Published in final edited form as:

Genet Epidemiol. 2009 ; 33(Suppl 1): S74–S80. doi:10.1002/gepi.20476.

Inclusion of *a Priori* Information in Genome-Wide Association Analysis

Nathan Tintle¹, Francesca Lantieri^{2,3}, Jérémie Lebrech^{4,5}, Melanie Sohns⁶, David Ballard⁷, and Heike Bickeböllner^{6,*}

¹ Department of Mathematics, Hope College, Holland, MI ² The Children's Hospital of Philadelphia, Philadelphia, PA ³ Università degli Studi di Genova, Genoa, Italy ⁴ Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands ⁵ Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands ⁶ Department of Genetic Epidemiology, University Medical Center Göttingen, Göttingen, Germany ⁷ Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT

Abstract

Genome-wide association studies (GWAS) continue to gain in popularity. To utilize the wealth of data created more effectively, a variety of methods have recently been proposed to include *a priori* information (e.g., biologically interpretable sets of genes, candidate gene information, or gene expression) in GWAS analysis. Six contributions to Genetic Analysis Workshop 16 Group 11 applied novel or recently proposed methods to GWAS of rheumatoid arthritis and heart disease related phenotypes. The results of these analyses were a variety of novel candidate genes and sets of genes, in addition to the validation of well known genotype-phenotype associations. However, because many methods are relatively new, they would benefit from further methodological research to ensure that they maintain type I error rates while increasing power to find additional associations. When methods have been adapted from other study types (e.g., gene expression data analysis or linkage analysis) the lessons learned there should be used to guide implementation of techniques. Lastly, many open research questions exist concerning the logistic details of the origin of the *a priori* information and the way to incorporate it. Overall, our group has demonstrated a strong potential for identifying novel genotype-phenotype relationships by including *a priori* data in the analysis of GWAS, while also uncovering a series of questions requiring further research.

Keywords

gene set analysis; external information; gene expression; hierarchical Bayesian model; candidate regions; candidate genes; pathway

INTRODUCTION

Conducting a genome-wide association study (GWAS) is an increasingly popular approach to identify the genetic components of disease. Reasons for this popularity include an increasing understanding of the human genome and the refinement and availability of high-throughput single-nucleotide polymorphism (SNP) genotyping technologies. However, for some diseases, achieving adequate statistical power in GWAS can still be prohibitive [Amos, 2007]. Power can be extremely low for SNPs contributing only weakly to disease,

even in the largest GWAS. This limitation is considerable because it has been argued that GWAS could be the answer to finding the many genes associated with complex diseases [Bossé et al., 2009].

Group 11 of Genetic Analysis Workshop 16 examined various methods of including *a priori* information into a GWAS under the theme of “incorporating information on gene expression/function/pathways into genome-wide searches”. The overall motivation of these approaches is straightforward: by including additional available information in the GWAS, there will be an increased ability to find significant genotype-phenotype associations that would not have been found using GWAS data alone. The methods examined here use *a priori* information from other association studies, other types of experimental data (e.g., gene expression), and information obtained from our increased understanding of the human genome (e.g., gene pathway information). Group 11 consists of six papers that can be classified into two overlapping subgroups depending on their main goal: gene set identification and gene/SNP prioritization. Three papers [Ballard et al., 2009; Sohns et al., 2009; Tintle et al., 2009] used or compared methods of using SNP-level statistics obtained in GWAS to identify *a priori* defined, biologically relevant sets of genes (“gene sets”) that are significantly associated with the phenotype. Four papers ranked genes or SNPs in terms of association with the phenotype using gene sets [Lebrec et al., 2009a; Sohns et al., 2009], gene expression information [Charlesworth et al., 2009], or prior association reports [Lantieri et al., 2009]. Table I summarizes the six contributions, including information on the data used, the methods and the goal of the analysis.

Contributors to Group 11 considered data from Problems 1, 2, and 3. Data for Problem 1 consist of whole genome (Illumina 550k chip) case-control association data from the North American Rheumatoid Arthritis Consortium (NARAC), while data for Problem 2 are family-based genome-wide (Affymetrix 550K chip) association data for cardiac-related phenotypes from the Framingham Heart Study (FHS). The Problem 3 dataset consists of simulated cardiac-related phenotypes for the FHS genotype data. More details are available elsewhere in this volume.

