Published in final edited form as:
*Genet Epidemiol*. 2011 ; 35(0 1): S29–S34. doi:10.1002/gepi.20646.

# Incorporating Biological Information into Association Studies of Sequencing Data

**Gary Chen**[1,*], **Peng Wei**[2,*], and **Anita L. DeStefano**[3]

[1]Division of Biostatistics, Department of Preventive Medicine, University of Southern California, Los Angeles, CA

[2]Division of Biostatistics and Human Genetics Center, University of Texas School of Public Health, Houston, TX

[3]Department of Biostatistics, School of Public Health, Boston University, Boston, MA; and Department of Neurology, Boston University School of Medicine, Boston, MA

## Abstract

We summarize the methodological contributions from Group 3 of Genetic Analysis Workshop 17 (GAW17). The overarching goal of these methods was the evaluation and enhancement of state-of-the-art approaches in integration of biological knowledge into association studies of rare variants. We found that methods loosely fell into three major categories: (1) hypothesis testing of index scores based on aggregating rare variants at the gene level, (2) variable selection techniques that incorporate biological prior information, and (3) novel approaches that integrate external (i.e., not provided by GAW17) prior information, such as pathway and single-nucleotide polymorphism (SNP) annotations. Commonalities among the findings from these contributions are that gene-based analysis of rare variants is advantageous to single-SNP analysis and that the minor allele frequency threshold to identify rare variants may influence power and thus needs to be carefully considered. A consistent increase in power was also identified by considering only nonsynonymous SNPs in the analyses. Overall, we found that no single method had an appreciable advantage over the other methods. However, methods that carried out sensitivity analyses by comparing biologically informative to noninformative prior probabilities demonstrated that integrating biological knowledge into statistical analyses always, at the least, enabled subtle improvements in the performance of any statistical method applied to these simulated data. Although these statistical improvements reflect the simulation model assumed for GAW17, our hope is that the simulation models provide a reasonable representation of the underlying biology and that these methods can thus be of utility in real data.

### Keywords

exome sequencing; pathway analysis; gene association

## Introduction

Genome-wide association studies (GWAS) that use data from single-nucleotide polymorphism (SNP) chips have been successful in identifying some genes or genomic regions that influence disease susceptibility or quantitative trait loci (QTLs). However,

**Corresponding Author:** Anita DeStefano, Boston University School of Public Health, 801 Massachusetts Avenue, 3rd Floor, Boston, MA 02118, adestef@bu.edu, Phone: 617-638-4057.
*These authors contributed equally to this work

GWAS using SNP chips are limited to analysis of relatively common variants and do not necessarily identify or even examine causal polymorphisms. Biological function is not a criterion for SNP selection on chips. Exome sequencing provides data that characterize all variants within the exome of each gene, enabling one to specifically seek functional variants with biological effects. Incorporating biological knowledge into association analysis of exome data may increase the power to detect such variants.

The mini-exome data of Genetic Analysis Workshop 17 (GAW17) present the challenge of trying to identify relevant genetic variants in a data set that includes both common and rare variants. A goal of many GAW17 participants was to try to overcome the statistical limitations that make it difficult to analyze rare variants. The GAW17 contributors to Group 3, reviewed in this summary paper, approached the challenge of rare variants by incorporating prior biological information into their analyses.

The Group 3 contributors applied a variety of approaches with a wide range in the definition of biological information. For some approaches, considering the gene rather than the SNP as the unit of analysis was the biological information incorporated. In other approaches the functionality of variants or the functionality of genes or gene sets was characterized. An overview of the contributions is given in Table I.

We organize this paper based on the broad themes identified in the analyses performed by the GAW17 participants in Group 3. These themes are (1) gene-based association analyses including rare variants, (2) gene-based approaches for variable selection, and (3) association analyses incorporating SNP or gene functionality. The first two themes include methods that use SNP location within or near a gene as part of the method. The only other biological information used in these approaches is whether or not the SNPs are nonsynonymous. The methods in the third category include more detailed biological information, such as the predicted functionality of SNPs and gene pathway information.

