2011

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 L. Tintle
Dordt College, nathan.tintle@dordt.edu

Hugues Aschard
Harvard School of Public Health

Inchi Hu
Hong Kong University of Science and Technology

Nora Nock
Case Western Reserve University

Haitian Wang
Hong Kong University of Science and Technology

Recommended Citation

Tintle, N., Aschard, H., Hu, I., Nock, N., Wang, H. and Pugh, E. (2011), 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. Genet. Epidemiol., 35: S56–S60. doi: 10.1002/gepi.20650

See next page for additional authors

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

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, Genetic Analysis Workshop 17, 1000 Genomes Project

Disciplines

Bioinformatics | Genetics and Genomics | Statistics and Probability

Comments

This is a pre-publication author manuscript of the following final, published article: Tintle, N., Aschard, H., Hu, I., Nock, N., Wang, H. and Pugh, E. (2011), 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. Genet. Epidemiol., 35: S56–S60. doi: 10.1002/gepi.20650

The definitive version is published by Wiley and available from Wiley Online Library (wileyonlinelibrary.com) at http://onlinelibrary.wiley.com/doi/10.1002/gepi.20650/abstract.

Authors

Nathan L. Tintle, Hugues Aschard, Inchi Hu, Nora Nock, Haitian Wang, and Elizabeth Pugh

Title: Inflated type I error rates when using aggregation methods to analyze rare variants in 1000 Genomes exon sequencing data in unrelated individuals: A summary report from Group 7 at Genetic Analysis Workshop 17

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

[$]Contact Author

1. Department of Mathematics, Hope College, Holland, MI, USA
2. Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA
3. Department of 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, 44106, USA
5. Center for Inherited Disease Research, School of Medicine, Johns Hopkins University, Baltimore (MD), USA

Running Title: Inflated type I error rates using aggregation methods

Contact author information:

Mailing address: Nathan Tintle, Department of Mathematics, Hope College, 27 Graves Place, Holland, MI 49423 USA

Phone: 616-395-7272

Email: tintle@hope.edu

Abstract:

As part of Genetic Analysis Workshop 17 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, namely that of type I error and false positive report probability rates high above those expected based on empirical type I error levels (as high as 50-90%). Two main causes emerged: population stratification and “long-range” correlation (gametic phase disequilibrium) between rare variants. Population stratification was expected due to a diverse sample. Correlation between rare variants was attributable to both random (e.g. nearly 10K of 25K markers were ‘private’ variants; small sample size of n=697) and non-random causes (more correlation was observed than was expected by random chance). Principal components analysis was used to control for population structure and helped 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 to first 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.

Key words:

population structure; correlated markers; next-generation sequencing

Background

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 SNP microarray data.

The methods used to analyze common (e.g. MAF>5%) SNPs 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 HWE, quality control measures, consideration of population stratification, consideration of LD between SNP markers, etc.; 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. Most GWAS studies, to date, report predominantly small effect sizes (e.g. median odds ratio (OR) of all reported ORs in the NHGRI GWAS catalog is 1.3; Hindorff et al. [2011]). Single marker association methods can only detect association 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 very strong associations, it is difficult to have sufficient power to identify true associations as statistically significant due to the severe penalty imposed by multiple-testing correction procedures. For example, to have sufficient (80%) power to find a SNP significant at the genome wide level (e.g. $1x10^{-7}$) with low minor allele frequency (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 4700 cases and 4700 controls if the actual ‘causal’ variant was typed, and potentially even more if the risk variant was not typed but in LD 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 (i.e. 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 attempts to aggregate true variant-phenotype associations across biologically meaningful sets (e.g., genes or sets of genes/pathways) in order to both intensify the strength of association while simultaneously substantially decreasing the number of tests conducted. In many ways this field of methods is in its infancy, though has been considered by earlier GAWs (e.g. Tintle et al. [2009]) and others (see Wang et al. [2010] for a recent review).

Interestingly, this class of aggregation methods for SNP microarray data was developed to combat similar problems 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: 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).