In this paper we describe key findings from our consideration of *a priori* data in GWAS. In the following sections we present logistic issues common to many of the contributions, gene-set identification approaches, and gene/SNP ranking methods. Analysis of real and simulated data demonstrates the potential of these methods to identify novel genotype-phenotype relationships.

METHODS

All papers used standard methods for limiting the SNPs used in the analysis to high-quality SNPs according to Hardy-Weinberg equilibrium, minor allele frequency, and missing data, though the cutoffs used in making these decisions varied between papers.

Logistic Issues

There were three main logistic issues common to members of our group: how to assign SNPs to genes, how to use single-SNP association tests to yield a gene level statistic, and where to obtain the *a priori* information for the analysis. In order to incorporate *a priori* information into GWAS, all participants of Group 11 attempted to combine a SNP-level/trait association signal with gene-level information. This required making a SNP-to-gene assignment and computing a gene-level summary measure of association. Ballard et al. [2009] only assigned intragenic SNPs to the gene of interest, other authors allowed intergenic SNPs to be assigned to their closest gene as long as they were located within 5 kb [Lantieri et al., 2009], 25 kb [Charlesworth et al., 2009], or 500 kb [Lebrec et al., 2009a];

Sohns et al., 2009; Tintle et al., 2009] of the start/end position of the neighboring gene. Charlesworth et al. [2009], Lebec et al. [2009a], Sohn et al. [2009], and Tintle et al. [2009] used the most significant SNP assigned to a gene as a summary measure of association for that gene and then attempted to account for different gene lengths. Ballard et al. [2009] used the *p*-value from the simultaneous addition of all intragenic SNPs in a multiple regression of case-control status onto the SNP genotypes as a gene-level score. Lantieri et al. [2009] and portions of Sohn et al. [2009] used the raw SNP-level association signal directly. Because *a priori* information can be obtained from a multitude of sources, Table I shows the sources used in the analyses considered here.

Gene set identification

Gene set analysis (GSA) was initially proposed for gene expression [Goeman and Buhlmann, 2007; Subramanian et al., 2005] and recently for GWAS [Chasman, 2008; Wang et al., 2007]. First, GSA combines gene-level statistics into a single statistic for each gene set. Gene sets are then evaluated for their statistical significance, in order to identify sets associated with the phenotype. Ballard et al. [2009], Sohn et al. [2009], and Tintle et al. [2009] considered GSA. Multiple GSA statistics were considered. Fisher's exact test (FET) and its asymptotic equivalent were considered by Tintle et al. [2009] and Ballard et al. [2009], respectively. Gene set enrichment analysis (GSEA) [Subramanian et al., 2005; Wang et al., 2007] was used by Sohn et al. [2009] and Tintle et al. [2009]. SUMSTAT, implemented as MAXMEAN [Efron and Tibshirani, 2007], was used by both Tintle et al. [2009] and Ballard et al. [2009] and SUMSQ [Dinu et al., 2007] was used by Tintle et al. [2009].

In addition to the choice of GSA statistic, three permutation methods were used to determine significance: 1) permutation of case-control status (phenotype, Sohn et al. [2009]), 2) selection of random gene sets (gene, Ballard et al. [2009], Tintle et al. [2009]; random sets must be the same size as the set of interest), and 3) shuffling SNP statistics across the SNPs (SNP). All three methods compute a *p*-value by comparing the observed statistic to statistics obtained from permutations. Tintle et al. [2009] uses an false-discovery rate (FDR) of 5% to assess statistical significance, Ballard et al. [2009] uses an FDR of 1%, and Sohn et al. [2009] uses the list of gene sets ranked by *p*-value. All three methods were used in a follow-up analysis presented in this paper that compares all methods and all permutation strategies on the same (NARAC) data.

SNP/gene prioritization based on *a priori* information

Instead of assessing statistical significance of gene sets, four papers [Charlesworth et al., 2009; Lantieri et al., 2009; Lebec et al., 2009a; Sohn et al., 2009] prioritized SNPs or genes using *a priori* information, namely, gene set, gene expression, and candidate genes/regions.

