2011

Identification of Genetic Association of Multiple Rare Variants Using Collapsing Methods

Yan V. Sun
Emory University

Yun Ju Sung
Washington University School of Medicine in St. Louis

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

Andreas Ziegler
University of Lubeck

Recommended Citation

Sun, Y. V., Sung, Y. J., Tintle, N. and Ziegler, A. (2011), Identification of genetic association of multiple rare variants using collapsing methods. Genet. Epidemiol., 35: S101–S106. doi: 10.1002/gepi.20658

Identification of Genetic Association of Multiple Rare Variants Using Collapsing Methods

Abstract

Next-generation sequencing technology allows investigation of both common and rare variants in humans. Exomes are sequenced on the population level or in families to further study the genetics of human diseases. Genetic Analysis Workshop 17 (GAW17) provided exomic data from the 1000 Genomes Project and simulated phenotypes. These data enabled evaluations of existing and newly developed statistical methods for rare variant sequence analysis for which standard statistical methods fail because of the rareness of the alleles. Various alternative approaches have been proposed that overcome the rareness problem by combining multiple rare variants within a gene. These approaches are termed collapsing methods, and our GAW17 group focused on studying the performance of existing and novel collapsing methods using rare variants. All tested methods performed similarly, as measured by type I error and power. Inflated type I error fractions were consistently observed and might be caused by gametic phase disequilibrium between causal and noncausal rare variants in this relatively small sample as well as by population stratification. Incorporating prior knowledge, such as appropriate covariates and information on functionality of SNPs, increased the power of detecting associated genes. Overall, collapsing rare variants can increase the power of identifying disease-associated genes. However, studying genetic associations of rare variants remains a challenging task that requires further development and improvement in data collection, management, analysis, and computation.

Keywords

1000 Genomes Project, association, collapsing methods, next-generation sequencing

Disciplines

Bioinformatics | Genetics and Genomics | Statistics and Probability

Comments

This is a pre-publication author manuscript of the following final, published article: Sun, Y. V., Sung, Y. J., Tintle, N. and Ziegler, A. (2011), Identification of genetic association of multiple rare variants using collapsing methods. Genet. Epidemiol., 35: S101–S106. doi: 10.1002/gepi.20658

The definitive version is published by Wiley and available from Wiley Online Library (wileyonlinelibrary.com) at http://onlinelibrary.wiley.com/doi/10.1002/gepi.20658/abstract

Identification of Genetic Association of Multiple Rare Variants Using Collapsing Methods

Yan V. Sun[1], Yun Ju Sung[2], Nathan Tintle[3], and Andreas Ziegler[4]

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA

[2]Division of Biostatistics, Washington University School of Medicine, St. Louis, MO

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

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

Running Title: Collapsing Methods for Rare Variants

Corresponding author:

Yan V. Sun

Department of Epidemiology

Rollins School of Public Health

Emory University

1518 Clifton Road NE #3049

Atlanta, GA 30322

Phone: (404) 727-9090

Fax: (404) 727-8737

Email address: yvsun@emory.edu

Abstract

Next-generation sequencing technology allows investigation of both common and rare variants in humans. Exomes are sequenced on the population level or in families to further study the genetics of human diseases. Genetic Analysis Workshop 17 (GAW17) provided exomic data from the 1000 Genomes Project and simulated phenotypes. These data enabled evaluations of existing and newly developed statistical methods for rare variant sequence analysis for which standard statistical methods fail because of the rareness of the alleles. Various alternative approaches have been proposed that overcome the rareness problem by combining multiple rare variants within a gene. These approaches are termed collapsing methods, and our GAW17 group focused on studying the performance of existing and novel collapsing methods using rare variants. All tested methods performed similarly, as measured by type I error and power. Inflated type I error fractions were consistently observed and might be caused by gametic phase disequilibrium between causal and noncausal rare variants in this relatively small sample as well as by population stratification. Incorporating prior knowledge, such as appropriate covariates and information on functionality of SNPs, increased the power of detecting associated genes. Overall, collapsing rare variants can increase the power of identifying disease-associated genes. However, studying genetic associations of rare variants remains a challenging task that requires further development and improvement in data collection, management, analysis, and computation.

