Regression and Data Mining Methods for Analyses of Multiple Rare Variants in the Genetic Analysis Workshop 17 Mini-Exome Data

Joan E. Bailey-Wilson,[1]* Jennifer S. Brennan,[2] Shelley B. Bull,[3,4] Robert Culverhouse,[5] Yoonhee Kim,[1] Yuan Jiang,[2] Jeesun Jung,[6,7] Qing Li,[1] Claudia Lamina,[8] Ying Liu,[9] Reedik Mägi,[10] Yue S. Niu,[11] Claire L. Simpson,[1] Libo Wang,[12] Yildiz E. Yilmaz,[3,4] Heping Zhang,[2] and Zhaogong Zhang[13]

[1]*Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD*
[2]*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT*
[3]*Samuel Lunenfeld Research Institute of Mount Sinai Hospital, Toronto, ON, Canada*
[4]*Dalla Lana School of Public Health, University of Toronto, ON, Canada*
[5]*Department of Medicine, Washington University School of Medicine, Saint Louis, MO*
[6]*Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN*
[7]*Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN*
[8]*Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria*
[9]*Department of Statistics, Columbia University, New York, NY*
[10]*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK*
[11]*Department of Mathematics, University of Arizona, Tucson, AZ*
[12]*Department of Statistics, Purdue University, West Lafayette, IN*
[13]*Department of Mathematical Sciences, Michigan Technological University, Houghton, MI*

Group 14 of Genetic Analysis Workshop 17 examined several issues related to analysis of complex traits using DNA sequence data. These issues included novel methods for analyzing rare genetic variants in an aggregated manner (often termed collapsing rare variants), evaluation of various study designs to increase power to detect effects of rare variants, and the use of machine learning approaches to model highly complex heterogeneous traits. Various published and novel methods for analyzing traits with extreme locus and allelic heterogeneity were applied to the simulated quantitative and disease phenotypes. Overall, we conclude that power is (as expected) dependent on locus-specific heritability or contribution to disease risk, large samples will be required to detect rare causal variants with small effect sizes, extreme phenotype sampling designs may increase power for smaller laboratory costs, methods that allow joint analysis of multiple variants per gene or pathway are more powerful in general than analyses of individual rare variants, population-specific analyses can be optimal when different subpopulations harbor private causal mutations, and machine learning methods may be useful for selecting subsets of predictors for follow-up in the presence of extreme locus heterogeneity and large numbers of potential predictors. *Genet. Epidemiol.* 35:S92–S100, 2011.

Key words: rare variants; LASSO; machine learning; random forests; logic regression; binary trees; Poisson regression; ISIS; classification trees; meta-analysis; extreme sampling

*Correspondence to: Joan E. Bailey-Wilson, NIH/NHGRI, 333 Cassell Drive, Suite 1200, Baltimore, MD 21224. E-mail: jebw@mail.nih.gov
Published online in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.20657

INTRODUCTION

The overarching theme of Genetic Analysis Workshop 17 (GAW17) was the comparison of statistical methods for detecting genetic contributions to variability of complex traits using whole-exome DNA sequence data. The papers contributed to GAW17 were grouped by theme for discussion and comparison of performance of methods. The contributors to Group 14, on regression and data mining methods for multiple rare variants, addressed several issues in the 13 papers submitted to the workshop. These issues included novel methods for analyzing rare genetic variants in an aggregated manner (often termed collapsing rare variants), evaluation of various study designs to increase power to detect effects of rare variants, and the use of machine learning approaches to model highly complex heterogeneous traits.

COLLAPSING RARE VARIANTS

Conventionally, rare variants refer to variants (single-nucleotide polymorphisms [SNPs]) with a minor allele frequency (MAF) of 0.5–1%, and very rare variants refer to SNPs with MAF <0.5%. In Mendelian diseases, a single causal gene or pathway often harbors large numbers of individually very rare causal variants. Evolutionary theory suggests that even for complex traits, it is likely that causal genes (and related pathways) will harbor multiple very rare variants (in the population) that will contribute to the susceptibility to the disease, with individual families each segregating different causal variants in the same gene (or pathway). Association methods have limited power for mapping very rare variants in population studies because of the small number of sampled individuals carrying such a variant. For these methods to have adequate efficiency to detect individual rare variants of moderate effect size, the sample size needs to increase substantially as the MAF decreases. In the general context of biostatistics, the problem is known as the imbalance in the covariate distributions. For binary traits with categorical covariates, the problem can be translated into low cell counts in a cross-tabulation table. Therefore grouping and collapsing rare variants within genes or pathways is the most feasible option to improve power when studying rare variants in a sample of individuals who may each harbor different rare causal variants.

The remaining question is, What is the most powerful way to generate a surrogate for the aggregated genetic effect from a set of rare variants? Usually, grouping is constructed on the basis of functional relevancy, physical proximity, or both. Once rare variants are grouped, their genotypic information is typically combined or collapsed into a univariate score. Subsequently, the association between this group of rare variants and the disease is studied by means of the association between the score and the disease traits. Dering et al. [2011] provide an excellent review of various published rare variant association methods, including single-marker, multimarker, and various collapsing strategies.

A popular alternative to collapsing genotypic information is to combine single-SNP statistics. Six work groups in Group 14 attempted to address these issues using novel methods, and two work groups evaluated previously published methods using the GAW17 data. The new methods all used weights to aggregate the effects of multiple markers. Some methods derived weights from the data simultaneously with the association analysis, whereas others assumed uniform weights or weights inversely proportional to MAF before performing the association tests. One can expect that the derived-weight approaches may lose some power because some of the data are used to estimate the weights. The uniform- or inverse-weight approaches rely on the accuracy of their assumptions concerning the effects of rare variants to gain power, but if these assumptions are incorrect, then their power may not be optimal.