Incorporating gene set/SNP information in the analysis of GWAS—Lebec et al. [2009a] and Sohn et al. [2009] used hierarchical Bayesian models to integrate GWAS with prior pathway (gene set) and/or SNP information (non-synonymous, 3'UTR, intron, etc.), in order to rank genes for association with rheumatoid arthritis (RA). Prior information was gathered in a matrix of binary indicators. The corresponding prior pathway effect coefficients were estimated via empirical Bayesian procedures. In the so-called linear regression on pathways (LRP) approach, Lebec et al. [2009a] linearly regressed gene-level summary data (positive log-odds ratio of the most significant gene-assigned SNP in an additive model) on pathway effects. Sohn et al. [2009] modeled the non-centrality parameter of the SNP-chi-squared test using hierarchical Bayesian prioritization (HBP) [Lewinger et al., 2007]. Both then used posterior probabilities of gene/SNP association with

RA for gene prioritization. Sohns et al. [2009] also applied GSEA by using the leading edge subset (LES) of GSEA, which pinpoints the subset of genes most strongly associated with RA. Sohns et al. [2009] compared the top-ranking genes obtained by GSEA, HBP and hybrid versions thereof. In a follow-up analysis presented in this paper, Sohns et al. [2009] refit the HBP model using the same eight pathways selected by LRP, in order to compare the priority lists generated by the two different empirical Bayesian models (LRP and HBP).

Incorporating gene expression data in the analysis of GWAS—Charlesworth et al. [2009] demonstrated that gene expression data can be incorporated into a GWAS by using high-density lipoprotein (HDL) cholesterol levels from untransformed lymphocytes. First, they used a measured genotype analysis to compute a gene-based test of association, adjusting the lowest p -value for the effective number of SNPs [Li and Ji, 2005] ($p_{corrected} = 1 - (1 - p_{nominal})^e$, where $p_{nominal}$ is the lowest p -value and e is the effective number of SNPs). Correlations between quantitative gene expression profiles and HDL-C were pre-calculated [Göring et al., 2007] and the two tests were combined using a Z-transform test [Whitlock et al., 2005].

Incorporating information on candidate genes/regions in the analysis of GWAS—Another source of external information is candidate genes/regions suggested as significant by prior studies. Lantieri et al. [2009] compared four different methods of utilizing candidate gene information in GWAS. All methods required the assignment of SNPs to either a “candidate SNP” set or a “non-candidate SNP” set. Two sets of candidate SNPs were used: 1) SNPs nearby to 64 genes reported as associated with RA in a public database [Becker et al., 2004]; 2) SNPs recently found associated with RA and nearby SNPs in linkage disequilibrium (LD) [Gregersen et al., 1987; Hinks et al., 2006; Plenge et al., 2007; Remmers et al., 2007; Wellcome Trust Case Control Consortium, 2007]. It is important to note that all SNPs in the known HLA region were eliminated from the entire analysis. Four methods of handling candidate SNPs in the context of GWAS were considered: 1) the posterior odds for association method (PO) [Curtis et al., 2007], 2) the false-positive report probability method (FPRP) [Wacholder et al., 2004], 3) prioritized subset analysis (PSA) [Li et al., 2008], and 4) the empirically corrected p -values by permutation method (EMP). Each of the first two methods (PO and FPRP) was implemented twice with different prior probabilities (PO1, PO2, FPRP1, FPRP2).

RESULTS

Gene set identification

In the simulated FHS data and using gene permutation, Tintle et al. [2009] found that all considered statistics controlled the type I error rate, and SUMSTAT and SUMSQ were more powerful than FET and GSEA. Using gene permutation for the real FHS data, SUMSTAT and SUMSQ found more sets significantly associated with diabetes and heart disease than FET and GSEA. Ballard et al. [2009] conducted a GSA on the NARAC data using the binomial test and SUMSTAT with gene permutation while Sohns et al. [2009] analyzed the same data using GSEA and phenotype permutation. The sets found to be significant have biological plausibility and there is high overlap between the significant sets, though many are part of the known HLA region. Ballard et al. [2009] also analyzed the data excluding the HLA region. In this situation, SUMSTAT found more sets than the binomial approach.