Key words: 1000 Genomes Project, association, collapsing methods, next-generation sequencing

Introduction

Genome-wide association studies have successfully identified hundreds of novel genetic loci associated with complex disease traits. In most cases, only a small portion of the heritability is explained by these associated common variants [Manolio et al., 2009; Eichler et al., 2010]. Although the summation of these associated loci may underestimate the total amount of heritability that common variants can explain [Yang et al., 2010], rare genetic variants might also contribute a sizable proportion of the genetic susceptibility to common diseases. In contrast to common variants associated with small effects, rare variants with putative functional change, such as nonsynonymous single-nucleotide polymorphisms (SNPs), are more likely to have a larger effect [Kryukov et al., 2007; Gorlov et al., 2008].

Current technology allows an exhaustive search for rare variants by sequencing the whole genome of a human being [Lee et al., 2010; Lupski et al., 2010; Sobreira et al., 2010]. Although the cost has drastically decreased in recent years, conducting a whole-genome sequencing project may not be cost-efficient for identifying functional rare variants associated with common diseases. The 30 million base pairs of the human exome account for about 1% of the whole human genome. With a fixed budget, an exome sequencing project can survey a much larger sample (required for detecting rare variants) with better coverage of the sequence reads (higher quality of the calls of rare variants). Several studies have demonstrated the utility of exome sequencing in identifying functional variants related to human diseases [Bilguvar et al., 2010; Gilissen et al., 2010; Ng et al., 2010a, 2010b; Walsh et al., 2010].

Because of the low allele frequencies, traditional regression-based methods do not work well with rare variants derived from sequencing data. To address the analytical challenge of identifying rare genetic variants associated with diseases, investigators have developed a number

of collapsing methods to summarize the individual rare variants in association analyses. These methods have been described in detail by Dering et al. [2011a]. Genetic Analysis Workshop 17 (GAW17) provided exome sequencing data from the 1000 Genomes Project and simulated phenotypic traits, both binary and quantitative [Almasy et al., 2011]. These simulated data sets with a large number of rare variants can be used to evaluate the existing and newly proposed methods to identify the associations of rare variants. For exome sequencing data, a natural unit for collapsing genetic variants is the gene. Although many of these collapsing methods can be used for common variants, GAW17 Group 15 focused on the methods' performance in identifying associations of rare variants, sometimes jointly with common variants.

All 12 contributions to GAW17 Group 15 used the simulated data of 697 unrelated individuals. Taking advantage of 200 simulated data sets (simulated phenotypes with measured genotypes from the 1000 Genomes Project) of one binary trait and three quantitative traits and the available underlying genetic models, many contributors assessed type I error and power of existing and novel collapsing methods. These evaluations and comparisons help us to understand the performance of these methods in terms of both type I error and power.

Several existing collapsing methods for rare variant analysis—collapsing and summation test (CAST), indicator coding test RVT2 (rare variant test 2), combined multivariate and collapsing (CMC) method, weighted-sum (WS) collapsing method, data-adaptive summation (aSUM), and variable threshold (VT) collapsing method—have been assessed by Group 15 contributors. The CMC, WS, aSUM, and VT methods also provide flexible frameworks with which to analyze collapsed rare variants and common variants jointly. These methods have been summarized by Dering et al. [2011a]. Here, we describe novel approaches suggested and explored by Group 15 contributors.