## Methods

### Gene-Based Association Analyses Including Rare Variants

A common approach to association analysis of rare variants is to collapse or combine rare variant data within individuals across or within genes [Dering et al., 2011]. Fan et al. [2011] compare three such approaches that are appropriate for dichotomous traits: the gene-based combined multivariate and collapsing (CMC) method of Li and Leal [2008], the weighted-sum (WS) method of Madsen and Browning [2009], and the replication-based (RB) approach of Ionita-Laza et al. [2011] in the set of unrelated subjects. The CMC and WS approaches are among the first collapsing methods for rare variants and are described in detail by Dering et al. [2011]. The more recent RB approach, which aggregates across individuals, is summarized here.

The RB approach is based on partitioning observed variants into two disjoint classes (likely to be a risk or protective) using a weighting scheme that reflects the difference in observed frequencies between case subjects and control subjects. The RB method addresses a limitation of the other methods by enabling testing across both risk and protective alleles. Let $S_i$ denote the test statistic where $i = R$ for assessing risk alleles and $i = P$ when assessing protective alleles. Thus

$$S_i = \sum_{k=0}^{N_r} \sum_{k'>k} -n_k^{k'} \log[p(k,k')], \quad (1)$$

where, when $i = R$ for a specific variant, $k$ is the number of occurrences of the minor allele in control subjects and $k'$ is the number of occurrences of the minor allele in case subjects. When $i = P$ for a specific variant, $k$ is the number of occurrences of the minor allele in case subjects and $k'$ is the number of occurrences of the minor allele in control subjects. The probability $p(k, k')$ is computed as described by Ionita-Laza et al. [2011]. The overall test statistic is $S = \max(S_R, S_P)$.

Fan et al. [2011] applied the CMC, WS, and RB approaches to the binary affection trait and dichotomized variables for Q1, Q2, and Q4. The dichotomized variables were generated by applying an extreme sampling design in which individuals in the upper quartile were considered case subjects and those in the lower quartile were considered control subjects. Fan and colleagues also applied an extension of the WS approach to quantitative traits [Price et al., 2010]. Analyses were applied using all SNPs that met a minor allele frequency (MAF) threshold criterion and were limited to nonsynonymous SNPs meeting the MAF criterion. Associated genes were determined using the first replicate, and then the remaining replicates were used to determine replicability of these findings.

In contrast, Yip et al. [2011] took advantage of the pedigree data provided by GAW17 and explored linkage and family-based association approaches. They applied multiple nonparametric methods, including goodness-of-fit, mean, and trend tests using affected sib pairs and Haseman-Elston and modified Haseman-Elston tests using all sib pairs. The Haseman-Elston approach was also applied to Q1. The linkage methods did not make use of biological information. Fine mapping of regions of interest identified by means of linkage analysis was conducted using the family-based association test (FBAT). A novel contribution was a modified FBAT in which the rare variants within offspring in a nuclear family were summed across a gene. To account for correlation among SNPs, Yip and colleagues used two versions of an empirical variance: one at the pedigree level and one at the family level.

Both Fan et al. [2011] and Yip et al. [2011] discuss the issue of MAF threshold definition for rare variants. Fan and co-workers considered 1% and 5%, whereas Yip and colleagues considered 1% and 3%.

## Gene-Based Approaches for Variable Selection

A first approach to gene-based variable selection is the two-step gene-based partial least-squares (GB-PLS) technique proposed by Turkmen and Lin [2011]. In the first step SNPs within a gene are related to a latent variable through a projection obtained by a partial least-squares method analogous to principal components analysis, providing an estimate of the "outer coefficients." To estimate the "inner coefficients," Turkmen and Lin then relate the latent variables, along with any nongenetic covariates, to a trait of interest. They present two subtle variations of the method: GB-PLS1 and GB-PLS2. To implement GB-PLS1, Turkmen and Lin first apply an established method, called partial least-squares path modeling (PLSPM), followed by an ordinary least-squares regression of the trait on the latent variables and nongenetic covariates in order to solve the outer and inner coefficients, respectively. The GB-PLS2 variant of their method is geared toward models of higher dimensions and uses the statistical machinery of partial least-squares and least absolute shrinkage and selection operator (LASSO) regression in place of PLSPM and ordinary regression.

Chen [2011] presents a Bayesian Markov chain Monte Carlo (MCMC) algorithm that is designed to stochastically sample model variables (e.g., SNPs or other covariates) from a posterior distribution for the purposes of variable selection. To summarize, the statistical model used in defining the posterior density function assumes that the regression

coefficients from a first-level regression (where a trait of interest is regressed on a random set of model variables) arise from a prior distribution whose means and variances are empirically estimated from a second-level regression (where regression coefficients from the first level are regressed on a user-defined design matrix called $Z$). Chen proceeds with an analysis in which the mutation class (i.e., synonymous versus nonsynonymous) is stored in $Z$ as an indicator variable, contributing to the prior means, and adjacency information (whether a set of SNPs are members of the same gene) is stored in a matrix called $A$, contributing to the prior variances.