Follow-up analysis

To enhance comparability between methods, SUMSTAT, SUMSQ, GSEA, and FET were applied to the RA data with and without HLA, using three permutation methods (SNP, gene, and phenotype) and incorporating 825 gene sets from the biological processes of Gene

Ontology (GO) [Harris et al., 2004]. Table II shows the number of significant sets ($p < 0.05$) for each method. SNP permutation in GSEA finds few sets as significant when the HLA region is included, and interpretation is conceptually difficult (e.g., possible issues with LD). SUMSTAT and SUMSQ identify most of the gene sets (SUMSQ: 673/825=82% and SUMSTAT: 704/825=85%) as significant when phenotype permutation is used. For genotype permutation, GSEA, SUMSTAT, and SUMSQ find more sets as significant after excluding the HLA region than on the whole genome. Corresponding to the results of Tintle et al. [2009] (simulated data), FET with gene sampling finds decreasing numbers of significant sets with increasing cutoff (data shown for only 2 of 4 cutoffs). This trend reverses when using phenotype permutation.

SNP/gene prioritization based on a priori information

Incorporating gene set information in the analysis of GWAS—In Sohns et al. [2009], gene lists obtained by the HBP-only strategies are almost identical to the initial ranking. The two GSEA strategies highlight many new genes and are more consistent with each other than the HBP-only methods. In their LRP approach, Lebrech et al. [2009a] considered 27 gene sets of the c5 MSigDB gene set definition [Subramanian et al., 2005] using forward-stepwise regression to select the eight most relevant gene sets. Genes were ranked according to their posterior probability. Lebrech et al. [2009a] varied the relative influence of the gene set information by giving more or less weight to it. With less weight, the top genes identified were in the HLA region due to the dominant HLA signal. Seventeen genes were common to the top 100 gene lists regardless of the weight. Among those were at least four genes likely to be associated with RA, including *CD40* [Raychaudhuri et al., 2008], which was not found without gene set integration.

Follow-up analysis—To allow for closer comparison of HBP and LRP, Sohns et al. [2009] refit the HBP model using the same eight gene sets excluding the functional SNP annotations and using the gene-level signal instead of the SNP signal. The gene set parameter estimates of LRP show consistency with the estimates of the linear HBP sub-model for the strength of association. Comparing the top 200 genes between the two approaches leads to different results depending on which weights the LRP method uses. Specifically, there is a high overlap (153 genes) between HBP and LRP, with less weight on gene set information. Of the 153, only 14 non-HLA genes are not in the initial top 200 genes. Using LRP with more weight on the gene set ranking reduces the overlap of the top 200 genes between LRP and HBP to 24 genes, only six of which are also in the initial ranking. After excluding the HLA region, the overlapping sets results described above reversed. Pathway coefficient parameter estimates remained similar between the two methods.

Incorporating gene expression data in the analysis of GWAS—In their association analysis of FHS data, Charlesworth et al. [2009] found 14 genes significantly associated with HDL cholesterol level at FDR 1%. After combining results from association analysis, and gene expression analysis, 39 genes are statistically significant at FDR 1%. Of those, seven would not have been identified by either approach alone. These 39 genes include some well known cholesterol-related genes as well as a substantial number of understudied genes.

Incorporating information about candidate genes/regions in the analysis of GWAS—Lantieri et al. [2009] compared four methods of handling “candidate SNP” sets in GWAS. Figure 1 illustrates the overlap of those with the standard analysis ignoring candidate regions for the less restrictive candidate set. The other candidate set (not shown) revealed similar patterns. For the comparisons, the 2000 most significant SNPs following

ranking by each method were chosen. Ignoring candidate SNP information (original GWA) and PSA including candidate gene information find the least number of sets to be significant. PO and FPRP strongly depend on the chosen prior probabilities, with FPRP more permissive than PO under the same prior probability.

DISCUSSION

Our group considered a variety of methods to utilize *a priori* information to increase the power of GWAS. In addition to common logistic issues, methods of determining gene set significance, gene prioritization, and re-ranking SNPs were considered.

