were included. A more detailed description of the data used for GAW17 is provided elsewhere [Almasy et al., 2011].

Group 7 Participants

Eleven separate work groups participated in the Group 7 discussion at the GAW17 meetings. Of these, three work groups did not submit their manuscripts for publication. The other eight work groups [Aschard et al., 2011; Hu et al., 2011; Nock and Zhang, 2011; Petersen et al., 2011; Scholz and Kirsten, 2011; Wang et al., 2011; Yang and Gu, 2011; Yang et al., 2011] plus one work group from Group 3 that joined our group after the workshop [Li et al., 2011] are published in the companion *BMC Proceedings* volume (v. 5, suppl. 9, 2011). Here, we summarize the methods and results from the nine published submissions using three broad categories based on the type of aggregation used. We note that, where appropriate, investigators used MAF cutoffs of 1% or 5% to classify SNPs as rare; for details, see the specific papers.

Gene-Level Aggregations

Five work groups considered various approaches to aggregating SNP variant information at the gene level [Aschard et al., 2011; Hu et al., 2011; Li et al., 2011; Scholz and Kirsten, 2011; Yang and Gu, 2011]. Aschard et al. [2011] compared rare variant signal to common variant signal using standard approaches and then also proposed a new test in which rare and common variant methods were combined using Fisher's combined probability test across genes. Li et al. [2011] used a multistep training, testing, and validation strategy involving a weighted collapsing approach [Dering et al., 2011] to identify genes associated with the disease phenotype and then compared their approach to logistic regression and a random forest. Scholz and Kirsten [2011] compared a variety of gene aggregation approaches, including the maximum statistic of all SNPs within a gene, Hotelling's test, multivariate

analysis, and the least absolute shrinkage and selection operator (LASSO) method, after collapsing all rare variants. Hu et al. [2011] applied their genetic risk score approach at both the gene level and the pathway level (described more fully in the next subsection). Last, Yang and Gu [2011] applied unweighted and weighted collapsing strategies [Dering et al., 2011] at the gene level, in addition to performing pathway analyses, as described in the next subsection.

Pathway-Level Aggregations

Five work groups considered approaches to aggregating SNP variant information at the pathway or gene set level. Two main types of approaches were considered: intermediate summarization (SNP-level data were first summarized at the gene level using a gene-level aggregation method and then aggregated to the pathway level) and direct summarization (SNP variant information was summarized directly to the pathway or gene set level). Two groups considered a direct summarization approach [Hu et al., 2011; Yang et al., 2011], two groups considered an intermediate summarization approach [Nock and Zhang, 2011; Yang and Gu, 2011], and one group considered both [Petersen et al., 2011]. Groups used a mix of true biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [Hu et al., 2011], chromosome-based sets [Yang et al., 2011], and synthetically created sets of genes containing known numbers of truly causal genes [Nock and Zhang, 2011; Petersen et al., 2011; Yang and Gu, 2011].

Yang and Gu [2011] and Petersen et al. [2011] used published intermediate summarization methods adapted for next-generation sequencing data. Namely, they first aggregated SNP scores to genes using the weighted and unweighted collapsing strategies (described earlier) and then applied gene set enrichment analysis (GSEA) [Petersen et al., 2011; Yang and Gu, 2011], variable set enrichment analysis (VSEA) [Yang and Gu, 2011], the Kolmogorov-Smirnov test [Petersen et al., 2011], and Fisher's combined probability test [Petersen et al., 2011] using simulated and/or real gene sets. Nock and Zhang [2011] first identified potentially interesting genes using regression and gene-level aggregation. Then they constructed latent variables to evaluate the aggregate effects of rare and common variants in potentially interesting genes. Finally, they used a structural equation model to model relationships between genes, covariates, and other constructs.

The three groups that used a direct summarization strategy each used different methods. Hu et al. [2011] counted the total number of rare variants possessed by each individual across all SNPs within a given pathway, creating an individual's genetic risk score for that pathway, which was then regressed onto each phenotype separately. Petersen et al. [2011] applied weighted and unweighted collapsing strategies (see Dering et al. [2011]) directly to sets of SNPs defined by pathways or gene sets instead of genes. Yang et al. [2011] used weighted and unweighted collapsing methods, along with a novel variation on the weighted strategy that tunes the weights empirically using the observed association with the phenotype.

Wang et al. [2011] did not consider aggregation but instead first applied a simple regression model for each SNP. Top-ranked SNPs were then subjected to sliced inverse regression (SIR), a dimension reduction technique.