GAW17 provided many participants with their first attempt to analyze genotypes derived from next-generation sequencing data, in the context of a simulated phenotype with known characteristics. In our group, eleven of twelve participating groups considered approaches to aggregating next-generation sequencing variant signals to increase power using a mix of methods including (1) those originally developed for common variants (2) those exclusively developed for

NGS data and (3) novel extensions of these methods. Additionally, many groups also considered the performance of single SNP analyses on next-generation sequencing data, using both previously proposed and novel methods. In this summary paper we summarize the main methods proposed and individual findings of Group 7 participants, while also painting 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 consists of 697 unrelated individuals genotyped at 24,487 autosomal SNPs, all located within one of 3205 different genes. Genotypes were called from whole exome reads obtained from the 1000 genomes project, including individuals of European (CEU, TSI), Asian (CHB, JPT) and African (YRI, LWK) ethnicity. No quality information (e.g. coverage depth, quality score) was provided for the genotypes, though 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 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 dataset) all positively increased the likelihood of a disease phenotype (y/n). 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 manuscripts participated in the Group 7 discussion at the GAW 17 workshop. Of these, three manuscripts were not submitted for publication [Ling et al., 2011, Peralta et al., 2011 and Zlojutro et al., 2011, personal communications], while the other eight [Aschard et al., 2011, Hu et al., 2011, Nock et al., 2011, Petersen et al., 2011, Scholz and Kirsten 2011, Wang et al., 2011, Yang et al., 2011, Yang and Gu 2011], plus one paper from Group 3 that joined our group after the workshop [Li et al., 2011], are published in the companion BMC Proceedings volume. The methods and results from the nine published submissions are summarized using three broad categories based on the type of aggregation used. We note that, where appropriate, manuscripts used MAF cutoffs of 1% or 5% to classify SNPs as rare; for specific details see the referenced manuscripts.

Gene level aggregations

Five manuscripts 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 and Yang and Gu, 2011]. Here we briefly describe the methods used in these five manuscripts. Aschard et al. [2011] compared rare variant signal to common variant signal using standard approaches and then also proposed a new test where rare and common variant methods were combined using fisher's combined probability test across genes. Li et al. [2011] used a multi-step 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 Lasso after collapsing all rare variants. Hu et al., [2011] applied their genetic risk score approach at both the gene and pathway level (described more fully in the next section). Lastly, Yang and Gu [2011] applied unweighted and weighted collapsing strategies (Dering et al. [2011]) at the gene level, in addition to doing pathway analyses as described in the next section.

Pathway level aggregations

Five members of our group considered approaches to aggregating SNP variant information at the pathway or gene set level. There were two main types of approaches considered: *intermediate summarization* (SNP level data is 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 is summarized directly to the pathway/gene-set level). Two groups considered a direct summarization approach [Hu et al., 2011 and Yang et al., 2011], two groups considered an intermediate summarization approach [Yang and Gu, 2011 and Nock et al., 2011] and one group considered both (Petersen et al., 2011). Groups used a mix of true biological pathways from 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 et al., 2011, Petersen et al. 2011, Yang and Gu, 2011].

Intermediate summarization

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 GSEA [Yang and Gu 2011, Petersen et al.], VSEA [Yang and Gu], K-S [Petersen et al.] and Fisher’s combined probability test [Petersen et al.] using simulated and/or real gene sets. Nock et al. [2011] first identified potentially interesting genes using regression and gene-level aggregation. Then, latent variables were constructed to evaluate the aggregate effects of rare and common variants in potentially interesting genes. Finally, a structural equation model was used to model relationships between genes, covariates and other constructs.

Direct summarization

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, comprising an individual's *genetic risk score* for that pathway which was regressed onto each phenotype separately. Petersen et al. [2011] also used a direct summarization strategy, whereby they apply weighted and unweighted collapsing strategies (see Dering et al. [2011]) directly to sets of SNPs defined by pathways/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.

No aggregation

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; dimension reduction technique).

False positive report probability and type I error rate

Most members of our group computed either the false positive report probability (percent of non-causal genes (or SNPs or pathways) among all genes (causal and non causal) that meet an arbitrary criterion for significance; the probability the null is true given that one has rejected the null) or the type I error rate (percent of all non-causal genes (or SNPs or pathways) that meet an arbitrary criteria for significance among all non causal genes (significant and non-significant);

the probability that one rejects the null given that the null is true) as part of their analysis. We note that we use the terms “type I error” and “power” in this manuscript rather liberally as their utilization across papers in our group varied, including estimates of these quantities across the 200 phenotype replicates, which all contain the same genotypes.

Results

Inflated false positives for gene level aggregations

All five 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% (2/4), Scholz and Kirsten [2011] had a false positive report probability that ranged between 94-98% and Li et al. found 8/10. Type I error rates were also inflated (Yang and Gu 8-11%; Aschard 9-20%) at the nominal 5% level. A variety of attempts at fixing the problem took place including the elimination of genes/SNPs showing spurious (see Luedtke et al. [2011]) association (Yang and Gu 2011, Scholz and Kirsten), use of principal components (Aschard et al. [2011]), genomic control (Aschard et al. [2011]) and pooling data across phenotype replicates [Li et al., 2011].