Collapsing Methods for Rare Variant Analysis: Cumulative Minor Allele Test {AU: Because there is only one subhead, we can combine the main head and the subhead. This is standard practice for journals and books.}

The cumulative minor allele test (CMAT) for rare variant analysis is derived from the chi-square statistic and compares the total number of rare variants present in the gene for case subjects and control subjects [Zawistowski et al., 2010]. For a sample with N_A case subjects and N_U control subjects, assume $F > 1$ variants in the region of interest (ROI), each with a weighting factor $w_j \geq 0$ ($j = 1, \ldots, F$). The CMAT statistic compares the proportion of rare alleles in the case subjects to the proportions in the control subjects as follows:

$$\Sigma_{\text{CMAT}} = \frac{N_A + N_U}{2N_A N_U \sum_j w_j} \frac{\left(m_A M_U - m_U M_A\right)^2}{\left(m_A + m_U\right)\left(M_A + M_U\right)}, \tag{1}$$

where

$$m_A = \sum_{i=1}^{N_A} \sum_{j=1}^{F} w_j X_{ij} \tag{2}$$

and

$$m_U = \sum_{i=1}^{N_U} \sum_{j=1}^{F} w_j Y_{ij} \tag{3}$$

are the weighted minor allele counts across all sites in the ROI for case and control subjects, respectively.

Because the genetic variants may be in linkage disequilibrium and have small counts, a permutation strategy that shuffles the case-control labels and maintains the correlation structure of the genetic data is used to determine the statistical significance of the CMAT statistic. Luedtke et al. [2011] evaluated CMAT along with the CMC, WS, and RVT2 methods for their performance in analyzing the dichotomized trait. One thousand permutations were performed to determine the empirical *p*-value of each gene, that is, the ROI [Luedtke et al., 2011].

Genetic Similarity and Distance

An alternative approach to studying genetic associations is to investigate the relationship between pairs of individuals. This approach can be particularly useful for identifying clusters of individuals in the phenotype-genotype space. Methods based on correlating genotypic similarity and phenotypic similarity have been developed for genetic epidemiological research [Shannon et al., 2002; Beckmann et al., 2005]. In GAW17 Group 15, two pairwise approaches, one based on the kernel function and the other based on the Mantel test, were implemented and assessed using the GAW17 data.

Kernel-Based Association Test

The kernel-based association test (KBAT) combines multiple genetic variants and reduces the degrees of freedom and was initially proposed to study the genetic association of common variants [Kwee et al., 2008]. KBAT extends the least-squares kernel machines for quantitative traits and the logistic kernel machine for dichotomized traits to study multivariable

associations [Kwee et al., 2008; Wu et al., 2010]. Li et al. [2011] implemented and investigated the KBAT because it is suitable for combining multiple rare variants within a ROI. For quantitative traits, β and h are estimated by maximizing the penalized likelihood function

$$J(h,\beta) = -\frac{1}{2}\sum_{i=1}^{n}\left[y_i - X_i^T\beta - h(G_{ik})\right]^2 - \frac{1}{2}\lambda\|h\|^2, \quad (4)$$

where λ is a tuning parameter. The solution to the nonparametric function $h(\cdot)$ can be expressed as

$$h(G) = \sum_{i=1}^{n}\alpha_i K(G, G_{ik}) \quad (5)$$

for a given kernel function $k(\cdot, \cdot)$. {AU: In the preceding sentence, should the k in $k(\cdot, \cdot)$ be capitalized, i.e., $K(\cdot, \cdot)$, to match Eq. 5 and the subsequent text?} The kernel function $K(G_{ik}, G_{jk})$ measures the genetic similarity between two individuals i and j at the SNPs in gene k. The estimates of β and α (equivalently, h) can be obtained by plugging $h(G)$ into the penalized likelihood function.