STUDY DESIGNS

Sequencing studies are expected to remain fairly expensive for several years, and large samples are needed to detect the effects of rare variants of small effect size. Therefore study designs that can increase power are desirable. Several contributors to Group 14 proposed ideas to incorporate information on extreme phenotype values at the design and analysis stages of their association studies. In addition, some work groups compared study designs in which analyses were performed on stratified subpopulations followed by meta-analysis with designs in which analysis were performed on the complete sample while adjusting for subpopulation membership. Four work groups focused in full or in part on optimal study design methods.

MACHINE LEARNING METHODS

Over the past decade, technological advances have led to the ability to measure large numbers of possible predictors (genotypes and environmental exposure values) that may play a role in variation of a quantitative trait or risk of a disease. This has led to the so-called small n, large P problem, where the number of potential predictors P to be tested for association with the trait is many orders of magnitude larger than the number of individuals n who are studied. When one considers possible interactions among these predictors, the number of tests becomes so large that traditional frequentist tests have no power after correction for multiple testing. The various highly nonparametric methods that model risk of disease or variation of a quantitative trait and that rely on bootstrapping approaches to rank potentially causal predictors have been termed machine learning methods. Dasgupta et al. [2011] provide a review of the published machine learning methods that were used at GAW17. Two work groups applied three existing machine learning methods to the GAW17 data.

In the remaining sections of this paper, we summarize the simulation model and the methods and results for each paper under the three mentioned categories and discuss the results in light of the data-generating model. Some interesting observations from the analysis are also highlighted for further follow-up.

DATA

Of the two genetic data sets provided for analysis at GAW17, all contributors to Group 14 used the set that consisted of 697 unrelated individuals sequenced by the 1000 Genomes Project from seven ethnically homogeneous samples representing three geographically distinct ancestries: Asia, Africa, and Europe. A subset of whole-exome DNA sequence data from the 1000 Genomes Project was used to simulate a common, complex disease trait and three related quantitative traits. All four traits were modeled to have complex causation, involving both genetic and environmental factors. The sequence data, the simulation process, and the generating models for the one qualitative and three quantitative traits are described by Almasy et al. [2011]. Briefly, in the trait models, at causal loci for each trait, all genetic effects were additive on the quantitative trait or liability scale and the allele with the lower frequency always had a positive effect on the quantitative trait value or on the disease risk. Overall, SNP locus-specific heritability (h^2) averaged over 200 replicates of the complete data were all extremely small, with few variants having a locus-specific $h^2 > 1\%$ [Wilson and Ziegler, 2011]. Quantitative trait Q4 had a heritability of 0.7, but none of this was due to genes in the mini-exome sequencing set. For Q1, a gene-environment interaction was simulated for the *KDR* gene and the Smoking covariate. Although population origin of the 1000 Genomes Project participants was not used in the phenotype simulations, the causal allele frequencies differed across the different ethnic groups. No epistatic interactions were simulated for any of the traits.

METHODS FOR COLLAPSING RARE VARIANTS

In GAW17 Group 14, six work groups proposed novel approaches to collapse rare variants, as summarized in

Table I. In addition, two work groups evaluated previously published collapsing methods. By collapsing multiple markers into one unit for analysis, each approach limited the number of conducted tests in an attempt to increase power compared to the situation in which all rare variants were tested individually. Only a short description of the methods and main results are presented here, and readers are referred to the original papers for other details of each method.

Three work groups [Brennan et al., 2011; Niu et al., 2011; Wang et al., 2011] aimed to generate a linear combination of genotypes within a gene and relied on various regression techniques to test for association of the trait with individual genes. Jung et al. [2011] used the count of rare variants within a gene as the surrogate for the gene, but in their association test this count became the response variable and the phenotype became the independent variable in a zero-inflated Poisson regression model. Jiang et al. [2011] proposed a two-step procedure that relied on classification trees to combine markers into multiple groups, some containing markers within the same gene (step 1) and others containing markers in genes physically close to each other (step 2). Then the collapsed group was tested for association with the phenotype using traditional regression analysis. These four methods focused on the grouping of markers, whereas Liu et al. [2011] aimed to divide samples into subgroups as well. They proposed a metric measure, denoted by a genotypic similarity score, to calculate the pairwise similarity among samples at a given marker. Then, unsupervised hierarchical clustering was applied to the aggregated similarity scores among markers within genes to divide the genes into groups. As a result, the samples were also divided into a fixed number of groups. To evaluate the existing collapsing methods, Culverhouse et al. [2011] used the data sets to compare the performance of two similar methods proposed by Li and Leal [2008]. Mägi et al. [2011], on the other hand, evaluated the effect of missing genotypes and therefore applied rare variant mutational load analysis on the provided data sets with some proportions of the genotypes randomly set as missing.

