Published in final edited form as:
*Genet Epidemiol*. 2011 ; 35(Suppl 1): S61–S66. doi:10.1002/gepi.20651.

# Multiple Testing in High-Throughput Sequence Data: Experiences from Group 8 of Genetic Analysis Workshop 17

**Inke R. König**[1], **Jeremie Nsengimana**[2], **Charalampos Papachristou**[3], **Matthew A. Simonson**[4], **Kai Wang**[5], and **Jason A. Weisburd**[6]

[1]Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Lübeck, Germany

[2]Section of Epidemiology and Biostatistics, Leeds Institute of Molecular Medicine, Cancer Genetics Building, St James's University Hospital, Leeds, United Kingdom

[3]Department of Mathematics, Physics, and Statistics, University of the Sciences, Philadelphia, PA

[4]Institute for Behavioral Genetics, University of Colorado at Boulder, Boulder, CO

[5]Department of Biostatistics, University of Iowa, Iowa City, IA

[6]Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY

## Abstract

The use of high-throughput sequence data in genetic epidemiology allows the investigation of common and rare variants in the entire genome, thus increasing the amount of information and the potential number of statistical tests performed within one study. As a consequence, the problem of multiple testing may become even more pressing than in previous studies. As an important challenge, the exact number of statistical tests depends on the actual statistical method used. Furthermore, many statistical approaches for the analysis of sequence data require permutation. Thus it may be difficult to also use permutation to estimate correct type I error levels as in genome-wide association studies. In view of this, a separate group at Genetic Analysis Workshop 17 was formed with a focus on multiple testing. Here, we present the approaches used for the workshop. Apart from tackling the multiple testing problem, the new group focused on different issues. Some contributors developed and investigated modifications of existing collapsing methods. Others aimed at improving the identification of functional variants through a reduction and analysis of the underlying data dimensions. Two research groups investigated the overall accumulation of rare variation across the genome and its value in predicting phenotypes. Finally, other investigators left the path of traditional statistical analyses by reversing null and alternative hypotheses and by proposing a novel resampling method. We describe and discuss all these approaches.

### Keywords

next-generation sequencing; resampling; collapsing methods; rare sequence variants

Corresponding Author: Prof. Dr. Inke R. König, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Str. 1, 23562 Lübeck, Germany, Phone: +49 451 500 5580 Fax: +49 451 500 2999, inke.koenig@imbs.uni-luebeck.de.

## Introduction

With the availability of high-throughput sequence data, an enormous expansion in the magnitude of data volume has been reached. Previously, the spectrum of genetic epidemiological studies extended analysis beyond investigating single candidate genes toward the analysis of variants distributed across the entire genome with increasing coverage; these analyses formed the basis for genome-wide association (GWA) studies. At each step of increased analytical resolution, the magnitude of data in these analyses also greatly increased. The most recent advance involves sequence data that cover common and rare variants in candidate gene, whole-exome, or even whole-genome data. As a consequence, the problem of multiple testing presents an even more pressing problem than in the era of GWA studies, where it is already of significant concern.

In view of this challenging situation for the analysis of high-throughput sequence data, a separate contribution group (Group 8) for Genetic Analysis Workshop 17 (GAW17) was formed with a focus on multiple testing. In the next section we briefly discuss multiple testing in sequence data in general and how this problem was addressed by the Group 8 contributors.

Apart from tackling the problem of multiple testing, the contributions of Group 8 focused on different issues. Broadly, the issues fell into four categories, each described in more detail in the further sections of this paper. In the first category, modifications of existing collapsing methods were developed and investigated [Barrett and Nsengimana, 2011; Chung et al., 2011; Huang et al., 2011]. In the second category, contributors aimed at improving the identification of functional variants by reducing the number of dimensions [Kwon et al., 2011; Pardy et al., 2011]. Two further contributions investigated the overall accumulation of rare variation across the genome and its value in predicting phenotypes [Howrigan et al., 2011; Wu et al., 2011]. The final category was composed of contributions that left the path of traditional statistical analyses either by reversing null and alternative hypotheses [Papachristou, 2011] or by proposing a novel resampling method [Wang and Huang, 2011].