Three common logistic issues were identified: assigning SNPs to genes, assigning SNP test statistics to genes, and the source of the *a priori* information. Using a window of $\pm X$ kb for assigning SNPs to genes requires selection of a reasonable window size. The implications of different sizes are unknown. Improvement to the window method would likely be possible if information about LD, enhancers, and repressors were readily accessible for all SNPs of interest and computationally efficient methods existed for assigning SNPs to genes based on this information. While in most cases the maximum SNP statistic for each gene was used to obtain a gene-level statistic, multiple methods were proposed to account for potential large-gene bias from this approach. Insights into this issue may be gleaned from other analytic methods for testing multi-locus effects in GWAS. Lastly, the source of the *a priori* information to be used is of paramount importance. For gene sets, gene expression, or candidate genes, there is little advice available as to how to choose the most relevant and accurate *a priori* information correctly and so arbitrary decisions must be made.

Using gene sets for GWAS is a new idea introduced by Wang et al. [2007]. However, GSA became popular for gene expression analysis several years earlier. In general, similar conclusions to those made in gene expression analysis were drawn based on applying these methods to the GWAS. While the results of the GSA methods are promising, there is a need for more comprehensive theoretical or simulation analysis to validate the ability of these methods to maintain type I error rates and provide increased power to find the consistent, but weak, effects that they purport to find. Lessons learned from GSA in gene expression (e.g., lack of robustness of cutoff-based methods [Allison et al., 2006, Tintle et al., 2008], choice of statistics, and permutation procedures [Efron and Tibshirani 2007, Goeman and Bulhmann 2007]) should be considered in GWAS. Specifically, we note that Efron and Tibshirani have discussed the problem of finding many, many sets as significant (e.g., SUMSTAT and SUMSQ) when using phenotype permutation. They propose a hybrid approach (a mix of both gene and phenotype permutation called restandardization) that has yet to be considered for GWAS.

In general, the SNP/gene prioritization methods result in a mix of known and novel genes identified as significant [Charlesworth et al., 2009; Lebrech et al., 2009a; Sohns et al., 2009] or give prominence to known genes otherwise overlooked in GWAS [Lantieri et al., 2009]. The two hierarchical Bayesian models (LRP and HBP) are similar (see also Chen and Witte [2007]): they both allow natural incorporation of multiple prior gene set effects with GWAS but they differ in their weights of GWAS data relative to prior pathway information. Interestingly, compared with similar methods used on linkage scans [Lebrech et al., 2009b], the proportion of total scan signal variation explained by the pathway information is strikingly small in the GWAS (1% in GWAS vs. 50% in linkage).

CONCLUSION

Utilization of *a priori* information in the analysis of GWAS data offers a promising approach to the inherent challenges of adequate genetic dissection of a trait in GWAS analyses. In our group, these techniques tended to produce novel genes or sets of genes while also validating prior, known, relationships of genes/gene sets with the phenotype of interest. Overall, many of the techniques are rather new and would benefit from simulation studies to validate their robustness, ability to maintain type I error rates, and ability to increase power of single gene/SNP approaches. Additionally, where appropriate, lessons from areas in which some aspects of these methods have been explored in greater detail (e.g., gene expression analysis; multi-locus SNP techniques) should be applied in the further characterization and refinement of these approaches. Lastly, many open research questions exist in the logistic details of how to obtain and include *a priori* information in the analysis. Overall, our group demonstrated very promising results from implementation of methods to include *a priori* data in GWAS, though we uncovered a variety of areas requiring further research.

Acknowledgments

We thank the Group 11 participants for their contributions. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. Authors of this manuscript were supported by the German National Genome Research Network (BMBF, grant 01GS0837) [MS and HB], the United States National Human Genome Research Institute (NIH-NHGRI, R15-HG004543) [NT], the Netherlands Organization for Health Research and Development (ZonMW, grant 40-00812-98-03016) [JL], and the European League Against Rheumatism (Gidael project) [JL].