ZERO-INFLATED POISSON REGRESSION FOR MULTIPLE RARE VARIANTS

Jung et al. [2011] applied a zero-inflated Poisson regression model that took into account the excess of 0's caused by the extremely low frequency of rare variants. To implement this, Jung and colleagues created an indicator variable for each SNP based on the presence or absence of the rare variant. Then, summation of the variables within a gene was used to collapse rare variants. The sum is

TABLE I. Novel methods to collapse rare variants proposed by GAW17 Group 14

Group 14 contribution	Method	Idea	Phenotype analyzed
Jung et al. [2011]	Zero-inflated Poisson regression	$Y_i = \sum_{k=1}^{n_j} V_{ijk}$, where $V_{ijk} = 1$ if rare variants are present and 0 otherwise for the kth SNP of the jth gene for the ith individual and n_j is the total number of SNPs on the jth gene	Affected, Q1, Q2, Q4
Brennan et al. [2011]	Least absolute shrinkage and selection operator (LASSO)	LASSO for generating linear combinations of multi-SNP genotypes	Affected
Wang et al. [2011]	Partial least-squares and penalized orthogonal-components regression	For each gene, $Y = \mu + \sum_{j=1}^{k} \beta_j X_j + \varepsilon$, where Y is the phenotype vector and $\{X_1, \ldots, X_k\}$ are the genotypes of k rare variants within the gene. Partial least-squares finds linear combinations of X that explain the covariance between Y and X as much as possible. The linear combination here is merely seen as a construct and is not used for making inferences about the importance of the rare variants	Affected
Niu et al. [2011]	Group iterative sure independence screening (ISIS)	For all the genes, $\mathbf{y} = \sum_{j=1}^{J} X_j \beta_j + \varepsilon$, where $\mathbf{y}$ is the phenotype vector, X_j are the SNP genotypes within the jth gene, and ε is a normally distributed random noise vector	Q1, Q2
Jiang et al. [2011]	Classification tree	Collapsed markers are defined based on SNP interactions detected by classification trees.	Affected
Liu et al. [2011]	Inverse-probability weighted clustering	The following similarity scores are used: $\begin{array}{llll} & & \text{Individual 1} & \\ & & a & A \\ \text{Individual 2} & a & 1/p_a^2 & -1/[p_a(1-p_a)] \\ & A & -1/[p_a(1-p_a)] & 1/(1-p_a)^2 \end{array}$ where p_a denotes the minor allele frequency of a; and $\text{sim}(i,j) = \sum_{k \in G} \text{sim}(i,j;k)$, where $\text{sim}(i, j; k)$ is the genotypic similarity score between two individuals i and j at SNP k.	Affected, Q1, Q2, Q4

essentially a count of all rare SNPs in a gene. Treating the counts as response variables, Jung and co-workers applied a zero-inflated Poisson model based on the results of collapsing the variants, with Affected status (or quantitative traits Q1, Q2, or Q4), Age, Sex, and Smoking status as covariates, at the same time adjusting for population substructure using PLINK [Purcell et al., 2007]. This method had over 90% power to detect *FLT1* for both Q1 and the disease trait at a *P*-value of at least 1.56×10^{-5}. *KDR* was detected at this *P*-value in more than 50% of replicates for Q1. However, many noncausal genes were also significant at these same *P*-values in large numbers of replicates.

LASSO FOR COLLAPSING MULTIPLE VARIANTS

Brennan et al. [2011] considered a two-step approach of analyzing rare variant data by incorporating the least absolute shrinkage and selection operator (LASSO) technique (reviewed by Dasgupta et al., 2011). In the first step, Brennan and colleagues conducted a gene-level screening on SNPs within each gene using the LASSO method. In particular, indicator variables were created on the basis of the rare variants. A selection model was used to generate a linear combination of these new constructs, subsequently producing a new marker that represented the SNP group. A new marker was accepted only if the corresponding linear combination was nonzero in at least 5% of the subjects. In the second step Brennan and colleagues screened the remaining genes by performing clustering analysis on the positions of the SNPs in the genes that were left after the first step. The LASSO method was used to generate a representative marker for each cluster. In Brennan and colleagues' analysis, the first simulated replicate analyzing the disease trait Affected was used to generate markers. The other 199 replications were used to evaluate the set of constructed markers using a classification tree method. Two causal genes were identified in 18% and 28% of replicates, and 16 noncausal genes were identified as significant in 18–55% of replicates. Future development of this method could involve the use of more data to generate markers or incorporation of pathway information in the second step in lieu of clustering genes by distance. In nonsimulated data analyses, the data could be split into training and testing subsets and permutation could be used to control type I error.

PARTIAL LEAST-SQUARES FOR COLLAPSING MULTIPLE VARIANTS

Partial least-squares components are also linear combinations of the predictor variables, which are constructed to maximize an objective criterion based on Cov(X^*w, Y), the sample covariance between X^*w (a linear combination of the original predictors) and the Affected trait. Wang et al. [2011] adopted partial least-squares components for collapsing multiple rare variants within genes and compared the method to analyses of single SNPs. Specifically, for each gene, Wang and colleagues constructed partial least-squares components using cross-validation. Then, they used the penalized orthogonal component regression estimation (POCRE) algorithm [Zhang et al., 2009] to analyze the collapsed "new markers" along with Age, Sex, and Smoking status as covariates. *FLT1* was the only causal locus that was estimated to have a nonzero effect in a large number of replicates (41% in SNP-only analyses and 17.5% in gene-based analyses) with three other causal loci having nonzero estimated effects in 3–10% of replicates. False-positive effects were replicated in at least 3% of replicates for nine genes. Future work will evaluate whether inclusion of pathway information improves performance.

GROUP ISIS FOR COLLAPSING RARE VARIANTS

Niu et al. [2011] used the group iterative sure independence screening (group ISIS) approach to select important genes and the SNPs within. The model considered is given in Table I. Niu and colleagues assumed that the model was bilevel sparse [Breheny and Huang, 2009], which means that only a small number of genes were related to the phenotype of interest and only a subset of the SNPs in these related genes were important. This is a reasonable assumption for high-dimension modeling when there is a group structure among the predictors. This method performed well in the low- and medium-noise cases when results were averaged across multiple replicates. However, when only a single replicate was used, it had lower power (e.g., only two causal genes, *FLT1* and *KDR*, were detected for Q1). As with many other methods, Niu and colleagues approach exhibited high false-positive rates when only a single replicate was used.

