Published in final edited form as:

Genet Epidemiol. 2011 ; 35(Suppl 1): S18–S21. doi:10.1002/gepi.20644.

Detecting Multiple Causal Rare Variants in Exome Sequence Data

Kenny Q. Ye¹ and Corinne D. Engelman²

¹Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY

²Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, Madison, WI

Abstract

Recent advances in sequencing technology have presented both opportunities and challenges, with limited statistical power to detect a single causal rare variant with practical sample sizes. To overcome this, the contributors to Group 1 of Genetic Analysis Workshop 17 sought to develop methods to detect the combined signal of multiple causal rare variants in a biologically meaningful way. The contributors used genes, genome location proximity, or genetic pathways as the basic unit in combining the information from multiple variants. Weaknesses of the exome sequence data and the relative strengths and weaknesses of the five approaches are discussed.

Keywords

Bayesian; pathways; simulated

Introduction

Recent advances in sequencing technology have provided genetic epidemiologists with new opportunities to investigate the role of rare variants in common diseases. Yet they also present a great challenge because the statistical power to detect a single causal rare variant is limited with practical sample sizes. To overcome the lack of statistical power in detecting a single causal rare variant, researchers have devised a natural strategy to detect the combined signal of multiple causal rare variants in a biologically meaningful way. The five papers summarized here, from Group 1 of Genetic Analysis Workshop 17 (GAW17), follow the same general strategy with somewhat different approaches to analyzing the provided data.

The GAW17 data are based on real exome sequence data of 3,205 genes from 697 individuals provided by the 1000 Genomes Project (<http://www.1000genomes.org>). The phenotypes were simulated with a relatively small number of causative variants with various frequencies and effect sizes across a small number of genes. There were 39 single-nucleotide polymorphisms (SNPs) in 9 genes for quantitative risk factor Q1, 72 SNPs in 13 genes for quantitative risk factor Q2, and 51 SNPs in 15 genes for a latent liability trait that was not included in the distributed phenotype data. The genes involved in the phenotypes were selected primarily from a few biological pathways. In addition, a binary phenotype of

Correspondence Author: Corinne D. Engelman, Department of Population Health Sciences, University of Wisconsin School of Medicine and Public Health, 610 Walnut Street, 1007A WARF, Madison, WI 53726-2397, Phone: (608) 265-5491, Fax: (608) 263-2820, cengelman@wisc.edu.

The authors contributed equally to this work

affection status was assigned based on a linear combination of Q1, Q2, the latent trait, and Q4, which is not associated with any genotype in the data set. Further details of the simulated data can be found elsewhere [Almasy et al., 2011].

Two different study designs were simulated in the GAW17 data: unrelated individuals and large families. All five Group 1 papers used only the unrelated individuals data set, and these contributions are summarized in Table I. All five groups of investigators attempted to detect combined signals of multiple variants in their analyses. Pradhan et al. [2011], Pungpapong et al. [2011], and Xu and George [2011] used genes as the basic unit in combining the information from multiple variants. Agne et al. [2011] used nonoverlapping regions based on genome location proximity, and Xing et al. [2011] used genetic pathways as the basic unit for the combined signal.

Four out of the five work groups, with the exception of Pradhan et al. [2011], used a single measure to summarize a set of SNPs that were biologically related. They then tested for association between phenotypes and these measures. Pradhan et al. [2011] simply used multiple regressions to entertain multiple variants.

Two of the five work groups, Agne et al. [2011] and Xu and George [2011], focused exclusively on rare variants and their association with the binary phenotype of affection status, using similar statistical approaches. The other three groups included common variants as part of the association analysis and focused only on the quantitative traits.

Two of the five work groups, Pradhan et al. [2011] and Pungpapong et al. [2011], used similar Bayesian frameworks to identify genes and SNPs associated with the phenotype. Unlike the other three work groups, their approach allowed multiple genes, pathways, and loci to be considered simultaneously when detecting their association with phenotypes. Although the frameworks were similar, the implementations were quite different.