## What Is the Problem of Multiple Testing in High-Throughput Sequence Data?

Although multiple testing is certainly a well-known problem and although various approaches have been used to deal with it in GWA studies [Ziegler and König, 2010, ch. 14], three features make multiple testing especially challenging with the use of high-throughput sequence data. First, the overall number of investigated variants is greatly increased in any genomic region, as described earlier. As a consequence, the number of statistical tests increases from one or a few tests at one extreme to millions of tests at the other extreme. The exact number of tests performed depends in part on the statistical method used. Although the naive approach would be to test all common and rare variants individually, the rarity of these variants makes this approach statistically problematic [Dering et al., 2011]. As a result, many statistical methods collapse across rare variants within a region of interest, resulting in fewer overall tests. However, deciding on the criteria for collapsing variants and how many tests should be performed depends on a number of factors. These factors include the fact that regions of interest defined by different criteria, as described by Dering et al. [2011], can result in a different number of tests in a study. Moreover, which variants within a region are collapsed varies as a result of different thresholds that define rare variation and the possible inclusion of common variants as well.

The second challenge with multiple testing in sequence data is that many of the popular test statistics used for rare variants rely on permutation to yield *p*-values, and permutation can be

computationally time-consuming [Dering et al., 2011]. For instance, some approaches use the maximum test statistic over various collapsing methods [e.g., Price et al., 2010], using permutation to estimate the significance. Furthermore, permutation is often required because the tests use estimates from observed data, such as the direction of the effect [e.g., Han and Pan, 2010]. As a consequence, time-consuming permutations already have to be performed just to estimate *p*-values; implementing permutation approaches on top of that to estimate the correct type I error level may be suitable in GWA studies, but it cannot be applied as easily to sequence data.

A final challenge may be specific to the data simulated for GAW17, which contain functional variants with a range of effect sizes, with most having small effects and thus low power [Almasy et al., 2011]. Therefore overly stringent control of the type I error may not be desirable, because it comes at the cost of missing true positives.

In our GAW17 group, investigators used four principal techniques to deal with multiple testing, with a combination of techniques applied by most: (1) classical Bonferroni adjustment for multiple testing [Barrett and Nsengimana, 2011]; (2) a type or modification of resampling [Pardy et al., 2011; Wang and Huang, 2011]; (3) procedures that reduced the number of overall tests, for instance, through functional classification [Howrigan et al., 2011], by using a multistage design [Barrett and Nsengimana, 2011; Howrigan et al., 2011; Pardy et al., 2011; Wu et al., 2011], or by reducing dimensionality [Chung et al., 2011; Huang et al., 2011; Kwon et al., 2011; Pardy et al., 2011]; and (4) circumvention of the problem altogether by reversing the null and alternative hypotheses [Papachristou, 2011] or by using machine learning approaches [Huang et al., 2011; Pardy et al., 2011; Wu et al., 2011]. Further details of these methods are given in the following subsections.

## How Can Collapsing Methods Be Improved?

Barrett and Nsengimana [2011] introduced two modifications to the standard collapsing method. First, they implemented a two-stage design to increase cost-effectiveness and to reduce the number of statistical tests. For this, they split each GAW17 data replicate into equally sized training and test data sets. They used training and test data sets from the same replicate of unrelated individuals to avoid the problem that genotypes were invariant across replicates. In the simple approach, in the first stage a simple score was computed as the total number of minor alleles from all variants in a gene. Association between this score and each of Q1, Q2, and Q4 was then tested in the training data set using linear regression, regressing the phenotype on the score and adjusting for population in three categories: 156 Europeans (CEPH [Utah residents] or Tuscans), 321 Asians (Chinese or Japanese), and 220 Africans (Luhya or Yoruba). Genes with $p \leq 0.01$ were taken forward to the second stage, in which the same linear regression analysis was performed on the selected genes in the test data set and a standard Bonferroni correction was applied to the number of genes taken forward. As a second modification to the standard collapsing methods (alternative method), different gene scores were used in the two stages. Specifically, in stage 1 the same gene score as described in the simple approach was used and, in addition, each variant effect was tested separately to ascertain the direction, not the significance, of its effect on the phenotype. In stage 2, instead of summing all minor variants in the gene, only those variants with the same direction as the overall gene effect were summed.