Li et al. [2011] implemented the kernel function based on genetic similarity measured by identity-by-state (IBS) sharing between two individuals i and j at the SNPs within gene k. To consider the potentially larger effect of the genetic variants with lower frequency, this flexible kernel function can be weighted by minor allele frequency (MAF) q, as follows:

$$K(G_{ik}, G_{jk}) = \frac{\sum_{l=1}^{s} w_{lk} \mathrm{IBS}(M_{lik}, M_{ljk})}{2s}, \quad (6)$$

where M_{rik} denotes the genotype of individual i at SNP r in gene k and w_{lk} is a weight based on q_{lk} (the MAF of SNP l within gene k) and equals $1/(q_{lk})^{1/2}$.

Mantel Test

Instead of directly testing the associations between rare variants or their collapsed summary statistics, Sun et al. [2011] implemented a collapsing method to examine the correlation between the genetic dissimilarity (i.e., the genetic distance) and phenotypic dissimilarity. This method is a Mantel-type statistic that tests the dependence between the elements of two matrices [Mantel, 1967]. The two matrices contain data from multiple variables obtained on a common sample of subjects, where the rows correspond to the subjects and the columns contain data on the two sets of variables X and Y. For n subjects with two variables X and Y, two distance matrices, each with $n(n-1)/2$ pairwise distances, are first calculated. The Mantel statistic is based on the cross-product term

$$Z = \sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij} Y_{ij}, \quad (7)$$

where n denotes the number of subjects in the distance matrices and X_{ij} and Y_{ij} are the pairwise distances between subjects i and j. The elements of a distance matrix are not independent, and determination of the type I error level for the correlation (i.e., Mantel's statistic Z) between two distance matrices is not straightforward. Therefore the significance level is usually evaluated

using a permutation procedure [Beckmann et al., 2005]. For a given gene, the genetic distance X_{ij} between each pair of subjects is calculated as the sum of difference of the additive effect on each rare SNP. For a SNP, the distance between two homozygotes is 2, but the distance is 1 between homozygote and heterozygote genotypes. The genetic distance between a pair of subjects on the gene level is the sum of the genetic distance of individual SNPs. For a gene involving two SNPs with alleles *A*/*a* and *B*/*b*, the genetic distance between a pair of individuals ranges from 0 (same genotype) to 4 (*AABB* vs. *aabb*). The phenotypic distance is the absolute difference of the phenotypic value between a pair of individuals ($|Y_i - Y_j|$).

Integrated Analysis of Both Rare and Common Variants

As shown in several previous studies [Li and Leal, 2008; Madsen and Browning, 2009], combining rare and common variants within a ROI can improve the power of identifying a disease-associated region when both types of genetic variants contribute to the disease. Two Group 15 studies proposed alternative approaches to jointly analyzing rare and common variants by applying variable selection and the least absolute shrinkage and selection operator (LASSO) [Dasgupta et al., 2011].

Data-Adaptive Forward Selection

Dai et al. [2011] proposed a three-step procedure that uses the associated common SNP as an anchor to select the rare variants. The variable selection procedure starts with selecting the most significant common SNP in the ROI. Then the rare SNPs that improve the goodness-of-fit are added to the model one at a time. The selection of rare SNPs repeats until no such rare SNP exists in the ROI. The goodness-of-fit is measured by the *F* statistic of a linear regression model

that combines the selected variants into a collapsing score. The final test statistic is the absolute value of the *t* statistics for the final linear regression model $E(y) = \alpha + \beta S_{\text{final}}$, where S_{final} is the final collapsed score including all selected common and rare variants. Without knowledge of the distribution of the final *t* statistic, a genome-wide permutation needs to be performed to evaluate the global empirical *p*-value. This data-adaptive forward selection procedure selectively chooses only variants that improve the joint association between the ROI and the disease trait.

LASSO

The LASSO is an efficient variable selection method for high-dimensional data analysis [Tibshirani, 1996]. Recently, the LASSO and its variants have been adapted to the analysis of high-dimensional genetic variants [Szymczak et al., 2009; Dasgupta et al., 2011]. Chen et al. [2011] applied LASSO regression to select common variants that should remain in the model. They performed a 10-fold cross-validation for estimating the shrinkage parameter (λ) for the LASSO. The common variants remaining in the model after LASSO selection and the covariates and rare variant score were then fitted in a multiple regression model. The joint genetic association of common and rare variants was tested using the partial *F* test.