References

- Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: From disarray to consolidation and consensus. *Nat Rev Genet.* 2006; 7:55–65. [PubMed: 16369572]
- Amos CI. Successful design and conduct of genome-wide association studies. *Hum Mol Genet.* 2007; 16(Spec No 2):R220–5. [PubMed: 17597095]
- Ballard DH, Apornetewan C, Lee JY, Lee JS, Wu Z, Zhao H. A pathway analysis applied to Genetic Analysis Workshop 16 genome-wide rheumatoid arthritis data. *BMC Proc.* 2009; 3(Suppl 7):S91. [PubMed: 20018088]
- Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. *Nat Genet.* 2004; 36:431–2. [PubMed: 15118671]
- Bossé Y, Bacot F, Montpetit A, Rung J, Qu HQ, Engert JC, Polychronakos C, Hudson TJ, Froguel P, Sladek R, Desrosiers M. Identification of susceptibility genes for complex diseases using pooling-based genome-wide association scans. *Hum Genet.* 2009; 125:305–18. [PubMed: 19184112]
- Charlesworth JC, Peralta JM, Drigalenko E, Göring HHH, Almasy L, Dyer TD, Blangero J. Toward the identification of causal genes in complex diseases: A gene-centric joint test of significance combining genomic and transcriptomic data. *BMC Proc.* 2009; 3(Suppl 7):S92. [PubMed: 20018089]
- Chasman DI. On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet Epidemiol.* 2008; 32:658–68. [PubMed: 18481796]
- Chen GK, Witte JS. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet.* 2007; 81:397–404. [PubMed: 17668389]
- Curtis D, Vine AE, Knight J. A pragmatic suggestion for dealing with results for candidate genes obtained from genome-wide association studies. *BMC Genet.* 2007; 8:20. [PubMed: 17490491]
- Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics.* 2007; 8:242. [PubMed: 17612399]
- Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat.* 2007; 1:107–29.

- Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: Methodological issues. *Bioinformatics*. 2007; 23:980–7. [PubMed: 17303618]
- Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat Genet*. 2007; 39:1208–16. [PubMed: 17873875]
- Gregersen PK, Silver J, Winchester RJ. The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum*. 1987; 30:1205–13. [PubMed: 2446635]
- Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2004; 32:D258–61. [PubMed: 14681407]
- Hinks A, Worthington J, Thomson W. The association of PTPN22 with rheumatoid arthritis and juvenile idiopathic arthritis. *Rheumatology (Oxford)*. 2006; 45:365–8. [PubMed: 16418195]
- Lantieri F, Jhun MA, Park J, Park T, Devoto M. Comparative analysis of different approaches for dealing with candidate regions in the context of a genome-wide association study. *BMC Proc*. 2009; 3(Suppl 7):S93. [PubMed: 20018090]
- Lebrec JJ, Huizinga TWJ, Toes REM, Houwing-Duistermaat JJ, van Houwelingen HC. Integration of gene ontogeny pathways with North American Rheumatoid Arthritis Consortium genome-wide association data via linear modeling. *BMC Proc*. 2009a; 3(Suppl 7):S94. [PubMed: 20018091]
- Lebrec JJ, Nishchenko I, van der Wijk HJ, Huizinga TW, van Houwelingen HC. A polygenic model for integration of linkage and pathway information. *Genet Epidemiol*. 2009b; 33:198–206. [PubMed: 18979499]
- Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol*. 2007; 31:871–82. [PubMed: 17654612]
- Li C, Li M, Lange EM, Watanabe RM. Prioritized subset analysis: Improving power in genome-wide association studies. *Hum Hered*. 2008; 65:129–41. [PubMed: 17934316]
- Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*. 2005; 95:221–7. [PubMed: 16077740]
- Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, Li W, Tan AK, Bonnard C, Ong RT, Thalamuthu A, Pettersson S, Liu C, Tian C, Chen WV, Carulli JP, Beckman EM, Altshuler D, Alfredsson L, Criswell LA, Amos CI, Seldin MF, Kastner DL, Klareskog L, Gregersen PK. TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *N Engl J Med*. 2007; 357:1199–209. [PubMed: 17804836]
- Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burt NP, Gianniny L, Korman BD, Padyukov L, Kurreeman FA, Chang M, Catanese JJ, Ding B, Wong S, van der Helm-van Mil AH, Neale BM, Coby J, Cui J, Tak PP, Wolbink GJ, Crusius JB, van der Horst-Bruinsma IE, Criswell LA, Amos CI, Seldin MF, Kastner DL, Ardlie KG, Alfredsson L, Costenbader KH, Altshuler D, Huizinga TW, Shadick NA, Weinblatt ME, de Vries N, Worthington J, Seielstad M, Toes RE, Karlson EW, Begovich AB, Klareskog L, Gregersen PK, Daly MJ, Plenge RM. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet*. 2008; 40:1216–23. [PubMed: 18794853]
- Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, de Bakker PI, Le JM, Lee HS, Batliwalla F, Li W, Masters SL, Booty MG, Carulli JP, Padyukov L, Alfredsson L, Klareskog L, Chen WV, Amos CI, Criswell LA, Seldin MF, Kastner DL, Gregersen PK. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med*. 2007; 357:977–86. [PubMed: 17804842]