Each of the five papers is described briefly in the “Methods and Results” section, roughly in the order of the complexity of their approaches. More detail can be found in the individual papers published in the full GAW17 proceedings.

Methods and Results

We start with two papers that are similar in several aspects. Both Agne et al. [2011] and Xu and George [2011] analyze the simulated affection status, and both methods are based on the number of rare variants in predefined genomic regions. Although both methods can accommodate different definitions of rare variants, Xu and George [2011] define a rare variant as one having a minor allele frequency (MAF) < 0.01 and Agne et al. [2011] count only private variants, which occur once in the set of 697 unrelated individuals. One big difference between the two work groups is that Agne et al. [2011] use nonoverlapping bins of a fixed number of SNPs (30 and 100 SNPs) as the predefined genomic regions, and Xu and George [2011] use genes as the predefined genomic regions. Note that the number of SNPs per gene ranges from 1 to 231 with more than one-third of the genes having only one SNP. The two methods also differ in how the test statistics are constructed. Xu and George [2011] count the number of rare variants for each individual and then use the common two-sample *T*-test statistic with pooled variance to compare the average number of rare variants per individual between the case subjects and control subjects. They use the statistic's normal approximation to obtain the *p*-values. For each fixed bin, Agne et al. [2011] compare the number of private variants in case subjects to the number of SNPs in both case subjects and control subjects. The statistical significance of this discrepancy is then evaluated by permuting the individual affection status a large number of times.

These two work groups present their results in different ways. Xu and George [2011] estimate type I error and power of their method in a straightforward way using associated genes and nonassociated genes. Their results are promising; the type I error was well controlled, and the average power over 200 replicates was 0.16 at the 0.01 significance level and 0.10 at the 0.001 significance level. They also reported the top 10 genes that were most frequently identified in the 200 replicates. Two Q1-associated genes, *FLT1* and *PIK3C2B*, were at the top of the list, appearing 136 times and 87 times, respectively, at the 0.001 significance level; these were true positives. However, the next eight genes, appearing between 33 and 79 times out of 200 replicates at the 0.001 significance level, were all false positives.

Agne et al. [2011] set the significance level for each permutation test at 0.05. To combine these simulation-by-simulation reports of significant regions, they use the concept of return frequency, a count of the number of times a region is significant out of 200 simulations. If a region had no causal variants, the chance of observing a certain return frequency was evaluated using a Poisson approximation. For both fixed-bin sizes of 30 and 100 SNPs, all the top 10 return regions had return frequencies greater than 0.25 (50 out of 200), which corresponded to a chance of 10^{-19} under the null hypothesis. Note that such a high statistical significance was obtained by aggregating the statistics for each of the 200 replicates. With a bin size of 100 SNPs, 3 of the top 10 regions contained 7 causal private SNPs, and with a bin size of 30 SNPs, 2 of the top 10 regions contained 3 causal private SNPs; both analyses exceeded the expected number of causal private variants discovered at random: 3.56 (100-SNP bin) and 1.10 (30-SNP bin). The false positives were rather high because the rest of the top regions contained no private causal variants.

Xing et al. [2011] apply a weighted group-wise association test to analyze the association of quantitative trait Q1 at the biological pathway level. They first create a genetic summary score by summing the number of minor alleles over all SNPs residing in a predefined biological pathway, either with equal weight or with a weighting scheme that extends the idea of Madsen and Browning [2009] to quantitative traits. The weight for each SNP is given by

$$w_j = \frac{1}{[2Np_j(1-p_j)]^{1/2}}, \quad (1)$$

where

$$p_j = \frac{\sum_{i=1}^N I_{ij} \delta(Y_i) + 1}{2 \sum_{i=1}^N \delta(Y_i) + 2}, \quad (2)$$

I_{ij} is the number of minor alleles of individual i at the SNP, $\delta(Y_i) = 1$ if the phenotype Y_i is within one standard deviation of the mean and $\delta(Y_i) = 0$ otherwise. This weighting scheme assigns a higher weight to SNPs that are rare among individuals with nonextreme phenotypes. Xing et al. [2011] then use a simple linear regression model to test the association between the phenotype and the genetic summary score. When the weighting scheme is used, permutation tests are used to obtain the p -values because the weight assignment uses phenotypes. For the unweighted summary score, standard methods for linear regression are used.