TREE METHOD FOR MULTIPLE VARIANTS

Jiang et al. [2011] used a two-step supervised recursive partitioning process to automatically detect SNP interactions and to define markers by these interactions. In the first step, separate trees were constructed to model Affected status (Y) using the SNPs contained in each gene on a chromosome. A new bilevel marker could be constructed: One level of the marker represented the paths in the tree that led to terminal nodes with a majority of case subjects, and the other level of the marker represented paths that led to terminal nodes with a majority of control subjects. A new single marker for each gene was recorded only so long as the least frequent level was not rare (a user-defined threshold; Jiang and colleagues used $1-0.99^2 = 0.0199$ in these analyses). In the second step, Jiang and colleagues performed a clustering analysis to group SNPs within nearby genes using the genes that did not result in a marker in step 1; a tree was built for the phenotype "Affected" for each cluster group. A new single marker for each group was recorded so long as it was not rare. In their analysis, Jiang and colleagues used the first replicate to generate markers. The remaining 199 phenotype replications were used to conduct logistic regressions, adjusting for Sex, Age, Smoking status, and ethnic group cluster covariates. Using a Bonferroni correction for the number of constructed markers used in the final analysis, Jiang and colleagues found that *FLT1* was significant in 10 out of 199 analyses and that many other markers were significant in individual replicates but were not replicated more than eight times. When the false discovery rate was used, *FLT1* was significant in 17 out of 199 replicates but at the expense of increasing false-positive rates. In nonsimulated data analyses, the data could be split into training and testing subsets, and permutation could be used to control type I error.

INVERSE-PROBABILITY WEIGHTED CLUSTERING

Liu et al. [2011] proposed a novel approach for gene-based grouping and collapsing of SNP genotypes. They defined inverse-probability weighted similarity scores to overweight genotypic differences observed for rare variants. Specifically, for an individual 1 with genotype *aa* and an individual 2 with genotype *Aa*, because *a* is the minor allele, the (*a*, *a*) match will dominate the (*a*, *A*) mismatch and these two individuals will have a high similarity score according to the proposed weighting scheme. Similarities at different SNP loci within a gene were aggregated by summation. Then Liu and colleagues used a bounded monotone-decreasing exponential transformation to convert the obtained similarity into distance. For each gene, unsupervised hierarchical clustering was done on the basis of the resultant multi-SNP weighted distances, and partitions of individuals were created for that gene by subsequently cutting the hierarchical clustering tree into a prespecified number of groups (partition sizes 5 to 10 were compared). Liu and colleagues then evaluated the association between a phenotype trait and the partition constructed from the genotypes using three association tests (one-way analysis of variance, chi-square test, and the partition retention method [Chernoff et al., 2009; Zheng et al., 2010]). For Q1, both *FLT1* and *KDR* were identified in more than 50% of replicates, with *FLT1* detected in all the replicates. For Q2, six genes that contained causal SNPs were identified as top genes, but only *VNN1* (22%) was identified in more than 5% of replicates. However, as with all the methods applied to the GAW17 data, some of the genes that were identified in multiple replicates were false positives.

EVALUATION OF PREVIOUSLY PUBLISHED COLLAPSING METHODS

In addition to the novel methods, Culverhouse et al. [2011] compared the performance of two published methods [Li and Leal, 2008] on Q1, Q4, and affection status. The first method was simply a count of how many rare alleles an individual carried for a particular gene. The second method was dichotomous, indicating whether or not an individual carried at least one rare allele in a particular gene. The two methods performed similarly in the GAW17 data, particularly for the strongest associations. This simply suggests that in these data the outliers in phenotype were not also outliers in terms of the count variable for any gene. Clearly, this cannot be generalized to other genetic models. In addition, Culverhouse and colleagues, like all other investigators using the GAW17 data, observed that there were many more highly significant false positives than expected for traits Q1 and affection status. However, in their analysis of Q4, the type I error rate did not appear inflated. Thus they concluded that it was unlikely that the inflation of type I error in the analyses of the other traits was a completely random effect of using multiple replicates of the same genotypes. They also noted that one outlier individual (with an extreme phenotype at Q1 in nearly every replicate) was the only carrier of rare variants in multiple genes, and this fact became significant only when this person was included in the data.

Mägi et al. [2011] used rare variant mutational load analysis [Li and Leal, 2008; Morris and Zeggini, 2010], as implemented in the program GRANVIL (http://www.well.ox.ac.uk/GRANVIL), to examine the 1,297 causal genes with at least two rare variants. Within the generalized linear model framework, the method collapses the variants for each individual and uses the counts of minor alleles divided by the number of called alleles for a set of markers (e.g., those belonging to the same gene) as one covariate, denoted as the mutational load, in the regression analysis. This approach differs from other collapsing methods using the count of rare variants [Jung et al., 2011] because of the additional loading weights based on call rates for each individual. Because rare variants are more prone to be sequenced unsuccessfully than common (MAF$>$0.05) variants are, Mägi and colleagues evaluated the power and type I error of these methods when various proportions of the genotype data were missing. As with other methods, at a nominal Bonferroni-corrected threshold of $P \leq 3.86 \times 10^{-5}$, they observed association for the Q1 causal *FLT1* gene in all replicates, for the Q1 causal *KDR* gene in about 23% of replicates, and for other causal genes in any of the traits in only small numbers of replicates. The observed power to detect association with causal genes was not dramatically affected by call rate. Similarly, the type I error rate for noncausal genes was relatively unaffected by the rate of missing genotypes but was somewhat inflated at all levels of call rates (as observed for all methods). Their results suggest that the GRANVIL approach for testing association with the mutational load of rare variants within a gene is relatively robust to missing genotype data, which occur either at random or with differential allele-specific failures.