False-Positive Report Probability and Type I Error Rate

As part of their analyses, most members of Group 7 computed either the false-positive report probability (percentage of noncausal genes [or SNPs or pathways] among all genes [causal and noncausal] that meet an arbitrary criterion for significance: the probability that the null hypothesis is true even though it has already been rejected) or the type I error rate

(percentage of all noncausal genes [or SNPs or pathways] that meet an arbitrary criterion for significance among all noncausal genes [significant and nonsignificant]: the probability that the null hypothesis is rejected even though it is true). We use the terms *type I error* and *power* rather liberally here because their use across the contributions from Group 7 varied, including estimates of these quantities across the 200 phenotype replicates, which all contain the same genotypes.

Results

Inflated False Positives

All five work groups using gene-level aggregation reported higher than expected type I error or false-positive report probabilities. Specifically, Hu et al. [2011] found a false-positive report probability of 50%, Scholz and Kirsten [2011] had a false-positive report probability that ranged between 94% and 98%, and Li et al. [2011] found 80%. Type I error rates were also inflated at the nominal 5% level (8–11% [Yang and Gu, 2011]; 9–20% [Aschard et al., 2011]). A variety of attempts were made to fix the problem, including the elimination of genes or SNPs that showed spurious associations [Yang and Gu, 2011; Scholz and Kirsten, 2011] (see Luedtke et al. [2011] for a discussion of spurious associations), use of principal components [Aschard et al., 2011], genomic control [Aschard et al., 2011], and pooling data across phenotype replicates [Li et al., 2011].

Similarly, all five groups that applied a pathway-level approach found inflated type I error rates. Specifically, the false-positive report probability tended to be quite high (65% [Yang et al., 2011]; 60.3% [Nock and Zhang, 2011]; and 86% and 33% [Hu et al., 2011]). Type I error rates were also inflated (5–9% [Yang and Gu, 2011]; up to 50% [Petersen et al., 2011]). Eliminating spurious genes showed substantial improvement in error rates [Petersen et al. 2011; Yang and Gu, 2011]. Using principal components did control the type I error rate but no significantly associations remained.

Wang et al. [2011], who did not use aggregation, looked first at the 30 most significant SNPs after application of a regression technique on the first 10 replicates and found that 80% of them were false positives. After the application of their dimension reduction approach using SIR, the average false-positive report probability dropped to 20% (4 out of 5 markers selected were causal).

Comparative Results of Aggregation Approaches

In the following paragraphs we provide a comparative analysis of the various aggregation approaches. All Group 7 contributions are considered, except for two groups that proposed novel approaches [Nock and Zhang, 2011; H. Wang et al., 2011] that were not directly compared to existing approaches.

For the gene-level aggregation results, Aschard et al. [2011] found that collapsing was outperformed by traditional multivariate approaches in gene-based association tests and showed weak power (9–23%) after controlling the type I error rate using genomic control so long as a gene included common variants (MAF > 1%). However, importantly, this power was essentially 0 when genes that included only rare variants were considered. Although Li et al. [2011] did not explicitly control the type I error rate, they used a receiver operating characteristic curve to compare methods, finding that a novel extension of the empirical Bayes risk prediction model provided the greatest area under the curve. Scholz and Kirsten [2011] compared methods and found that genes with multiple independent causal variants were better detected by multivariate methods (after collapsing rare variants), whereas genes with a single causal variant were better detected using the maximum association statistic

within a gene—findings that were true regardless of the methods applied to control the type I error rate.

For the pathway-level aggregation studies, Hu et al. [2011] and Yang and Gu [2011] explored multiple levels of aggregation, and both found that aggregating directly to the pathway level yielded more power than first aggregating at the gene level.

Among groups that compared different pathway-level aggregation methods, Yang and Gu [2011] found that the novel VSEA method outperformed the standard GSEA method. Petersen et al. [2011] found that, when summarizing significance at the gene level first, Fisher's combined probability test outperformed GSEA and the Kolmogorov-Smirnov approach, whereas direct application of the weighted-sum method on all SNPs from the pathway tended to yield the most power. Yang et al. [2011] had preliminary evidence that a Markov chain Monte Carlo (MCMC) method might outperform the genetic algorithms they considered, but they acknowledged that this could have been due to particular choices they made about how the genetic algorithm procedures were implemented.

Discussion

In general, Group 7 participants observed highly inflated type I errors (as high as 50%), high false-positive report probabilities (up to 90%), and low power to detect the simulated causal variants. Even after attempts to control type I errors (e.g., correcting for population stratification), power was generally quite low. One exception that resulted in high power was the use of collapsing methods (pathway or gene level) in which the collapsed variants included a high proportion of causal variants with higher MAFs (e.g., *FLT1* and *KDR*). Because only nonsynonymous SNPs predicted by SIFT [Ng and Henikoff, 2001] were included in the causal model, reducing the analysis to only nonsynonymous SNPs was thought to improve power, although the results showed only modest improvement [Scholz and Kirsten, 2011; Luedtke et al., 2011].