Another novel approach [H. Zhou, personal communication, 2010] is a variation on the LASSO method that appends gene-specific Euclidean, or L2 norm, penalties (the square root of a sum of the squares of β coefficients) on top of the global standard L1 norm penalty (the sum of the absolute value of β coefficients) to the loss function for ordinary or logistic regression [Zhou et al., 2010]. In other words, the likelihood is reduced to a greater extent for large values of either the L1 or the L2 norm. Zhou et al. [2010] uses the L1 norm to encourage sparse models so that most coefficients stay parked at 0. Once a common, easily detectable SNP within a gene enters the model, the L2 penalty relaxes for that gene, allowing correlated rare SNPs of weak effect within that gene to enter the model. Zhou and colleagues' method is more appealing than some of the burden-style collapsing methods described earlier because the method is agnostic to the sign of the coefficients, so that for any given effect size, both protective and risk SNPs that are rare can be detected with equal probability. The method is implemented in the software Mendel, which is publicly available at http://www.genetics.ucla.edu/software [Zhou et al., 2011].

All the variable selection methods were applied to traits Q1 and Q2 and evaluated across the 200 replicates.

## Incorporating SNP or Gene Functionality into Association Analyses

The contributions within this theme expand the type of a priori biological information used to include (1) predicted deleterious effects of nonsynonymous variants at the SNP level and (2) curated gene sets (e.g., MSigDB), canonical pathways (e.g., Ingenuity Pathway Analysis [IPA] and Kyoto Encyclopedia of Genes and Genomes [KEGG]), and gene networks (e.g., protein-protein-interaction [PPI] networks) at the gene level.

Wei et al. [2011] compared weighting schemes based on the functional predictions of nonsynonymous SNPs using two popular algorithms (SIFT [Kumar et al., 2009] and PolyPhen-2 [Adzhubei et al., 2010]) in the gene-based variable threshold (VT) association test of Price et al. [2010]. Four versions of the VT test, varying by the weights assigned to the SNPs within a gene, are applied to the binary trait Affected: (1) grouping all SNPs with equal weights, (2) grouping only nonsynonymous SNPs with equal weights, (3) using predicted damaging scores by SIFT as weights, and (4) using predicted damaging scores by PolyPhen-2 as weights. A two-component normal mixture model is also proposed to combine the potentially heterogeneous test results based on different functional prediction algorithms.

Ngwa et al. [2011] propose using gene set information to identify disease-associated pathways, aiming at combining weak to moderate SNP-level association signals. Pathway-based association analysis of sequencing data entails two steps: SNP- or gene-level association analysis followed by pathway enrichment analysis. Ngwa and colleagues studied and compared three pathway analysis methods: gene set enrichment analysis (GSEA), empirical enrichment analysis, and IPA. The first step is the same for all methods: single-SNP regression-based association testing for common SNPs. For rare SNPs (MAF < 1%) an aggregate rare variant is defined as the count of rare alleles within a gene, and regression-

based association is performed on the rare variant. Each gene is then assigned the most significant association *p*-value, either from a single common SNP or from the aggregate rare variant from that gene. In the subsequent pathway-level analysis, the GSEA uses a running sum statistic to test the overrepresentation of each set's genes at the top of a ranked gene list by their *p*-value, whereas IPA uses Fisher's exact test to identify canonical pathways that are significantly enriched with genes with association *p*-values less than 0.01. Finally, the empirical enrichment analysis uses a permutation test to determine whether a gene set contains a higher proportion of genes with low association *p*-values (<0.01) than would be expected by chance. Ngwa and colleagues apply the three methods to traits Q1 and Q4 to evaluate the power to detect the vascular endothelial growth factor (VEGF) pathway (the causal gene set for Q1 in the simulation model) and the type I error, respectively.

After understanding the role of the VEGF pathway in the simulating model, Fan et al. [2011] also implemented the four rare variant methods (CMC, WS, RB, and VT) with a pathway focus by aggregating across all the genes within the VEGF pathway, rather than aggregating within each gene.