STUDY DESIGNS TO INCREASE POWER TO DETECT RARE CAUSAL VARIANTS

Several methods were proposed to increase power when analyzing sequence data by utilizing information on extreme phenotype values or by stratifying on subpopulations.

OVERSAMPLING SCHEMES BASED ON EXTREME VALUES

In studies where resources are limited, sequencing might be restricted to a subset of the entire cohort. One natural choice is to sequence the upper and lower tails of a quantitative trait distribution. But is this really the best sampling design? If a subset is sampled from the entire cohort, what analysis method should be used? Yilmaz and Bull [2011] evaluated these questions by comparing the following sampling schemes, among others: a 50% simple random sample, an extreme phenotype sampling (taking the upper and lower 25% of the quantitative trait distribution), and a 50% sampling design that gave each individual a nonzero chance to be drawn but with higher probability assigned to those with extreme values. To avoid bias and inflation of type I error, statistical methods have to account for the specific sample ascertainment scheme. These methods include an inverse regression conditioning on the phenotype and standard survey methods with inverse-probability weighting.

As expected, the simple random sample design could not be recommended because of a high loss of power. The quantitative-trait-dependent sampling designs investigated by Yilmaz and Bull [2011], though, did emerge as cost-effective alternatives with the oversampling or complete selection of the extremes of the distribution performing well when analyzed with Poisson regression or linear regression with inverse-probability weighted estimation. The simulation study results suggest that the quantitative-trait-dependent selection designs generally yield greater than 50% relative efficiency compared to using the entire cohort, implying cost-effectiveness of 50% sample selection and worthwhile reduction of sequencing costs.

It was clear, however, that linear regression using the complete distribution of the quantitative trait was more powerful than any of the methods using only 50% of the data. Was the power loss of extreme-value approaches just a result of a smaller sample size? Lamina [2011] investigated this question by restricting the analysis to the tails of the distribution (bottom 10% versus top 10%) using methods specifically designed for the detection of rare variants. Rare variants were collapsed within genes by means of an indicator variable that coded whether the individual had at least one rare variant in the gene. Within this analysis, the sample size was increased (1) by combining pairs of replicates into a single sample, thus doubling the number of individuals with extreme phenotype values in each of 100 combined replicates, and (2) by widening the tails to include the top and bottom 20 or 30% of the sample. Thus Lamina analyzed extreme tail samples, which corresponded to only 20%, 40%, or 60% of the original sample size. Linear regression approaches applied to the extreme tail subsets showed inflated type I error, but methods appropriate for case-control tests and adjusted for population stratification exhibited better control of false-positive rates. Linear regression of the complete data set had higher power than any approach using only the upper and lower 10% of one sample (i.e., using only 20% of the original sample size). However, when sample size was increased by taking the extreme top and bottom 10% from a larger sample (i.e., from two combined replicates; 40% of the original sample size), Lamina observed comparable or even higher power compared to analyzing the complete sample. Increasing the sample size by adding values closer to the median of the quantitative trait distribution, however, did not improve the power. Lamina concluded that enriching the analysis by gathering a greater proportion of individuals with extreme values in the phenotype of interest than in the general population led to a higher power to detect rare variants compared to analyzing a population-based sample with equivalent sample size.

CANDIDATE GENE ANALYSIS DESIGN GUIDED BY EXTREME VALUES

Both Yilmaz and Bull [2011] and Lamina [2011] proposed a specific selection of individuals based on extreme values. In contrast, Zhang et al. [2011] used information on extreme values for the selection of candidate genes. Their proposed hybrid approach combines an iterative regression strategy with an extreme-value strategy. The extreme-value strategy is a gene-based method whereby a gene is identified as a candidate if at least one individual with extreme values ($>95\%$ quantile) has at least one rare variant within the gene. This strategy addresses the multiple testing problem by testing only candidate genes. The iterative regression involves a variant-based multimarker score test aimed at identifying a group of significant variants. The hybrid approach combining both strategies was shown to outperform the reference method—the combined multivariate and collapsing (CMC) method [Li and Leal, 2008]—with regard to power. However, work remains to be done to determine the optimal value of the number of SNPs to include in step 1 of the iterative procedure, because too small a number will exclude causal SNPs but too large a number will also cause power loss by the addition of noise terms. More important, the validity of approaches to utilize extreme trait values to increase power depends on the degree of truth in the assumption that rare variants with the largest effect drive the quantitative trait values to the tail. Depending on the trait, one may run the risk of missing causal variants that increase the variance, but not the mean value of the trait, by targeting only variants associated with the extreme trait values. Zhang and colleagues also showed that a small number of phenotypic outlier individuals caused inflated type I errors in Q1. This problem was resolved after transforming the data, highlighting the importance of removing or adjusting for outliers in these types of analyses.

The papers discussed in this section used information about extreme phenotype values in the sampling design, analysis stage, or candidate gene selection process. Overall, these work groups showed that it was useful to incorporate some kind of extreme-value approach for the detection of rare variants if appropriate statistical methods were used. The question remains of how extreme values should be defined. Depending on the phenotype of interest, one must balance the informative value of restricting samples to the far extremes with the ability to obtain an adequate sample size.