Xing et al. [2011] use this method to evaluate 809 pathways defined by PharmGKB and 3,009 pathways defined by GO (Genetic Ontology). Bonferroni correction was used to adjust for multiple testing. Among 12 significant PharmGKB pathways, most were somewhat related to the vascular endothelial growth factor (VEGF) pathway, which includes most of the nine Q1-associated genes. Although the relationship between the significant GO processes and the VEGF pathway was less clear, the GO pathways contained as many Q1-associated genes as the significant PharmGKB pathways. One interesting observation from the results is that the unweighted score appears to work as well as, if not better than, the weighted summary score. This can be explained by the fact that in the simulation the effect size of the causal variants was not related to the MAF, and the weighting scheme improves the power only if rare variants tend to have a larger effect.

Unlike the three contributions just discussed, the approach used by Pradhan et al. [2011] considers the effects of multiple genes simultaneously in a multiple linear regression setting, and unlike four other papers discussed here, the genotypes are not summarized before their associations with the phenotypes are assessed. Pradhan and colleagues' approach is based on a Bayesian framework of model selection as proposed by Yoon [2006]. They first consider a model space in which each model under consideration contains a fixed number of factors (usually small and set by the users). The factors are then ranked according to their marginal posterior probabilities.

Pradhan and colleagues analyze the data at two different levels, first by treating individual SNPs as factors and then by treating genes as factors. When a gene is treated as a factor, all variants in the gene enter the model as a group. Because each gene contains a different number of SNPs, the model with larger genes tends to fit the data better so that the large genes are favored if flat prior probabilities are assigned to each model. In an attempt to correct such bias, the prior probability of a model is set to be proportional to $e^{-k/2}$, where k is the total number of model parameters, following the same reasoning of the Bayesian information criterion (BIC). By using standard conjugate prior probabilities for the regression model, Pradhan and colleagues could compute the posterior probability of a model efficiently with a closed-form expression. Because there was a large number of candidate models, they used a straightforward Metropolis-Hasting algorithm to estimate the marginal posterior probabilities.

Pradhan et al. [2011] applied their method to the quantitative trait Q1. The results were reported in two ways. First, they looked at the top 10 genes (or top 20 SNPs) ranked by the marginal posterior probabilities. Second, they looked at the genes (or SNPs) whose marginal posterior probability passed a certain threshold at 0.5 or 0.1. Using gene-level analysis with the model size fixed at three factors and a threshold of marginal posterior probability of 0.5, Pradhan and colleagues' methods detected *FLT1*, a Q1-associated gene, in each of the 10 replicates; *KDR*, another Q1-associated gene, in 4 out of 10 replicates; and some non-Q1-related genes in 5 out of 10 replicates. Relaxing the threshold to 0.1 did not detect Q1-associated genes much more often but greatly increased the number of false positives. Besides *FLT1* and *KDR*, three other Q1-associated genes were among the top 10 genes at least once out of 10 replicates. Their SNP-level analysis consistently identified several Q1-associated SNPs in *FLT1* and *KDR*. In general, the identified variants had higher MAFs than nonidentified causal variants.

The work by Pungpapong et al. [2011] is unique in the sense that it uses a two-stage approach and that at each stage the phenotype information is used. In the first stage, Pungpapong and colleagues use penalized orthogonal-components regression (POCRE) [Zhang et al., 2009] to build the model for each genetic region using a training data set, and then they calculate the predicted values in the test data set. They use these values as gene-

level markers. POCRE itself has a function of variable selection so that the “markers” of a gene might not be constructed from all SNPs in the gene. In the second stage, all gene-level markers, together with some covariates, are entered into an empirical Bayesian variable selection process developed by Johnstone and Silverman [2004, 2005]. Unlike the full Bayesian methods used by Pradhan et al. [2011], in which the users specified all prior parameters, empirical Bayesian methods determine some prior parameters from the data.