Barrett and Nsengimana [2011] found that both strategies were comparable, with an overall low power but adequate type I error. However, for one gene, for which most of the effect was due to very rare variants, the simple strategy outperformed the alternative approach. The alternative method sought to improve power by removing variants with an opposing effect from the gene score; the data were not favorable to this, because all simulated variants had a

positive effect on the phenotypes and no gene contained variants with opposing effects [Almasy et al., 2011].

Chung et al. [2011] based their approach on aggregating the genotype dissimilarities between individuals across an entire DNA sequence region in a distance matrix to capture an entire region simultaneously [Schork et al., 2008]. For this, Euclidean distances *d* were calculated using numerically coded genotypes of the 13 risk genes for Q2 for all possible pairs of unrelated individuals:

$$d(\mathbf{a},\mathbf{b}) = \left\|\mathbf{a}-\mathbf{b}\right\| = [(\mathbf{a}-\mathbf{b})\cdot(\mathbf{a}-\mathbf{b})]^{1/2}, \quad (1)$$

where the Euclidean distance is defined as the L2 norm between two individual genotype vectors **a** and **b**. For each gene, an $n \times n$ genotypic distance matrix **D** and a $n \times 1$ phenotype matrix **X** were constructed and used to calculate a pseudo-*F* statistic under the regression model that includes the trait as the independent variable. Each of the 13 × 200 tests underwent 1,000 permutations in which the rows and columns of its raw genotype matrix were shuffled at random. The empirical *p*-value was determined as the frequency of observing more extreme pseudo-*F* statistics in permutations than in the actual gene case. These analyses were performed using either all variants within a gene or only rare variants with minor allele frequency less than 0.01. Similarly, 508 control genes for Q2 were selected and tested using all 200 replications.

Chung et al. [2011] compared this approach using either all variants or only rare variants with the collapsing method of Li and Leal [2008] and the Mantel test [Mantel, 1967]. The Mantel test measures the correlation between two distance matrices, for which here the phenotypic distance was correlated with the genotypic distance based on the Euclidean distance measure as before. Chung and colleagues found that the frequency of false positives in their approach and in the collapsing method was slightly inflated to a similar degree. The Mantel test was somewhat less inflated but identified fewer variants. Overall, there was no best-performing method across all genes. An advantage of Chung et al.'s approach is that, in principle, the unit of the test can be defined flexibly; for instance, it can be defined to be composed of a functional domain instead of a sequence of adjacent variants. However, the method is computationally intensive and thus inappropriate for large studies.

Huang et al. [2011] focused on the problem that, in many collapsing methods, the collapsing criterion is rather subjective. The core of their approach was that variants were collapsed randomly, as follows: Within one gene, an integer *S* was randomly drawn to indicate the number of variants to be included in the first subset. Then, *S* variants within the gene were drawn randomly to make up the first subset, and the remaining variants constituted the second subset. Within each of the two subsets, the variants were then collapsed according to the method of Li and Leal [2008]. This procedure was repeated many times, and in each repetition, important variants were selected by applying a variable selection algorithm on the resulting random subsets. Specifically, Huang and colleagues used the least absolute shrinkage and selection operator (LASSO) [Dasgupta et al., 2011; Tibshirani, 1996] as the selection algorithm with the phenotype Q1 and counted the frequencies with which single variants were selected across all replications of the random sampling. Without formal control of type I error, they assumed that variants that were more frequently selected would be more likely to be important. The result of their procedure showed that, within the top ten ranked genes, three of them actually contained functional variants.

## Does Dimensionality Reduction Help to Identify Functional Variants?

Some Group 8 contributors specifically aimed to alleviate the multiple testing problem in the data by reducing the number of statistical tests performed. For instance, Kwon et al. [2011] tested all of the variants simultaneously in unrelated individuals. Obviously, the number of observations was much smaller than the number of predictors in the data, so the classical approaches did not yield stable estimates. Based on their work for a previous Genetic Analysis Workshop for dichotomous endpoints [Kwon et al., 2009], the investigators applied singular value decomposition to reduce the dimension of the design matrix in the regression model. Within a Bayesian regression, the model was evaluated using a Markov chain Monte Carlo procedure with Gibbs sampler, and *p*-values were derived from permutation. The performance of the proposed method was evaluated in the first 10 replicates both with a single single-nucleotide polymorphism (SNP) association analysis and with a standard penalized regression [Dasgupta et al., 2011; Tibshirani, 1996]. Overall, the penalized regression and the novel approach faired similarly. Specifically, for the analysis of Q1, the positive predictive values were 19% for the penalized regression and 20% for the novel approach but only 4% for the single-SNP analysis. On the other hand, the negative predictive values were greater than 99% for all approaches. Compared with the penalized regression, the novel approach had the advantage of being more computationally efficient.