META-ANALYSIS DESIGN ADJUSTED FOR POPULATION SUBSTRUCTURE

Additional questions concerning study design for rare variants were addressed by Culverhouse et al. [2011]. The first question examined whether a larger sample of unrelated individuals would have given more power to detect rare variants. The second question was whether it was more powerful to analyze the complete set of unrelated individuals (a sample that was made up of distinct population-specific subsamples) while adjusting for population structure through the use of covariates or to analyze each subpopulation separately. When analyzing the complete sample for Q1, only *FLT1* was significant at $P<1\times10^{-6}$ in all replicates, and no other causal gene was significant in at least 50% of replicates. As sample size was increased using meta-analysis of 10 and all 200 replicates, Culverhouse and colleagues detected some additional causal loci, thus showing that increased sample size can detect some of the variants with lower locus-specific heritability. However, when subpopulations were analyzed (combining 10 replicates by meta-analysis), Culverhouse and colleagues detected additional causal loci in the subpopulations. *FLT1* contained multiple rare variants with relatively large effect on Q1. Because some causal variants were present in all subpopulations except the Luhya (representing more than 84% of the data), pooling the data and using Population as a covariate maximized power. However, the single causal variant for Q1 in *VEGFA* was a private mutation in the Luhya population. Meta-analysis of the Luhya subjects alone (total $N = 5{,}400$) resulted in extremely significant association ($P = 2.1 \times 10^{-94}$), whereas it required meta-analysis of 50 replicates of the full data (total $N = 34{,}850$) before this gene surpassed the 10^{-6} significance threshold ($P = 1.4 \times 10^{-14}$). These results show that including population as a covariate is not always an effective substitute for analyzing the subpopulations separately and suggest that population-specific analyses may help to detect genes with causal variants private to a single population, as has been suggested by Keen-Kim et al. [2006].

MACHINE LEARNING METHODS FOR RARE VARIANT ANALYSIS

It has been suggested that machine learning methods may be particularly powerful when epistatic interactions exist between predictors [Dasgupta et al., 2011]. In the Group 14 contributions, three machine learning methods were evaluated.

MACHINE LEARNING METHODS TO SCREEN FOR CAUSAL VARIANTS

Kim et al. [2011] applied random forests [Breiman, 2001] and logic regression [Ruczinski et al., 2004] to Q2 in the 321 members of the Asian subpopulation (to avoid issues of population substructure) to screen for loci that contributed to variation in this trait. Kim and colleagues selected a subset of predictors that they had determined made the strongest contributions to the final models and evaluated the performance of each method by assessing how often the true causal versus noncausal variants were included in the top-ranked subset of predictors across 200 replicates. The same metric was also applied to traditional single-marker analysis results from simple linear regression for comparison purposes. They compared analyses of individual SNP genotypes using dominant coding (only in random forests) or additive coding (only in simple linear regression), using indicator variables generated by collapsing rare variants within genes or a pathway [Li and Leal, 2008], and recoding common variants using dominant and recessive coding.

When rare variants were collapsed within genes, random forests and logic regression outperformed simple linear regression when the percentage of causal variants ranked in the top 10% of predictors across at least 10 replicates was compared. Random forests ranked the same causal gene among the top 10% of predictors in up to 40% of replicates, and logic regression selected the same causal gene among the 10 predictors in the final logic regression model in up to 53% of replicates. Random forests showed similar performance, although at lower levels of replication, for analysis of uncollapsed SNP variants and of collapsed SNPs by pathway variants. However, for logic regression, Kim et al. [2011] observed elevated replication of noncausal genes in the top-ranked predictors; they showed by small-scale permutation tests that logic regression had little potential to select the true causal variants over what was expected by chance in these data.

Kim et al. [2011] found that all three methods selected larger proportions of causal variants than noncausal variants in their top-ranked predictors (random forests and simple linear regression) or in the final logic regression models across multiple replicates, indicating statistical validity. However, for all three methods, when analyzing the uncollapsed and gene-collapsed variants, most causal variants were not ranked in the top 10% of predictors in large proportions of replicates. This low power is most likely due to the simulation model for Q2 in the Asian population, which was essentially polygenic, with small locus-specific heritabilities of any causal variants and no epistatic interactions. One causal variant had an average (across 200 replicates) locus-specific h^2 [Falconer and Mackay, 1996] of 0.04, five causal variants had locus-specific h^2 between 0.011 and 0.017, and the remaining causal variants had locus-specific $h^2 < 0.01$. Thus there was little power to detect causal variants in this small Asian sample using any statistical method and, as expected, the loci with higher locus-specific h^2 were included in the top predictors in the largest numbers of replicates. These results suggest that for polygenic models with small locus-specific h^2, such as the one simulated here, (1) larger samples would be desirable and (2) in a similarly small study, at least the top 5–10% of predictors should be used in follow-up studies to give a reasonably high probability that at least one causal locus is selected.

BAYESIAN NETWORK STRUCTURE LEARNING TO DETECT CAUSAL SNPS IN CANDIDATE GENES

Schlosberg et al. [2011] applied Bayesian network structure learning (BNSL) [Needham et al., 2007] for a different purpose. Rather than attempting to screen all variants to detect those that were most likely to contribute to the etiology of a trait, they focused on the situation in which target genes that harbor causal variants for the disease trait Affected have already been chosen for resequencing; the goal was to detect true causal SNPs from among the measured variants in these candidate genes. Examining all the available SNPs in the known causal genes, BNSL produced a Bayesian network from which two subsets (the Markov blanket and the descendants of Affected) of SNPs were extracted and then measured for statistical significance using the hypergeometric distribution. As applied to these data, overall the method did not demonstrate the ability to detect more causal SNPs than would be selected by chance, although Schlosberg and colleagues' analyses did suggest that improved performance for this highly polygenic trait could be attained in larger sample sizes. Their analyses of ethnic subgroups also suggests that BNSL may be a valuable strategy for real data if different causal variants exist across subpopulations.