Another pathway approach is an interactome (i.e., the landscape of all possible interacting proteins) analysis to identify densely connected genes in a gene network that is overrepresented by disease-associated genes [R. Massanet-Villa, personal communication, 2010]. First, SNP-level association tests are performed, and a subnetwork connecting all the genes that contain significant SNPs is built based on the BioGRID PPI network, which consists of 30,887 interactions involving 9,379 proteins. Second, spectral clustering is applied to partition the subnetwork into densely connected clusters, followed by Gene Ontology (GO) enrichment analysis to characterize each cluster.

## Results

### Gene-Based Association Analyses Including Rare Variants

Comparison of the CMC, WS, and RB methods applied to affection status and the dichotomized quantitative traits found high correlation across the approaches, with no one of these approaches showing an advantage in these data for the gene-based approach [Fan et al., 2011]. Six true disease genes for Q1, five for Q2, and three for Affected status were detected in the top 50 most significant genes using the first simulated data set. Many of the disease genes were highly replicable in 200 simulations with power greater than 50%, indicating a potentially true signal. However, there were numerous consistent false-positive genes. For example, among the top genes identified in the first simulation, *GOLGA1* for Affected status, *PPP1R14BP1* for Q1, and *MAP3K8* for Q2 were additionally identified in more than 50% of the remaining replicates.

The power to detect truly associated genes was increased when only nonsynonymous SNPs were analyzed. Power was also higher when examining Q1 and Q2 as dichotomous variables generated on the basis of sampling from the extremes design as opposed to applying the WS approach to the continuous measures of all individuals [Fan et al., 2011].

Yip et al. [2011] found wide variability in the results from different linkage methods. The linkage methods that aggregated results across pedigrees failed to identify any of the causal genes among the top candidates. In contrast, causal genes were identified when genetic heterogeneity was considered in a pedigree-stratified analysis. Fine mapping of linkage regions by FBAT identified an improvement in the ability to identify causal variants when Yip and colleagues' gene-based modified FBAT was applied to the empirical variance at the family level. Further research is required to identify the best variance adjustment.

Both Fan et al. [2011] and Yip et al. [2011] considered different MAF thresholds and consistently found higher power when considering a less stringent definition of the MAF. This difference was more pronounced for specific genes, such as *FLT1*.

## Gene-Based Approaches for Variable Selection

Turkmen and Lin [2011] did not find appreciable differences in performance between the two variations of the GB-PLS analysis. However, they did successfully demonstrate that by aggregating association evidence across SNPs within a gene, their approach was more powerful than if they had ignored joint information within genes in their analyses.

The results from a sensitivity analysis for the Bayesian MCMC approach in which varying levels of biological information were compared indicate that incorporating informative prior distributions through $Z$ and $A$ provides better statistical properties in terms of power across a wide range of false-positive rates than incorporating noninformative prior distributions does. In support of Turkmen and Lin's [2011] postulation that gene-level information is helpful, Chen [2011] demonstrated that the $A$ matrix allowed SNPs within a gene to distribute evidence for association among themselves, a desirable property when one considers that some rare variants can potentially increase their probability of being detected by borrowing evidence from nearby common variants with moderate effect sizes.

Like Chen [2011], H. Zhou [personal communication, 2010] also performed sensitivity analyses across different prior specifications, generating receiver operating characteristic (ROC) plots that illustrated power across different type I error rates. The best model was the standard approach that Zhou proposed, which was to include both the L1 and the L2 norms in the penalty. The next best models in descending order were the L2-norms-only model, the L1-norm-only model, and, the worse performing model, a conventional SNP-by-SNP univariate test.

## Incorporating SNP or Gene Functionality into Association Analyses

The SIFT and PolyPhen-2 prediction scores were found to be overall positively correlated, and the qualitative predictions of being damaging were in high concordance (odds ratio = 5; $p < 10^{-16}$). Although concordance was high, it was not perfect; hence the two algorithms' predictions disagreed with each other on a number of SNPs. As a result, the SIFT-based and the PolyPhen-2-based VT test results could also differ; for example, the top 100 genes identified by each method shared about 70% of genes in common. On the other hand, the SIFT-based VT test and the test incorporating only synonymous or nonsynonymous information shared 85% of genes among their top 100 gene lists. The proposed mixture model was found to be effective in combining heterogeneous test results. In particular, some genes with moderately small $p$-values from both SIFT-based and PolyPhen-2-based VT tests could be boosted to have higher ranks by the mixture-model-based combined analysis.