Inflated false positives for pathway level aggregations

Similarly, all five groups applying a pathway level approach found inflated type I error rates. Specifically, false positive report probability tended to be quite high (13/20 Yang et al. [2010], 60.3% for Nock et al. [2011], and 62/72 and 1/3 for Hu et al. [2011]. Type I error rates were also inflated (5-9% Yang and Gu; Petersen et al. (up to 50%)). Eliminating spurious genes [Petersen et al. 2011 and Yang and Gu 2011] showed substantial improvement in error rates. Using principal components did control the type I error rate but no significantly associations remained.

Inflated false positives for non-aggregating analysis

Wang et al. [2011] 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 are causal).

Comparative results of aggregation approaches

The following sections provide a comparative analysis of the various aggregation approaches. All manuscripts are considered except two groups which proposed novel approaches [Nock et al., 2011 and Wang et al., 2011] that were not directly compared to existing approaches.

Gene level aggregation

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 as long as a gene included common variants (MAF>1%). However, importantly, this power was essentially zero when considering genes that included only rare variants. While Li et al. [2011] did not explicitly control the type I error rate, they used an ROC curve to compare methods, finding that a novel extension of the Empirical Bayes Risk Prediction model provided the highest AUC (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.

Comparing gene and pathway level aggregation

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.

Pathway level aggregation

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., found that, when summarizing significance at the gene level first, Fishers combined probability test outperformed GSEA and the K-S approach, while 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 MCMC may outperform the genetic algorithms (GA) they considered, but they acknowledge that this could be due to particular choices made about the implementation of the GA procedures.

Discussion

In general, Group 7 participants observed very 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 results of collapsing methods (pathway or gene level) where the collapsed variants include a very high proportion of causal variants with higher MAF (e.g. FLT1 and KDR). Since only non-synonymous SNPs predicted by SIFT [Ng and Henikoff, 2001] were included in the causal model, reducing the analysis to only non-synonymous SNPs was thought to improve power, though results showed only modest improvement [Scholz et al., 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 will briefly describe in the following sections.

Population stratification

When the first round of rare variant collapsing methods were proposed and published, little was made of the issue of population stratification. While 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 this sample was large since subjects were taken from 6 populations. Many members of our group considered population stratification in their analysis most commonly through the use of principal components. In general, this 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 impacted much by LD when aggregating at the gene or pathway level unless there is LD between genes or between pathways, something that is usually assumed to be a relatively small problem [Li and Leal, 2008], though it is unclear if rare variant methods correctly account for LD blocks within genes/sets. In the data considered for this workshop, 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 (identified by both Li et al. [2011] and Luedtke et al. [2011]) was a non-causal gene showing strong association with Q1 and contained a SNP that was strongly correlated with a causal SNP in KDR (r=-0.20, p=9x10^{-8}). This correlation between markers that are located far apart in the genome is not LD 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 and non-random 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 are perfectly correlated decreases as the allele frequency increases. As noted earlier, when a large number of private variants are in a sample (e.g. 9433 on 697 individuals), most private variants are perfectly correlated with many other private variants (the 9433 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 non-random correlation. In follow-up analyses both Luedtke et al. [2011] and Aschard et al. [2011] demonstrated that correlations between SNPs were significantly beyond what is 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 non-random correlation between genotypes. Regardless of the cause, few methods have been proposed that handle correlated markers for next-generation sequencing data. One such approach with promising results has been proposed by Wang et al. [2011], in which they 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 has 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 groups using a variety of approaches, optimal analysis methods are very dependent upon number, strength and MAF of markers associated with disease. In this dataset, a few markers with MAF>1% had strong association with the phenotype, and so were able to be 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 since the signal is so strong that it remains significant even after applying very stringent correction. However, 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 (>50%) power is only observed when sets of genes contain a large fraction of causal genes: it is unclear if this will be the case in practice. Second, as is often the case with a major new technological breakthrough, there are a host of good and promising ideas about how to analyze next-generation sequencing data, and there were many novel, interesting and still promising approaches proposed by the members of our group. Unfortunately, we can make few conclusive statements about many of the approaches due to inflated type I errors.