Table I summarizes the analyses of rare variants performed by GAW17 Group 15 contributors. Both the quantitative traits and the dichotomized trait were analyzed. Because the contributors decided to be either blinded or unblinded to the simulation answers, the analytical strategies discussed during the GAW17 meetings were heterogeneous. However, all contributors chose to use similar analytical approaches in their final contributions. Given the causal genetic associations simulated in 200 replicates, all work groups evaluated the performance of existing or novel approaches by testing type I error fraction and power [Chen et al., 2011; Dai et al.,

2011; Dering et al., 2011b; Luedtke et al., 2011; Sun et al., 2011] or receiver operating characteristic (ROC) curves with similar measurements [Li et al., 2011; Lin et al., 2011; Sung et al., 2011]. Because all causal SNPs were nonsynonymous in the simulation model, six out of nine work groups examined the performance of collapsing methods by including nonsynonymous SNPs only. Almost all contributors implemented permutation tests to determine the statistical significance resulting from the nonstandard distribution of the test statistics derived from the collapsing methods. The inclusion of covariates was also considered to assess its impact on the performance of these methods.

Results

After extensive investigations of the collapsing methods for rare variant analysis, we observed several common themes in our group. Although the power can be improved under specific scenarios, such as filtering nonsynonymous SNPs and inclusion of appropriate covariates, the overall performance of all tested methods was similarly poor. By adjusting for multiple testing of thousands of genes, all collapsing methods were underpowered to detect genes with causal rare variants in 697 unrelated samples except for a few top genes, such as *FLT1* and *KDR* for the simulated quantitative trait Q1.

We also observed surprisingly high type I error fractions for Q1 and Q2 across all tested methods. For Q4, which did not have any causal genetic variants simulated, the type I error fraction of the tested methods was not inflated. Two work groups in Group 15 further investigated the potential causes of the inflated type I error [Luedtke et al., 2011; Sung et al., 2011] and identified hundreds of SNPs in gametic phase disequilibrium with the causal rare variants. Gametic phase disequilibrium, also called gametic disequilibrium, is the nonrandom

correlation between genetic loci. Here, we use the term to define such nonrandom correlation between loci that are located beyond a local haplotype block, sometimes on different chromosomes. In extreme cases, a noncausal SNP had identical genotypes as the causal SNP for all 697 individuals.

Another potential cause of inflated type I error is population structure. Principal components analysis (PCA), which is used to adjust for population structure, reduces type I error [He et al., 2011; Luedtke et al., 2011], but the effectiveness of controlling the false positives is influenced by the MAF [He et al., 2011]. Specifically, PCA reduced false-positive fractions more effectively in common SNPs (MAF > 0.05) than in rare SNPs (MAF < 0.01). Unfortunately, although false-positive fractions were reduced, the power to detect true associations was also reduced by using PCA.

Not surprisingly, we confirmed that the power for identifying associations of rare variants can be improved by including the appropriate prior knowledge. For instance, incorporating the correct covariates in the model increased both sensitivity and specificity [Lin et al., 2011]. In addition, the power to identify associated genes increased by selecting only the nonsynonymous SNPs for all collapsing methods, because all causal rare variants were nonsynonymous SNPs in the simulation model. Meanwhile, the synonymous-SNPs-only test served as a negative control for assessing the false-positive fraction. Given that all simulated effects of rare variants were deteriorating, the data-adaptive methods, such as the aSUM method, did not perform better than the non-data-adaptive methods.

We also observed interesting features of some novel approaches. The forward selection method combining both common and rare variants achieved substantially higher power than

other methods, which considered either rare or common variants regardless of their associations with the outcome [Dai et al., 2011].