DISCUSSION

The traits simulated for GAW17 were extremely complex, exhibiting both locus and allelic heterogeneity, wide variation in locus-specific allele frequency and effect size, and private causal variants in certain subpopulations. Many of the lessons learned from the analyses of these data with novel methods are reiterations of classic results in statistical genetics. Power and type I error depend on sample size, effect size, degree of heterogeneity, and data characteristics that violate assumptions of the statistical methods. Readers should keep the simulation model in mind as we evaluate the performance of each analysis method.

The simulated traits were highly polygenic with different distributions of rare variants across subpopulations and small genetic effects of individual causal variants. The observed results were highly dependent on locus-specific heritabilities of the causal variants. For quantitative trait Q1, there were 39 causal SNPs in 9 genes, with mean SNP locus-specific h^2, averaged over 200 replicates of the complete data, ranging from 0.0009 to 0.152 [Wilson and Ziegler, 2011]. Only 11 and 5 of the Q1 causal SNPs exhibited locus-specific h^2 greater than 1% and 3%, respectively. The five SNPs with the largest effect size were located in the *FLT1* gene ($h^2 = 0.152, 0.083, 0.037$) and the *KDR* gene ($h^2 = 0.031$, 0.031), For Q2, there were 72 causal SNPs in 13 genes, with SNP locus-specific heritability ranging from 0.00098 to 0.016 [Wilson and Ziegler, 2011], with only 2 SNPs (one in *VNN1* and one in *VNN3*) having locus-specific h^2 greater than 1%. The Affected status trait is due to a combination of signals from Q1, Q2, Q4, and a latent disease liability. No simulated causal SNP is expected to have a large effect on risk of affection.

RESULTS CONSISTENT WITH LOCUS EFFECT SIZE

The Affected status trait was analyzed by several work groups comparing different collapsing methods. Multiple methods consistently detected causal effects for a few causal genes, such as *FLT1* and *KDR*, but with different observed power in the different methods. For example, partial least-squares identified *FLT1* as associated with the disease trait in 35 out of the 200 replicates, whereas inverse-probability weighted clustering detected *FLT1* and *KDR* in more than 50% of the simulated replicates. These two genes are causal for Q1 and harbor causal variants with the highest locus-specific h^2 in all the GAW17 simulated data. Indeed, most analyses of Q1 tended to detect *FLT1* with adequate power in all replicates at stringent significance thresholds. Unlike Q1, causal SNPs for Q2 and those that are specific to Affected status had much smaller effect sizes and were harder to map. Overall, one can conclude that power is dependent on locus-specific heritability, which is dependent on both allele frequency in the population being studied and on the effect of the locus on variation of the trait or risk of the disease (a classic result that applies to the novel methods presented here).

INCREASED POWER BY COLLAPSING RARE VARIANTS AND STRATEGIC SAMPLING

As expected, collapsing rare variants either within genes or within pathways increased the power to detect causal variants given the extreme allelic heterogeneity simulated for these traits. Also as expected, increasing sample sizes did increase power to detect loci with smaller locus-specific h^2, although extremely large samples will be required to detect effects of rare variants with small effect sizes. In GAW17 all rare variants have positive effects. This may not be the case in real association studies. It is worth applying the collapsing methods utilized in GAW17 to real sequence data and/or to more complex simulated data sets and evaluating the performance therein.

A machine learning method, random forests, performed as well as or somewhat better than simple linear regression when analyzing collapsed variants for selecting a fairly large subset of predictors for follow-up in additional studies in the extremely polygenic Q2 trait. Additional evaluation of such methods in the presence of epistatic interactions would be useful.

Different methods for collapsing or aggregating variants within a candidate locus were most powerful for different loci, especially when the genetic signals were not strong. This suggests that a multiple method approach may be valuable for future association studies of DNA sequence data and complex traits.

Furthermore, most of the proposed methods have fewer underlying assumptions than existing ones [Li and Leal, 2008] and show promise for detecting at least some loci in realistic situations. Different study designs may be powerful to detect different types of loci. In particular, sampling designs can affect power where sampling from extremes can be beneficial if analyzed properly by the various collapsing methods. Analyses from Group 14 showed that even with the identical genetic model applied to multiple subpopulations, sample size was not the only factor determining power. If rare causal variants are private to a subpopulation, then stratified analysis may be more powerful than a combined analysis, despite a considerable decrease in sample size. Thus studies conducted in different populations may be required to detect many rare variants. The complex models utilized in GAW17 show clearly that small samples will have low power to detect loci with small effects on a trait and that even large samples may not be adequate to reliably detect rare variants with extremely small effects on variation of the trait or risk for a disease.

POTENTIAL CULPRITS FOR TYPE I ERROR

An interesting observation is the phenomenon that was labeled consistent false positives during the GAW17 conference. Specifically, a few noncausal genes were identified by more than one method consistently. Possible reasons for this, based on the analysis of different groups, include (1) population stratification; (2) the detection in multiple replicates of noncausal SNPs that are strongly associated with causal variants, because the same genotypes were used in all of the 200 replications; (3) outliers and genetic disequilibrium. Given the small sample size and the large number of SNPs, the contributors to GAW17 observed that a set of noncausal SNPs were perfectly correlated with some of the rare causal variants, which thus led to some of the false-positive results. When applying the various collapsing methods to rare variant sequence data, as a precaution, one should take additional steps to address the aforementioned issues. For example, whenever population information is available, existing methods such as EigenSoft [Price et al., 2006] or PLINK [Purcell et al., 2007] can be used to adjust for population stratification. Removing outliers or applying appropriate transformations might also help to improve the performance of the proposed methods. The GAW17 results also suggest that evaluation of the correlation structure among associated loci is a reasonable strategy when interpreting the results of analyses of DNA sequence data, particularly in small samples. If strong correlations are observed among multiple rare associated variants that are distant (or on separate chromosomes), this may indicate that several variants are all highly correlated with a private causal variant in a single individual in the sample.