Two main reasons for type I errors were addressed during our group discussions and in our papers—population stratification and correlation between markers—which we briefly describe in the following subsections.

Population Stratification

When the first rare variant collapsing methods were proposed and published, little was made of the issue of population stratification. Although the treatment of population stratification and covariates is a reasonably straightforward issue in regression-based approaches on SNP microarray data (common variants), little has been published on best practices for handling population stratification in next-generation sequencing data. The level of population stratification in the GAW17 sample was large because subjects were taken from seven populations. Many members of our group considered population stratification in their analyses, most commonly through the use of principal components. In general, this approach reduced the type I error rate; however, many group members still saw increased type I errors even after accounting for population stratification.

Correlation Between Markers

In the analysis of SNP microarray data, understanding and leveraging marker correlation is a critical part of the design and analysis of most studies, with most of the effort involving linkage disequilibrium (correlation of markers generally located in close proximity on the genome). However, collapsing and aggregating methods should not be affected much by linkage disequilibrium when aggregating at the gene or pathway level unless there is linkage disequilibrium between genes or between pathways, something that is usually assumed to be

a relatively small problem [Li and Leal, 2008], although it is unclear whether rare variant methods correctly account for linkage disequilibrium blocks within genes or gene sets. In the data considered for GAW17, however, there appears to be correlation between markers, genes, and pathways that are located far apart on the genome. For example, one of the causal variants (private variant C4S1877) is identical to 27 other SNPs in different genes. In another case, *BUD13* was identified by both Li et al. [2011] and Luedtke et al. [2011] as a noncausal gene showing strong association with Q1, and it contained a SNP that was strongly correlated with a causal SNP in *KDR* ($r = -0.20$, $p = 9 \times 10^{-8}$). This correlation between markers that are located far apart in the genome is not linkage disequilibrium in the truest sense and was identified as either long-range correlation or gametic phase disequilibrium at the workshop. There are two possible explanations for the observed correlations: random causes and nonrandom causes.

Random correlation between markers is a phenomenon that is unique to rare variants in next-generation sequencing data. For example, the chance that two randomly selected markers will be perfectly correlated decreases as the allele frequency increases. As noted earlier, when a large number of private variants are in a sample (e.g., 9,433 variants in 697 individuals), most of them are perfectly correlated with many other private variants (the 9,433 markers were distributed across 685 individuals, yielding only 685 distinguishable markers). Random correlation between markers is typically not an issue in analysis of common variants.

Another explanation is nonrandom correlation. In follow-up analyses both Luedtke et al. [2011] and Aschard et al. [2011] demonstrated that correlations between SNPs were significantly beyond what was expected due to random chance alone (detailed results not shown). Further information on data production and cleaning would be needed to explain the cause of nonrandom correlation between genotypes.

Regardless of the cause, few methods have been proposed that can handle correlated markers for next-generation sequencing data. One such approach with promising results is that of Wang et al. [2011], who suggest using a multiple regression approach of individual correlated SNPs on the phenotype to identify SNPs with the strongest marginal effect (detailed results available from the authors).

Where We Stand Now

When type I errors are not well controlled, evaluation of power becomes meaningless; thus our group had little ability to report on the relative value of various methods in terms of statistical power. However, two themes are worth noting. First, as corroborated by a number of GAW17 groups using a variety of approaches, optimal analysis methods are quite dependent on the number, strength, and MAF of markers associated with disease. In this data set, a few markers with MAF > 1% had strong association with the phenotype, and so they were found by many groups, across replicates, using SNP-based, gene-based, or pathway-based approaches. In fact, SNP-based methods showed some value for these SNPs because the signal was so strong that it remained significant even after applying stringent corrections. However, work groups struggled to detect markers with lower MAF and weaker effects, especially for those genes containing few causal SNPs. Pathway methods lend a partial solution. However, high power (>50%) is observed only when sets of genes contain a large fraction of causal genes. It is unclear whether this will be the case in practice. Second, as is often the case with a major new technological breakthrough, a host of good and promising ideas have been offered about how to analyze next-generation sequencing data, and many novel, interesting, and still promising approaches were proposed by the members of our group. Unfortunately, we can make few conclusive statements about many of the approaches because of inflated type I errors.