Discussion

Inflation of Type I Error

Overall we found that collapsing methods had limited capability to identify most causal genes because of low power and high type I error fraction. Inflated type I error fractions were consistently observed for simulated traits using all collapsing methods. The expected type I error for Q4, which has no causal variants, suggested that the inflation was not due to the statistical methods. A large number of SNPs in noncausal genes were in gametic phase disequilibrium with causal SNPs simulated for Q1 and Q2 [Li et al., 2011; Luedtke et al., 2011; Sung et al., 2011]; that is, they were highly correlated with causal SNPs. These highly correlated or even identical genotypes may account for the large number of false-positive genes for Q1, Q2, and the affection status. Because the number of rare variants (especially the private mutations observed only once in the sample) was much larger than the sample size, it was more likely that causal and noncausal variants shared the same or similar genotype distributions. For the sample size of 697, at least two private mutations were on the same individual's genome among every 698 rare variants with private mutations. These pairs of rare variants, which may include simulated causal variants, had perfect correlation. The spurious correlations among causal and noncausal rare variants could have been caused by gametic phase disequilibrium and population stratification [Luedtke et al., 2011]. A further analysis showed a higher amount of gametic phase disequilibrium than random chance, which suggests potential genotyping errors from the exomic sequencing; however, no

information about genotype calling, cleaning, or other preprocessing steps were available for these data to allow us to further pinpoint the cause.

Permutation and Computation

Because of the unknown distribution of the test statistics and because of potential correlation among genetic variants, the Group 15 contributors needed to use permutation to accurately determine the statistical significance for most collapsing methods of rare variant analysis. Permutation is a computationally expensive procedure that generates the null distribution of the test statistic by deconstructing the relationship between predictors and outcome repeatedly. Permutation of high-dimensional data requires a large amount of computational resources. To achieve a significant empirical *p*-value adjusted for multiple testing, investigators need hundreds of thousands of permutations to scan the human exome. In a simplified example, a Bonferroni-corrected *p*-value of 0.05 for 20,000 human genes requires at least 400,000 permutations (1/(0.05/20,000)). Although the total number of human genes is approximately fixed in the equation, other factors (e.g., a non-gene-based ROI) might increase the total number of permutations needed.

Improvements in both software and hardware can address this computational challenge. Well-implemented algorithms can greatly speed up the computation of high-throughput data for genetic epidemiological studies [Schwarz et al., 2010]. A large number of permutations can also be easily parsed into smaller jobs to take advantage of parallel computing available in all high-performance computing facilities and smaller scale computer clusters. The multicore design of the latest CPUs and graphic processing units (GPUs) combined with parallel programming

techniques will help genetic epidemiologists to address computational challenges in analyzing rare variants and sequencing data.

On the other hand, genetic epidemiologists start to face another issue of data management: how to securely transfer, store, and back up the large amount of sequencing data (terabytes rather than gigabytes). A well-designed computation environment will be critical to conquer these technical bottlenecks for analyzing next-generation sequencing data. More important, like all new forms of data for genetic epidemiological studies, the next-generation sequencing data need to be carefully examined and cleaned. The issue of data quality cannot be overemphasized at the early stage of the analysis of next-generation sequencing data. The valid scientific findings have to rely on high-quality sequencing data that are likely to be more critical for studying rare variants.

Real Data Analysis: The More We Know, the More We Know What We Don't Know

One of the biggest questions still unanswered after extensive studies of simulated data is how to best analyze the real data with many rare variants from sequencing projects. With limited simulation models, we are not able to evaluate all scenarios where these assumptions are violated. In practice, we do not know the location and effect of noncoding causal variants (e.g., *cis*-regulatory elements), the proportion of causal variants that change protein coding (nonsynonymous vs. synonymous SNPs), the weight of a genetic variant relative to the allele frequency, the direction of the genetic effects (deteriorating vs. protective), or the boundary of an ROI. In that regard, the data-adaptive approaches and methods that require minimal assumptions are more favorable. Several interesting directions should be considered or combined in the future analysis of rare variants. The aSUM statistic addresses the issue that causal variants may not

have effects all in one direction. Unfortunately, the GAW17 data simulated only deteriorating effects that do not allow a formal test to examine the benefit of this flexible approach. The weighted KBAT seems to be less sensitive to the assumption of lower MAF and larger effect size. It suffers less power loss than the WS method when this assumption is violated [Li et al., 2011].