With the aim of identifying common variants associated with the phenotypes, Pardy et al. [2011] applied the LASSO procedure to select important predictors [Dasgupta et al., 2011; Tibshirani, 1996]. For this, adjustment by clinical parameters or prefiltering was necessary to yield an estimate of the shrinkage parameter. For prefiltering, random forests were used to discard unimportant variants. In more detail, subsamples were constructed in each replicate by randomly drawing 348 unrelated individuals without replacement. This was repeated 10 times, and the analyses were performed on each of the resulting subsamples. Then, random forests [Dasgupta et al., 2011; Schwarz et al., 2010] were used to identify possibly important variants based on the Gini index, which is a measure of impurity *i* for a single node and is defined as

$$i=1-\sum_{j} p(j)^2. \tag{2}$$

Here, $p(j)$ is the proportion of observations that are labeled with class $j$ in that node. The most important variants, about 100, were then forwarded to the LASSO analysis with a cross-validation for determining the penalty factor [Tibshirani, 1996] to further select variables. Finally, linear or logistic regression models were developed with backwards selection. The final model for Q1 included 18 variants, of which 3 were true positives; identification of variants for the other phenotypes was less successful.

## Does the Accumulation of Rare Variation Predict Phenotypes?

Instead of identifying functional variants, two groups of contributors investigated the merit of including rare variants to predict phenotypes. Thus each group adopted a broader approach to determining the role that rare variants play in the development of complex diseases. For example, Howrigan et al. [2011] tested the assumption that overall mutational load in an individual, rather than a few causal variants, would associate with affection status and that this association would be stronger for increasingly rare variants. They constructed a phenotype of affection status based on all replicates of unrelated individuals and tested for association with mutational load. To define mutational load, they used different thresholds

and weighting schemes, most of which are described by Price et al. [2010]. These procedures included (1) the simple count of minor alleles at all variants, (2) a count of only variants that meet a certain allele frequency threshold, (3) an inverse weighting score of minor alleles based on their frequency, (4) variable allele frequency thresholds determined by permutation tests, (5) a minor allele count of nonsynonymous and synonymous variants, and (6) functional weighting according to amino acid changes. For all these criteria, Howrigan and colleagues performed regression analyses of affection status on the defined minor allele count or score, sex, and ethnicity. Across the different thresholds and weighting schemes, there was a consistent main effect of overall load that was positively associated with affection status. This result was expected because most inclusion criteria included the simulated effects. This effect was not driven by a small number of variants with large effects, and removal of the most significant genes still found an effect, albeit reduced. The differences between the procedures were negligible. However, focusing specifically on rare alleles did not result in stronger associations, resulting in no clear support for a mutational load hypothesis. However, further conclusions are certainly restricted by the underlying simulated model.

Using a different approach, Wu et al. [2011] found that consideration of rare alleles did result in a more precise prediction of the affection status. Their procedure was composed of two stages. In the first stage, all variants were tested for association with affection status using Fisher's exact test for common variants and collapsing methods for rare variants in unrelated case and control subjects. Only variants with some evidence of association were then taken forward to the second stage. In the second stage, Wu and colleagues used support vector machines [Hastie et al., 2009] to build prediction models based on the common variants from stage 1 and the covariates Age, Sex, and Smoking status. Then, rare variants were added to the models to investigate their additive value. To generate training and test data, Wu et al. [2011] considered two strategies. In the first approach, the prediction model was developed on the first replicate and tested on the remaining replicates. This led to the result that adding rare variants to the models slightly improved the prediction. Increasing the number of variants in the model by adding less significant variants in the first stage did not lead to improved prediction. According to the second strategy, every replicate was split into a training data set and a test data set and the analyses were performed on the combination across replicates. In this approach, the improvement by adding rare variants was greater, and the prediction was better by using more variants from stage 1.