Conclusions

As we saw with the advent of the analysis of SNP microarray data a decade ago, as we enter the next-generation sequencing era there are more questions and ideas than there are concrete answers about best-practices for the analysis of these data. However, our groups 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. Additionally, multiple groups found evidence of correlation between markers located far apart in the genome attributable to both random and non-random causes. While no conclusive statements can be made, control of population structure and evaluation of random and non-random 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.

Acknowledgements

We thank the Group 7 participants for their contributions and to Mark Zlojutro, Pingzhao Hu and Markus Scholz for helpful comments on this manuscript. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. Authors of this manuscript were supported by the National Human Genome Research

Institute R15-HG004543 [N.T.], HHSN268200782096C. [E.P.], R21DK084529 and Foundation Bettencourt-Schueller [H.A.] and NCI K07CA129162 [N.L.N.].

References

Aschard H, Qiu W, Pasaniuc B, Zaitlen N, Cho MH, Carey V. 2011. Combining effects from rare and common genetic variants in exome-wide association study of sequence data. BMC Proc XX:XX

Almasy L, et al., 2011. Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc XX:XX.

Dering C, Pugh E, Ziegler A. 2011. Statistical analysis of rare sequence variants: An overview of collapsing methods. Genetic Epidemiol, GAW SUPPL.

Hindorff LA, Junkins HA, Hall PN, Mehta JP, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed March 25, 2011.

Hu P, Xu W, Cheng L, Xing X, Paterson AD. 2011. Pathway-based joint effect analysis of rare genetic variants using GAW17 exon sequence data. BMC Proc XX:XX

Laurie CC et al., 2010.Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic Epidemiol 34,591-602.

Li G, Ferguson J, Zheng W, Lee JS, Zhang X, Li L, Kang J, Yan X, Zhao H. 2011. Large-scale risk prediction applied to genetic analysis workshop 17 mini-exome sequence data. BMC Proc XX:XX

Li and Leal, 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83(3):311-321.

Ling H, Hetrick K, Schenll A, Pugh E. 2011. Personal Communication.

Lindgren CM, Heid IM, Randall JC, Lamina C, Steinthorsdottir V, et al.. 2009. Genome-Wide Association Scan Meta-Analysis Identifies Three Loci Influencing Adiposity and Fat Distribution. PLoS Genet 5(6): e1000508.

Luedtke A, Powers S, Petersen A, Sitarik A, Bekmetjev A, Tintle NL. 2011. Evaluating methods for the analysis of rare variants in sequence data. BMC Proc XX:XX.

Ng PC & Henikoff S. 2001 Predicting deleterious amino acid substitutions. Genome res 11(5):863-74.

Nock, NL and Zhang, LX. 2011. Evaluating aggregate effects of rare and common variants in the 1000 genomes exon sequencing GAW17 data using latent variable structural equation modeling. BMC Proc XX:XX

Peralta JM, Bernardini MCS, Vorst HR. 2011. Personal Communication.

Petersen A, Sitarik A, Luedtke A, Powers S, Bekmetjev A, Tintle NL. 2011. Evaluating methods for combining rare variant data in pathway-based tests of genetic association. BMC Proc XX:XX

Purcell S, Cherny SS, Sham PC. 2003. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Case-control calculator for discrete traits. Bioinformatics. 19(1):149-150. Calculator available at: http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html Accessed March 28, 2011.

Scholz M , Kirsten H. 2011. Comparison of scoring methods for the detection of causal genes with or without rare variants. BMC Proc XX:XX

Speliotes EK, Willer CJ, Berndt SI, et al., 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42(11):937-48.

Tintle, N.L., Lantieri F., Lebrec, J., Sohns, M., Ballard, D., Bickeböller, H. (2009) "Inclusion of *a priori* information in genome-wide association analysis" Genetic Epidemiol 33(S1):S74-S80.

Wang H, Huang C-H, Lo S-H, Zheng T, Hu I. 2011. New insights on old methods in identifying causal rare variants. BMC Proc XX:XX

Wang K, Li M and Hakonarson H 2010. “Analysing biological pathways in genome-wide association studies” Nat Rev Genet. 11(12):843-854.

Yang W and Gu CC. 2011. Enrichment analysis of genetic association in genes and pathways by aggregating signals from both rare and common variants. BMC Proc XX:XX

Yang F, Kang CJ, Marjoram P. 2011. Methods for detecting associations between phenotype and aggregations of rare variants. BMC Proc XX:XX

Zlojutro M, Kent JW, Dyer T, Blangero J and Almasy L. 2011. Personal Communication.