All Group 15 contributors chose the gene, a natural unit of the human genome, as the ROI for collapsing genetic variants. However, the ROI can also be based on nongenic regions, such as transcriptional regulatory regions, functional domains within a gene, or a set of related genes (e.g., a protein complex or a pathway) [Dering et al., 2011a; Tintle and Pugh, 2011]. The benefits of choosing alternative ROIs cannot be assessed using the GAW17 data because of the limitation of the simulation. In real data analysis, using an alternative collapsing unit may help to identify disease-related molecular mechanisms by enriching the genetic signal.

Sequencing data provide the ultimate resolution of the genetic variants in human DNA. With large populations sequenced, eventually we will be able to identify the causal genetic associations on the human genome. With the improvements in measurement, data processing, management, and methods of analysis, we will be able to further understand the genetic causes of complex diseases and develop strategies to deliver better prevention, diagnosis, and treatment to the public.

Acknowledgments

The Genetic Analysis Workshops are supported by National Institutes of Health (NIH) grant R01 GM031575. YVS was supported in part by NIH grant HL100245 from the National Heart, Lung, and Blood Institute. YJS was supported by NIH grants HL54473, HL45670, and GM28719. NT

was supported by NIH–National Human Genome Research Institute grant R15 HG004543. The work of AZ was supported by an intramural grant from the University of Lübeck.

References

{AU: This paper doesn't exist.}{AU: This paper doesn't exist either.}

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

Beckmann L, Thomas DC, Fischer C, Chang-Claude J. 2005. Haplotype sharing analysis using Mantel statistics. Hum Hered 59:67–78.

Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoglu D, Tuysuz B, Caglayan AO, Gokben S, et al. 2010. Whole-exome sequencing identifies recessive *WDR62* mutations in severe brain malformations. Nature 467:207–10.

Chen H, Hendricks AE, Cheng Y, Cupples LA, Dupuis J, Liu C-T. 2011. Comparison of statistical approaches to rare variant analysis for quantitative traits. BMC Proc 5(suppl 9):S113.

Dai Y, Guo L, Dong J, Jiang R. 2011. Improved power by collapsing rare and common variants based on a data-adaptive forward selection strategy. BMC Proc 5(suppl 9):S114.

Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. 2011. Brief review of regression-based and machine learning methods in genetic epidemiology: the GAW17 experience. Genet Epidemiol X(suppl X):X–X.

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

Dering C, Ziegler A, König IR, Hemmelmann C. 2011b. Comparison of collapsing methods for the statistical analysis of rare variants. BMC Proc 5(suppl 9):S115.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–50.

Gilissen C, Arts HH, Hoischen A, Spruijt L, Mans DA, Arts P, van Lier B, Steehouwer M, van Reeuwijk J, Kant SG, et al. 2010. Exome sequencing identifies *WDR35* variants involved in Sensenbrenner syndrome. Am J Hum Genet 87:418–23.

Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82:100–12.

He H, Zhang X, Ding L, Mersha TB, Kurowski BG, Martin LJ. 2011. Effect of population stratification analysis on false-positive rates for common and rare variants. BMC Proc 5(suppl 9):S116.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80:727–39.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet 82:386–97.

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, et al. 2010. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 465:473–7.

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

Li L, Zheng W, Lee JS, Zhang X, Ferguson J, Yan X, Zhao H. 2011. Collapsing-based and kernel-based single-gene analyses applied to Genetic Analysis Workshop 17 mini-exome data. BMC Proc 5(suppl 9):S117.