GSEA and IPA were found to have high power for detecting the VEGF pathway (91.2% and 93.0%, respectively) and a slightly deflated type I error rate, whereas the empirical enrichment method had a lower power (42.9%) and maintained the nominal type I error rate.

In their post hoc analysis of the VEGF pathway, Fan et al. [2011] collapsed rare SNPs across the genes in the pathway and compared tests of association. They found that the WS and RB approaches performed substantially better than the CMC method, probably because of the increased power of WS approaches for large genomic regions. Consistent with their findings for gene-based analysis, association was more significant when only nonsynonymous SNPs were included.

In the interactome analysis of Massanet-Villa [personal communication, 2010], only SNPs with MAF > 0.02 were considered, leading to a large number of true disease-associated genes being absent in the subsequent network-based analysis. Using more sophisticated gene-based association tests involving rare variants might improve the power of the interactome analysis and reduce the false-positive rate.

## Discussion

The contributors to Group 3 at GAW17 had different goals in incorporating biological information into their analyses of the mini-exome data. These goals included identifying single genes associated with disease, identifying pathways associated with disease, and selecting variables (genetic variants) that explained the variability in trait outcomes. Hence the approaches varied widely. The contributors also considered different types of biological information, including SNP location within a gene, whether a SNP was synonymous or nonsynonymous, the predicted functionality of nonsynonymous SNPs, and gene function as it defined participation in a gene or PPI pathway.

Despite the diversity of approaches, we identified three commonalities that indicate that it might be advantageous to incorporate biological information. The first of these is that identifying which SNPs map near or within a specific gene is important. This knowledge is essential for methods that use such information for aggregating or collapsing rare variants or for defining prior probabilities in the context of a Bayesian analysis. The Group 3 contributors provided new methods for incorporating this information in family-based association and variable selection approaches, such as the modified FBAT, GB-PLS, and Bayesian MCMC methods. Among the approaches explored by this group, we could identify no substantial advantage for a specific rare variant association method aggregating across a gene. The single-gene approaches have been examined in more detail elsewhere [Sun et al., 2011; Tintle et al., 2011].

As a second commonality, these contributors identified the importance of selecting the correct threshold for the MAF in defining rare variants. Many contributors identified differences in power with different thresholds. These differences might be driven by the simulation model, which includes causal variants with MAFs between 1% and 5%, resulting in a loss of power with a 1% threshold. Pan and Shen [2011] recently proposed a class of adaptive tests, including the VT method of Price et al. [2010] as a special case, to allow the MAF threshold for rare variants to be data adaptive. In light of the disadvantage of arbitrarily selecting a fixed threshold, the adaptive tests are appealing.

The third commonality is that several contributors in this group identified increased power when the analysis was restricted to only nonsynonymous SNPs, indicating the importance of differentiating synonymous from nonsynonymous SNPs. This finding may also be driven by the simulation model, in which all causal SNPs were nonsynonymous. However, the simulation model was designed to reflect biological realities, and hence this strategy may increase the power to identify causal genes in real data. Recently proposed association tests for rare variants [Pan and Shen, 2011; Wu et al., 2011] have demonstrated that using informative but not necessarily perfect external information, such as predicted functions of SNPs, in aggregating methods may substantially improve the statistical power.

The contributors to Group 3 also highlight the diversity in sources of biological information and the ongoing process of biological discovery and understanding, for example, the differences in predicted functionality of SNPs from SIFT versus PolyPhen and the differences in pathways defined in MSigDB versus KEGG versus IPA. Wei et al. [2011] demonstrated that methods that combine different sources of information may be useful.

In summary, the Group 3 contributions to GAW17 indicate that (1) gene-based analysis of rare variants is better than single-SNP analysis but that (2) the MAF threshold to identify rare variants may influence power and (3) power may be increased by considering only nonsynonymous SNPs. No single method held an appreciable advantage over the other methods. However, methods that carried out sensitivity analyses by comparing biologically informative to noninformative prior distributions demonstrated that integration of biological knowledge into statistical analyses can always, at the least, enable subtle improvements in the performance (i.e., true- and false-positive rates) of any statistical method under the GAW17 simulating model.