CONCLUSIONS

As expected, the different methods used by Group 14 contributors showed variation in the causal loci detected. Different research groups studied different phenotypes and used different thresholds for declaring significance. However, some consistent findings are that (1) power and type I error depend on sample size, effect size, degree of heterogeneity, and classic causes of errors; (2) analyses of very rare variants requires some form of grouping; and (3) rare variants introduce novel sources of error as a result of co-occurrence of private mutations within single individuals in small samples, suggesting that large samples of sequenced unrelated individuals will be needed to avoid high false-positive associations of rare variants with disease.

ACKNOWLEDGMENTS

The Genetic Analysis Workshops are supported by National Institutes of Health (NIH) grant R01 GM031575. This work was supported in part by the Intramural Research Program of the National Human Genome Research Institute, NIH grant T32 MH-14235 and R21 DA033827.

REFERENCES

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

Breheny P, Huang J. 2009. Penalized methods for bi-level variable selection. Stat Interface 2:369–380.

Breiman L. 2001. Random forests. Mach Learn 45:5–32.

Brennan J, He Y, Calixte R, Nyirabahizi E, Jiang Y, Zhang H. 2011. A LASSO-based approach to analyzing rare variants in genetic association studies. BMC Proc 5:S100.

Chernoff H, Lo S, Zheng T. 2009. Discovering influential variables: a method of partitions. Ann Appl Stat 3:1335–1369.

Culverhouse R, Hinrichs A, Suarez BK. 2011. Stratify or adjust? Dealing with multiple populations when evaluating rare variants. BMC Proc 5:S101.

Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. 2011. Brief review of machine learning methods in genetic epidemiology: the Genetic Analysis Workshop 17 experience. Genet Epidemiol, this issue.

Dering C, Hemmelmann C, Pugh E, Ziegler A. 2011. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol, this issue.

Falconer D, Mackay T. 1996. Introduction to Quantitative Genetics. Harlow, England: Pearson/Prentice Hall.

Jiang Y, Brennan J, Calixte R, He Y, Nyirabahizi E, Zhang H. 2011. Novel tree-based method to generate markers from rare variant data. BMC Proc 5:S102.

Jung J, Dantzer J, Liu Y. 2011. Identification of multiple rare variants associated with a disease. BMC Proc 5:S103.

Keen-Kim D, Mathews CA, Reus VI, Lowe TL, Herrera LD, Budman CL, Gross-Tsur V, Pulver AE, Bruun RD, Erenberg G, Naarden A, Sabatti C, Freimer NB. 2006. Overrepresentation of rare variants in a specific ethnic group may confuse interpretation of association analyses. Hum Mol Genet 15:3324–3328.

Kim Y, Li Q, Cropp C, Sung H, Cai J, Simpson C, Perry B, Dasgupta A, Malley J, Wilson A, Bailey-Wilson JE. 2011. Performance of random forests and logic regression methods using mini-exome sequence data. BMC Proc 5:S104.

Lamina C. 2011. Digging into the extremes: a useful approach for the analysis of rare variants with continuous traits? BMC Proc 5:S105.

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 83:311–321.

Liu Y, Huang CH, Hu I, Lo S-H, Zheng T. 2011. Association screening for genes with multiple potentially rare variants: an inverse-probability weighted clustering approach. BMC Proc 5:S106.

Mägi R, Kumar A, Morris A. 2011. Assessing the impact of missing genotype data in rare variant association analysis. BMC Proc 5:S107.

Morris AP, Zeggini E. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34:188–193.

Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. 2007. A primer on learning in Bayesian networks for computational biology. PLoS Comput Biol 3:e129.

Niu YS, Hao N, An L. 2011. Detection of rare functional variants using group ISIS. BMC Proc 5:S108.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575.

Ruczinski I, Kooperberg C, LeBlanc M. 2004. Exploring interactions in high dimensional genomic data: an overview of logic regression, with applications. J Multivar Anal 90:178–195.

Schlosberg CE, Schwantes-An T-H, Duan W, Saccone N. 2011. Application of Bayesian network structure learning to identify causal variant SNPs from resequencing data. BMC Proc 5:S109.

Wang L, Pungpapong V, Lin Y, Zhang M, Zhang D. 2011. Genome-wide case-control study in GAW17 using coalesced rare variants. BMC Proc 5:S110.

Wilson AF, Ziegler A. 2011. Lessons learned from the Genetic Analysis Workshop 17: transitioning from genome-wide association studies to whole-genome statistical genetic analysis. Genet Epidemiol, this issue.

Yilmaz YE, Bull SB. 2011. Are quantitative trait-dependent sampling designs cost-effective for analysis of rare and common variants? BMC Proc 5:S111.

Zhang D, Lin Y, Zhang M. 2009. Penalized orthogonal-components regression for large *p*, small *n* data. Electron J Stat 3:781–796.

Zhang Z, Sha Q, Wang X, Zhang S. 2011. Detection of rare variant effects in association studies: extreme values, iterative regression, and a hybrid approach. BMC Proc 5:S112.

Zheng T, Chernoff H, Hu I, Ionita-Laza I, Lo S. 2010. Discovering influential variables: a general computer intensive method for common genetic disorders. In: Lu H, Scholkopf B, Zhao H, editors. Handbook of Computational Statistics: Statistical Bioinformatics. New York: Springer.