Pungpapong et al. [2011] apply their approach to Q1 and use one replicate as the training set to obtain the gene-level markers. They then report the number of times a gene shows a nonzero effect from the empirical Bayes variable selection process. Only 4 genes appeared more than 5 times out of 200 replicates. Among them, *FLT1* appeared in all the replicates, *KDR* in 26%, *ARNT* in 6%, and *RIPK3* in 3%. All but *RIPK3* are Q1-associated genes. In addition, 98 noncausal genes were identified out of 200 replicates. Pungpapong and colleagues’ results are, in general, comparable to those of Pradhan et al. [2011] in terms of power and false positives.

Discussion

The five papers summarized here explore how rare variants play a role in common diseases. For most cases, we cannot directly compare the results between the different approaches because the work groups focused on different traits and different types of biological signals. Nevertheless, we attempt to discuss the relative strengths and weaknesses of the approaches.

We begin with some weaknesses of the GAW17 data. First, the low-coverage sequencing may have been insufficient for accurate genotype calling at many loci. Furthermore, the moderate sample size of 697 unrelated individuals is less than the sample size in most genome-wide association studies in reality and is thus underpowered. The relatively small sample size also leads to many rare variants being highly correlated with each other in the data. For example, the lone SNP in *VEGFC*, a private and causal SNP for Q1, was observed in an individual with 12 other private SNPs. The small sample size may explain some nonrandom type I error observed by Xu and George [2011] and Agne et al. [2011]. That is, the same gene (or locus) was wrongly identified as significant in many replicates, which could have resulted from one or more SNPs in this gene (or locus) being highly correlated with one or more causal SNPs. When evaluating the results of these papers, another fact to keep in mind is that in the simulated data, the minor alleles of the causal variants were all positively associated with the disease or disease traits. How much this reflects the biological truth is very much debatable. Had they not reflected the true biology, the results of Agne et al. [2011], Xing et al. [2011] and Xu and George [2011] would likely have been overoptimistic.

Several contributors to Group 1 pointed out that the large disparity in the number of SNPs in each gene, resulting, in part, from varying gene size, could be problematic if a gene was treated as the basic unit in the analysis, because genes with more SNPs tended to be favored over genes with only one SNP. The same problem also existed, but to a lesser degree, if the analysis treated a pathway as the basic unit. To overcome this problem, Pradhan et al. [2011] assigned lower prior probabilities to genes with more SNPs. However, they also reported that their adjustment, which follows the argument of the BIC, showed uneven effects under different parameter settings. The fixed-size bin approach used by Agne et al. [2011] directly targeted this problem. Instead of making statistical inferences on genes or pathways, they worked on loci, a more traditional concept in association and linkage studies, and collapsed private variants across each nonoverlapping bin with a fixed number of SNPs. In that way, the statistical power was even for each genetic region. However, the GAW17 data are not ideal for this approach because the sequencing was performed on selected gene-coding

regions, not across the entire genome. Therefore some of the bins may span a large area of the genome, which may partly explain the low detection rate in Agne and colleagues' results. On the other hand, they did manage to find causal regions not identified by other work groups, including the gene-level analysis by Xu and George [2011] on the same traits using a similar statistic. Agne et al.'s method also has the flexibility to include SNPs located in intergenic regions, because the genetic variants are not combined by gene or pathway but by adjacent SNPs, regardless of whether they are intragenic or intergenic.