## Acknowledgments

The Genetic Analysis Workshops are supported by National Institutes of Health (NIH) grant R01 GM031575. Support was also provided by NIH grants R01 HL106034 (PW), R01 AG031287-02 (ALD), R01 AG033193-03 (ALD), R01 ES016813 (GC), and R01 ES015090 (GC) and by a PRIME grant from the University of Texas School of Public Health (PW).

## References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Meth. 2010; 2010; 7:248–249.

Chen GK. Enhancing the discovery of rare disease variants through hierarchical modeling. BMC Proc. 2011; suppl 9(5):S16. [PubMed: 22373042]

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

Fan R, Huang C-H, Lo S-H, Zheng T, Ionita-Laza I. Identifying rare disease variants in the Genetic Analysis Workshop 17 simulated data: a comparison of several statistical approaches. BMC Proc. 2011; 5(suppl 9):S17. [PubMed: 22373071]

Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. 2011; 7 e1001289.

Kumar P, Henikoff S, Ng PC. Predicting the effects of coding nonsynonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4:1073–1081. [PubMed: 19561590]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–321. [PubMed: 18691683]

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

Ngwa JS, Manning AK, Grimsby JL, Lu C, Zhuang WV, DeStefano AL. Pathway analysis following association study. BMC Proc. 2011; 5(suppl 9):S18. [PubMed: 22373100]

Pan W, Shen X. Adaptive tests for association analysis of rare variants. Genet Epidemiol. 2011; 35:381–388. [PubMed: 21520272]

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86:832–838. [PubMed: 20471002]

Sun YV, Sung YJ, Tintle N, Ziegler A. Identification of genetic association of multiple rare variants using collapsing methods. Genet Epidemiol. 2011; X(suppl X):X–X.

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

Turkmen AS, Lin S. Gene-based partial least-squares approaches for detecting rare variant associations with complex traits. BMC Proc. 2011; 5(suppl 9):S19. [PubMed: 22373126]

Wei P, Liu X, Fu YX. Incorporating predicted functions of nonsynonymous variants into gene-based analysis of exome sequencing data: a comparative study. BMC Proc. 2011; 5(suppl 9):S20. [PubMed: 22373178]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data using the sequence kernel association test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]

Yip WK, De G, Raby BA, Laird N. Identifying causal rare variants of disease through family-based analysis of GAW17 data set. BMC Proc. 2011; 5(suppl 9):S21. [PubMed: 22373204]

Zhou H, Alexander DH, Sehl ME, Sinsheimer JS, Sobel EM, Lange K. Penalized regression for genome-wide association screening of sequence data. Pac Symp Biocomput. 2011; 2011:106–117. [PubMed: 21121038]

Zhou H, Sehl ME, Sinsheimer JS, Lange K. Association screening of common and rare genetic variants by penalized regression. Bioinformatics. 2010; 26:2375–2382. [PubMed: 20693321]

**Table I**

Overview of Group 3 contributions

| Contribution | Phenotype | Replicates | External information | Statistical methods | Knowledge of the simulation model |
|---|---|---|---|---|---|
| Chen [2011] | Q1, Q2 | All | Predicted functions of SNPs: synonymous/ nonsynonymous | Bayesian hierarchical model | Yes |
| Fan et al. [2011] | Q1, Q2, Affected, dichotomized Q1 and Q2 | All | Predicted functions of SNPs: synonymous/ nonsynonymous; VEGF pathway | Combined multivariate and collapsing, weighted sum, replication-based method | Yes |
| Massanet-Villa [personal communication, 2010] | Affected | 1 | Protein interactions | Hypergeometric distribution | No |
| Ngwa et al. [2011] | Q1, Q4 | All | Gene sets: MSigDB, canonical pathways in IPA | Linear regression-based association test, GSEA, empirical enrichment, Fisher's exact test | Yes |
| Turkmen and Lin [2011] | Q1, Q2 | All | SNPs within a gene | Gene-based partial least-squares method | Yes |
| Wei et al. [2011] | Affected | 1 | Predicted functions of SNPs: synonymous/ nonsynonymous, PolyPhen-2, SIFT | Variable threshold association test, normal mixture model | No |
| Yip et al. [2011][a] | Q1 | All | SNPs within a gene | Nonparametric linkage analysis, modified FBAT | Yes |
| Zhou [personal communication, 2010] | Q1, Q2 | All | SNPs within a gene | LASSO regression with gene-specific L2 norm penalties | Yes |

[a]Family-based data were used.