## How Can the Multiple Testing Problem Be Circumvented?

The remaining contributors to Group 8 used general alternative approaches to try to solve the multiple testing problem in unconventional ways. Papachristou [2011] circumvented the problem of multiple testing by reversing the null hypothesis and the alternative hypothesis. This approach was applied to the analysis of the quantitative phenotype Q2 with family-based data. Papachristou's aim was to construct a confidence set of genetic loci that contributed at least a predetermined percentage $h$ to the overall genetic variation of a quantitative phenotype. This confidence set inference (CSI) method was developed in the framework of generalized linear mixed models. Specifically, for every variant that was tested, the null hypothesis was that the variant was a quantitative trait locus contributing at least $h$ of the total genetic variance, and the alternative hypothesis was that this locus contributed less than $h$. Thus the set of variants for which the null hypothesis was not rejected at level $\alpha$ constitutes a $1 - \alpha$ confidence set of loci contributing at least $h$ percent of the total genetic variance. The likelihood was formulated in a linear mixed models context, where the phenotype was explained by the effects of the covariates plus the effect resulting from the specific variant plus a random polygenic effect plus a random residual effect. The likelihood was maximized to obtain parameter estimates. The resulting estimates were then

used to test the described hypotheses and to construct a confidence set by aggregating all variants for which the null hypothesis was not rejected. Because the sample size was rather small, it was artificially increased by combining data from two, three, or four consecutive replicates of family data. Different values of *h* were used to balance false discovery rates and true discovery rates. The results showed that power was problematic in the data. Therefore, with small sample sizes, *h* had to be set to high values in order to control the false discovery rate, but this resulted in a true discovery rate of nearly 0. With greater sample sizes, however, more reasonable values of *h* could be applied, yielding acceptable levels of the false discovery rate and a satisfactory true discovery rate.

One of the main advantages of Papachristou's approach is that in most of the analyzed replicates, the confidence set included only causative variants, thus demonstrating an ability to separate functional SNPs from nearby SNPs in potentially high linkage disequilibrium. This result occurred because, by design, the CSI method tested each SNP to see whether it was a potential quantitative trait locus (QTL) with a specific contribution to the phenotype, rather than whether it was associated with the phenotype. Nevertheless, among other things, the magnitude of the contribution of a SNP to the variability of a quantitative trait depended on the frequency of its minor allele. Thus Papachristou's method can easily separate the functional SNP from the nearby ones in high linkage disequilibrium so long as the minor allele frequencies of the causative and the noncausative SNPs are different. Also, the approach can be applied to families of arbitrary size and structure. Challenges of the approach are that the total genetic variance needs to be known in advance and that *h* needs to be predefined.

Finally, Wang and Huang [2011] proposed a novel resampling method for GWA studies that can also be applied to high-dimensional data. The method is built on the idea that often a genetic score is used as a test statistic that is standardized to be compared to a known distribution, and the distribution of this score can be derived from the score at all variants as a reference to avoid making distributional assumptions. Using the quantitative phenotype Q2 in replicate 2 of unrelated individuals together with the covariates Sex, Age, and Smoking, Wang and Huang defined a statistic *S* as the sample mean of ($y^*g^*$), where $y^*$ is the residuals of the phenotype *y* after removing covariate effects and $g^*$ is the residuals of the genotype *g* after removing covariates effects. To assess the genome-wide significance of *S*, Wang and Huang generated the null distribution for *S* according to the following procedure: (1) The residuals $y^*$ and $g^*$ were computed. (2) A variant was randomly selected from a set of variants under the null hypothesis. This set was determined by using a histogram of *p*-values, as described by Storey and Tibshirani [2003]. In this application, this determination led to all variants being used. (3) Either $y^*$ or $g^*$ was permuted across all subjects. (4) The statistic *S* for the specific variant was computed. Steps 2 to 4 were repeated *K* times; in the presented analysis, *K* was set to 10,000,000. Then, the *p*-value of *S* was given by the proportion of $S_k > S$. Application of this procedure led to the identification of 18 variants with resampling *p*-values smaller than $5 \times 10^{-4}$. However, none of them were causal. Still, the procedure has potential in that it is quite fast and applicable to collapsing methods for identifying rare variants and other statistics.