Lin WY, Zhang B, Yi N, Gao G, Liu N. 2011. Evaluation of pooled association tests for rare variants identification. BMC Proc 5(suppl 9):S118.

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 5(suppl 9):S119.

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, et al. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. New Engl J Med 362:1181–91.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5:e1000384.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. Nature 461:747–53.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–20.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, et al. 2010a. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. Nat Genet 42:790–3.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, et al. 2010b. Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 42:30–5.

Schwarz DF, Konig IR, Ziegler A. 2010. On safari to Random Jungle: a fast implementation of random forests for high-dimensional data. Bioinformatics 26:1752–8.

Shannon WD, Watson MA, Perry A, Rich K. 2002. Mantel statistics to correlate gene expression levels from microarrays with clinical covariates. Genet Epidemiol 23:87–96.

Sobreira NL, Cirulli ET, Avramopoulos D, Wohler E, Oswald GL, Stevens EL, Ge D, Shianna KV, Smith JP, Maia JM, et al. 2010. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. PLoS Genet 6:e1000991.

Sun YV, Zhao W, Shedden KA, Kardia SL. 2011. Identification of genes associated with complex traits by testing the genetic dissimilarity between individuals. BMC Proc 5(suppl 9):S120.

Sung YJ, Rice TK, Rao DC. 2011. Application of collapsing methods for continuous traits to the Genetic Analysis Workshop 17 exome sequence data. BMC Proc 5(suppl 9):S121.

Szymczak S, Biernacka JM, Cordell HJ, Gonzalez-Recio O, Konig IR, Zhang H, Sun YV. 2009. Machine learning in genome-wide association studies. Genet Epidemiol 33(suppl 1): S51–7.

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

Tintle N, Pugh E. 2011. Inflated type I error rates when using aggregation methods to analyze rare variant data in unrelated individuals: a summary report from Group 7 at Genetic Analysis Workshop 17. Genet Epidemiol X(suppl X):X–X.

Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, et al. 2010. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am J Hum Genet 87:90–4.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet 86:929–42.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–9.

Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. 2010. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87:604–17.

Table I. Overview of Group 15 contributions

Contribution	Phenotype	Collapsing method	Rare SNP MAF (%)	SNP filter	Type I error	Power	Permutation
Chen et al. [2011]	Q1, Q4	RVT, aSUM, CMC, LASSO	0.5, 1, 5	All	Y	Y	Y
Dai et al. [2011]	Q1, Q2	Forward selection, CAST, WS, RVT	1	All	Y	Y	Y
Dering et al. [2011b]	Affected	CAST, CMC, WS, RVT	1, 2, 3, 5, 7, 10	All, nsyn, syn	Y	Y	Y
He et al. [2011]	Q1	RVT, aSUM	1	nsyn	Y	Y	Y
Li et al. [2011]	Q1, Q2, Q4, Affected	CMC, WS, KBAT, Bayesian mixed-effects model	5	nsyn, syn	Y	Y	Y
Lin et al. [2011]	Q1, Q2, Q4, Affected	VT, WS, RVT	1, 5	All	Y	Y	Y
Luedtke et al. [2011]	Affected	CMC, WS, RVT, CMAT	5	All, nsyn	Y	Y	Y
Sun et al. [2011]	Q1, Q4, Affected	Mantel test	5	All, nsyn	Y	Y	Y
Sung et al. [2011]	Q1, Q2	RVT	5	All, nsyn	Y	Y	N

CAST, collapsing and summation test; CMC, combined multivariate and collapsing method; WS, weighted-sum method; aSUM, adaptive summation method; RVT, rare variant test; VT, variable threshold method; CMAT, cumulative minor allele test; KBAT, kernel-based association test; nsyn, nonsynonymous; syn, synonymous.