Conclusions

As we saw with the advent of the analysis of SNP microarray data a decade ago, in this early stage of the next-generation sequencing era there are more questions and ideas than there are concrete answers about best practices for the analysis of the GAW17 data. However, Group 7 contributors have identified a major issue in the analysis of the GAW17 next-generation sequencing data, namely, that of type I errors and false-positive probabilities high above those expected based on empirical type I error levels. Certain themes emerged as best practices for the handling of this type I error problem, including the use of principal components analysis to control for population structure. In addition, several work groups found evidence of correlation between markers located far apart in the genome that can be attributed to both random and nonrandom causes. Although no conclusive statements can be made, control of population structure and evaluation of random and nonrandom correlation between markers may help to control inflation of type I errors. Further work is needed to first identify best practices for the control of type I errors in the analysis of next-generation sequencing data and then to explore and compare the many promising new approaches for identifying markers associated with disease phenotypes.

Acknowledgments

We thank the Group 7 participants for their contributions and Mark Zlojutro, Pingzhao Hu, and Markus Scholz for helpful comments on this manuscript. The Genetic Analysis Workshops are supported by National Institutes of Health (NIH) grant R01 GM031575 from the National Institute of General Medical Sciences. NT is supported by National Human Genome Research Institute (NHGRI) grant R15-HG004543, EP is supported by NHGRI grants HHSN268200782096C and R21-DK084529, HA is supported by Foundation Bettencourt-Schueller, and NLN is supported by National Cancer Institute grant K07CA129162.

References

Almasy L, Dyer TD, Peralta JM, Kent JW Jr, Charlesworth JC, Curran JE, Blangero J. Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc. 2011; 5(suppl 9):S2.

Aschard H, Qiu W, Pasaniuc B, Zaitlen N, Cho MH, Carey V. Combining effects from rare and common genetic variants in an exome-wide association study of sequence data. BMC Proc. 2011; 5(suppl 9):S44.

Dering C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol. 2011; X(suppl X):XX–XX.

Hindorff, LA.; Junkins, HA.; Hall, PN.; Mehta, JP.; Manolio, TA. [accessed March 25, 2011] A catalog of published genome-wide association studies. 2011. http://www.genome.gov/gwastudies

Hu P, Xu W, Cheng L, Xing X, Paterson AD. Pathway-based joint effects analysis of rare genetic variants using Genetic Analysis Workshop 17 exon sequence data. BMC Proc. 2011; 5(suppl 9):S45.

Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol. 2010; 34:591–602. [PubMed: 20718045]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–21. [PubMed: 18691683]

Li G, Ferguson J, Zheng W, Lee JS, Zhang X, Li L, Kang J, Yan X, Zhao H. Large-scale risk prediction applied to Genetic Analysis Workshop 17 mini-exome sequence data. BMC Proc. 2011; 5(suppl 9):S46.

Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, Qi L, Speliotes EK, Thorleifsson G, Willer CJ, Herrera BM, et al. Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. PLoS Genet. 2009; 5:e1000508. [PubMed: 19557161]

Luedtke A, Powers S, Petersen A, Sitarik A, Bekmetjev A, Tintle NL. Evaluating methods for the analysis of rare variants in sequence data. BMC Proc. 2011; 5(suppl 9):S119.

Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001; 11:863–74. [PubMed: 11337480]

Nock NL, Zhang LX. Evaluating aggregate effects of rare and common variants in the 1000 Genomes Project exon sequencing data using latent variable structural equation modeling. BMC Proc. 2011; 5(suppl 9):S47.

Petersen A, Sitarik A, Luedtke A, Powers S, Bekmetjev A, Tintle NL. Evaluating methods for combining rare variant data in pathway-based tests of genetic association. BMC Proc. 2011; 5(suppl 9):S48.

Purcell, S.; Cherny, SS.; Sham, PC. [accessed March 28, 2011] Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits—case-control calculator for discrete traits; Bioinformatics. 2003. p. 149-50.Calculator available at http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html

Scholz M, Kirsten H. Comparison of scoring methods for the detection of causal genes with or without rare variants. BMC Proc. 2011; 5(suppl 9):S49.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL, Lindgren CM, Luan J, Mägi R, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010; 42:937–48. [PubMed: 20935630]

Tintle NL, Lantieri F, Lebrec J, Sohns M, Ballard D, Bickeböller H. Inclusion of a priori information in genome-wide association analysis. Genet Epidemiol. 2009; 33(suppl 1):S74–S80. [PubMed: 19924705]

Wang H, Huang C-H, Lo S-H, Zheng T, Hu I. New insights into old methods for identifying causal rare variants. BMC Proc. 2011; 5(suppl 9):S50.

Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. Nat Rev Genet. 2010; 11:843–54. [PubMed: 21085203]

Yang F, Kang CJ, Marjoram P. Methods for detecting associations between phenotype and aggregations of rare variants. BMC Proc. 2011; 5(suppl 9):S51.

Yang W, Gu CC. Enrichment analysis of genetic association in genes and pathways by aggregating signals from both rare and common variants. BMC Proc. 2011; 5(suppl 9):S52.