## Discussion

The common theme of the contributions in GAW17 Group 8 was multiple testing, although it was not the original intent of any of the contributors. We have identified three major challenges in dealing with multiple testing in high-throughput sequence data: (1) the number of tests performed, (2) the computational demand of permutations, and (3) the low power. The first challenge can naturally be met be reducing the number of tests in the experiment. Indeed, this is the path that was followed most frequently in this and other GAW17

contribution groups. Most prominently, all collapsing methods reduce the number of tests [e.g., Barrett and Nsengimana, 2011; Howrigan et al., 2011; Huang et al., 2011]. In addition, Group 8 investigators used different multistage designs or techniques to reduce the dimensionality, which also led to a decrease in the number of tests [e.g., Barrett and Nsengimana, 2011; Huang et al., 2011]. The second challenge of computational load resulting from permutations can be solved in principle by using more computing power. More elegantly, different statistics can be used that do not rely on permutation, which was done implicitly in a number of contributions. Finally, different, more efficient permutation schemes can be used. The third challenge of low power was an overarching topic at GAW17. In addition to simulated small effects, investigators identified overall problems with the type I error frequencies, as described by Tintle et al. [2011]. As a result, many contributors who tried to control the type I error reduced the power even further.

At this point, we can conclude that, although the multiple testing problem persists, it might not be qualitatively different from the issue surrounding GWA studies. The proposed solutions mostly aim at reducing the number of tests, and some of the procedures that include prefiltering steps are also applicable to other kinds of data. What is still required is more computing power to allow for extensive permutations and, more important, larger sample sizes for a reasonable power.

Because of the nature of the simulated data, the proposed modifications of collapsing methods in Group 8 suffered from low power and thus yielded promising but not extremely successful results. In comparison, the different contributions may be interesting for different types of data and applications: A strict control of the type I error frequency can be obtained by using the approach of Barrett and Nsengimana [2011]. As a flexible but computationally intensive procedure, the approach of Chung et al. [2011] may be especially interesting for smaller studies. Finally, preselection of possibly interesting variants in large studies but without formal error control might be obtained by using the approach of Huang et al. [2011].

In an attempt to reduce the number of tests performed and thereby curtail the need for multiplicity adjustment, Kwon et al. [2011] and Pardy et al. [2011] considered dimension reduction techniques, such as singular value decomposition or a modified version of the LASSO. The results of the analyses from both methods suggest that, although dimensionality reduction can certainly be useful and sensible when it comes to multiple testing, it may not necessarily lead to an increase in power. Nevertheless, because many parameters can be varied in these procedures, further in-depth investigations are needed to fully understand the best possible application.

Whether rare variants add information concerning the development of affection is difficult to answer from the use of one simulated data set. Unless the contribution of these rare variants to the phenotype is substantial, including them in the analyses may not be beneficial, especially if the inclusion of such variants comes at the cost of higher computational load. Further analyses in real data sets are required, because the conclusion is limited by the simulation for the workshop.

Papachristou [2011] proposed a novel approach for targeting functional variants with a specific contribution to quantitative phenotypes. The results of the analyses showed, as expected, that with small sample sizes the method did not have enough power to identify trait loci and maintain a low false discovery rate. Nevertheless, with larger sample sizes, the method yielded reasonable power and false discovery rate, and it also displayed an increased ability to distinguish functional SNPs from the nearby loci with which the QTLs were in high linkage disequilibrium.

Determining the null distribution of each marker tested in a GWA study can be cumbersome, especially in the context of permutation tests. Wang and Huang [2011] described a novel resampling method that is applicable to high-dimensionality data, and it can be used to compute a reference distribution for a genome scan of any arbitrary SNP under the assumption of no association with the phenotype. Even though application of the method to the simulated data resulted in a significant false discovery frequency, the method has some potential as a result of its computational efficiency and its ability to easily adapt to any test statistic.

In general, although the focus and applied methods were quite different across the contributions, power was a general issue, because many of the simulated loci remained undetected. Similar problems were observed in other GAW17 groups. Barrett and Nsengimana [2011] found that the strongest determinants of statistical power to detect a gene effect were the sum of minor allele frequencies in that gene, the average effect of its variants, and the total number of these variants if the total variant count was used as a gene score. It was not possible to compare the results from different groups because the approaches used varied in their applicability, with different methods being suitable to different types of data, different study designs and sizes, and different underlying functional effects. Thus it is likely that no single method will eventually emerge as the optimum solution. Further in-depth investigations of data simulated under different scenarios and of real data sets will be required to fully address the issues brought up at this meeting. It is quite clear, though, that identifying rare variants such as those with frequency less than 0.001 require either large data sets or samples that are enriched in those variants. New developments in the design, analysis, and interpretation of these studies are therefore needed.