Three contributions, Pradhan et al. [2011], Pungpapong et al. [2011], and Xing et al. [2011], analyzed the quantitative trait Q1. The results of the first two work groups are mostly comparable. Both detected *FLT1* in a vast majority of the replicates and *KDR* in 25–40% of the replicates. Both genes contain a large number of causal SNPs (11 for *FLT1* and 10 for *KDR*), including the four most common SNPs among causal variants. Xing et al. [2011] also reported these two genes, which, together with three other causal genes (*HIF1A*, *VEGFA*, and *ARNT*), are among the 10 genes most frequently included in significant pathways. *HIF1A*, *VEGFA*, and *ARNT* were also detected by Pradhan et al. [2011] in at least 1 of the 10 replicates, and *ARNT* was the third most frequently gene seen by Pungpapong et al. [2011]. Because Xing et al. [2011] did not show actual frequencies of these genes, we were unable to judge whether pathway-level analysis enhances the detection of genes with smaller signals but in the same pathway with genes with strong signals. *FLT1* was also reported by Xu and George [2011] to be the most frequently detected gene by a good margin in their analysis on the affection status. The second most frequently detected gene in their analysis was *PIK3C2B*, which contains 24 causal SNPs, although most have a small effect size and 15 are private variants. It does come as a bit of a surprise that this gene went undetected by Agne et al. [2011]. Overall, it is clear that genes with a large number of causal SNPs are more likely to be detected, which is not a surprise and follows the conventional wisdom.

In conclusion, we have summarized the work of five GAW17 Group 1 contributions, whose investigators sought to develop methods to detect the combined signal of multiple causal rare variants in a biologically meaningful way. The work groups used genes, genome location proximity, or genetic pathways as the basic unit in combining the information from multiple variants. These methods show a great deal of promise in exploring how rare variants play a role in common diseases.

Acknowledgments

The Genetic Analysis Workshops are supported by National Institutes of Health grant R01 GM031575. We thank Group 1 participants for their contributions.

References

- Agne M, Huang C-H, Hu I, Wang H, Zheng T, Lo S-H. Identifying influential regions in extremely rare variants using a fixed-bin approach. *BMC Proc.* 2011; 5(suppl 9):S3.
- Almasy LA, Dyer TD, Peralta JM, Kent JW Jr, Charlesworth JC, Curran JE, Blangero J. Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proc.* 2011; 5(suppl 9):S2.
- Johnstone IM, Silverman BW. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Ann Stat.* 2004; 32:1594–1649.
- Johnstone IM, Silverman BW. Ebayesthresh: R programs for empirical Bayes thresholding. *J Stat Softw.* 2005; 12:1–38.
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
- Pradhan K, Yoon SC, Wang T, Ye K. Identification of genes and variants associated with quantitative traits using Bayesian factor screening. *BMC Proc.* 2011; 5(suppl 9):S4.

- Pungpapong V, Wang L, Lin Y, Zhang D, Zhang M. Genome-wide association analysis of GAW17 data using empirical Bayes variable selection. BMC Proc. 2011; 5(suppl 9):S5.
- Xing C, Satten GA, Allen AS. A weighted accumulation test for associating rare genetic variation with quantitative phenotypes. BMC Proc. 2011; 5(suppl 9):S6.
- Xu H, George V. A gene-based approach for testing association of rare alleles. BMC Proc. 2011; 5(suppl 9):S7.
- Yoon, S. PhD dissertation. State University of New York; Stony Brook: 2006. Bayesian Factor Screening.
- Zhang M, Lin Y, Wang L, Pungpapong V, Fleet JC, Zhang D. Case-control genome-wide association study of rheumatoid arthritis from Genetic Analysis Workshop 16 using penalized orthogonal-components regression-linear discriminant analysis. BMC Proc. 2009; 3(suppl 7):S17. [PubMed: 20018006]

Table 1

Overview of Group 1 contributions

Contribution	Phenotype	Unit to combine variants	Minor allele frequency threshold	Analytic approach
Agne et al. [2011]	Affected	Nonoverlapping bins of a fixed number of SNPs	Private	Number of private variants in case subjects/number of SNPs in case subjects and control subjects; permutation tests
Pradhan et al. [2011]	Q1	Gene	None	Bayesian; multiple linear regression
Pungpapong et al. [2011]	Q1	Gene	None	Penalized orthogonal-components regression (POCRE); empirical Bayesian variable selection
Xing et al. [2011]	Q1	Genetic pathways	None	Weighted group-wise association test using linear regression and permutation tests
Xu and George [2011]	Affected	Gene	<0.01	T test