## Acknowledgments

The Genetic Analysis Workshops are supported by National Institutes of Health grant R01 GM031575. MAS was supported by an institutional training grant from the National Institute of Mental Health (T32 MH016880, John K. Hewitt, Director). JN has financial support from Cancer Research UK. We thank the additional Group 8 contributing members, including Jennifer H. Barrett, Doyoung Chung, Daniel P. Howrigan, Xin Huang, and Christopher Pardy.

## References

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

Barrett JH, Nsengimana J. Two-stage analyses of sequence variants in association with quantitative traits. BMC Proc. 2011; 5(suppl 9):S53.

Chung D, Zhang Q, Kraja AT, Borecki IB, Province MA. Distance-based phenotypic association analysis of DNA sequence data. BMC Proc. 2011; 5(suppl 9):S54.

Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genet Epidemiol. 2011; 35(suppl 1):S5–11. [PubMed: 22128059]

Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol. 2011; 35(suppl 1):S12–7. [PubMed: 22128052]

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70:42–54. [PubMed: 20413981]

Hastie, T.; Tibshirani, R.; Friedman, J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.

Howrigan DP, Simonson MA, Kamens HM, Stephens SH, Wills AG, Ehringer MA, Keller MC, McQueen MB. Mutational load analysis of unrelated individuals. BMC Proc. 2011; 5(suppl 9):S55.

Huang X, Fang Y, Wang J. Identification of functional rare variants in genome-wide association studies using stability selection based on random collapsing. BMC Proc. 2011; 5(suppl 9):S56.

Kwon S, Cui J, Rhodes SL, Tsiang D, Rotter JI, Guo X. Application of Bayesian classification with singular value decomposition method in genome-wide association studies. BMC Proc. 2009; 3(suppl 7):S9. [PubMed: 20018086]

Kwon S, Yan X, Cui J, Yao J, Yang K, Tsiang D, Li X, Rotter JI, Guo X. Application of Bayesian regression with singular value decomposition method in association studies for sequence data. BMC Proc. 2011; 5(suppl 9):S57.

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]

Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967; 27(2):209–20. [PubMed: 6018555]

Papachristou C. Confidence set of putative quantitative trait loci in whole genome scans with application to the Genetic Analysis Workshop 17 simulated data. BMC Proc. 2011; 5(suppl 9):S58.

Pardy C, Motyer A, Wilson S. Resampling procedures to identify important SNPs using a consensus approach. BMC Proc. 2011; 5(suppl 9):S59.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86:832–8. [PubMed: 20471002]

Schork N, Wessel J, Malo N. DNA sequence-based phenotypic association analysis. Adv Genet. 2008; 60:195–217. [PubMed: 18358322]

Schwarz DF, König IR, Ziegler A. On safari to Random Jungle: a fast implementation of random forests for high dimensional data. Bioinformatics. 2010; 26:1752–8. [PubMed: 20505004]

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003; 100:9440–5. [PubMed: 12883005]

Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc B. 1996; 58:267–88.

Tintle N, Aschard H, Hu I, Nock N, Wang H, Pugh E. Inflated type I error rates when using aggregation methods to analyze rare variants in 1000 Genomes Project exon sequencing data in unrelated individuals: a summary report from Group 7 at Genetic Analysis Workshop 17. Genet Epidemiol. 2011; 35(suppl 1):S56–60. [PubMed: 22128060]

Wang K, Huang J. Treating phenotype as given: a simple resampling method for genome-wide association studies. BMC Proc. 2011; 5(suppl 9):S60.

Wu C, Walsh KM, DeWan AT, Hoh J, Wang Z. Disease risk prediction with rare and common variants. BMC Proc. 2011; 5(suppl 9):S61.

Ziegler, A.; König, IR. A statistical approach to genetic epidemiology: concepts and applications. Weinheim, Germany: Wiley-VCH; 2010.