- Sohns M, Rosenberger A, Bickeböllner H. Integration of a priori gene set information into genome-wide association studies. *BMC Proc.* 2009; 3(Suppl 7):S95. [PubMed: 20018092]
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA.* 2005; 102:15545–50. [PubMed: 16199517]
- The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007; 447:661–78. [PubMed: 17554300]
- Tintle NL, Best AA, DeJongh M, Van Bruggen D, Heffron F, Porwollik S, Taylor RC. Gene set analyses for interpreting microarray experiments on prokaryotic organisms. *BMC Bioinformatics.* 2008; 9:469. [PubMed: 18986519]
- Tintle NL, Borchers B, Brown M, Bekmetjev A. Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. *BMC Proc.* 2009; 3(Suppl 7):S96. [PubMed: 20018093]
- Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *J Natl Cancer Inst.* 2004; 96:434–42. [PubMed: 15026468]
- Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet.* 2007; 81:1278–83. [PubMed: 17966091]
- Whitlock MC. Combining probability from independent tests: The weighted Z-method is superior to Fisher's approach. *J Evol Biol.* 2005; 18:1368–73. [PubMed: 16135132]

Figure 1. Comparison of four methods of incorporating information on candidate genes/regions Venn diagram with the number of SNPs selected from a set of RA candidate markers using a variety of methods. Total SNPs identified by each method is provided to the right of the diagram in parentheses following the method name. GWA, genome-wide association; EMP, empirically corrected p -value; PO, posterior odds; FPRP, false-positive report probability. For PO and FPRP, two different prior probabilities are used (0.1 and 0.01, denoted 1 and 2, respectively).

Table 1

Characteristics of the 6 GAW16 contributions of Group 11

Author	External information						Test ¹	HLA ²
	Dataset	Sample	Type	Source	Ranking	SNP/gene mapping	Measure of association	
Ballard et al.	NARAC (Problem 1)	862 RA cases, 1140 controls	Gene sets	KEGG, GenMapp, Biocarta	No	Intragenic	Regression	Yes and no
Charlesworth et al.	Framingham (real) (Problem 2)	6334 in families	Gene expression data	Array Express	Yes	Gene±25 kb	Omnibus test of quantitative association	N/A
Leventieri et al.	NARAC (Problem 1)	868 RA cases, 1194 controls	Candidate genes, candidate SNPs	NCBI GAD, Literature, LD, NCBI OMIM, HapMap	Yes	Gene±5 kb SNP±1 mb	Allelic test	No
Leurec et al.	NARAC (Problem 1)	868 RA cases, 1194 controls	Gene sets	MsigDB (GO)	Yes	Gene±500 kb	Trend test	Yes
Sims et al.	NARAC (Problem 1)	868 RA cases, 1194 controls	Gene sets, SNP info	Gen-Gen package (KEGG, Biocarta, GO)	Yes	Gene±500 kb	Trend test	Yes
Tintle et al.	Framingham (real and simulated) (Problems 2 & 3)	957 unrelated	Gene sets	MsigDB (Cytogenetic band, GO)	No	Gene±500 kb	Allelic test	N/A

¹ Test method to assess significance or yield ranking

² HLA region included in the analysis yes or no (applies only to NARAC data)

Table II

Number of significant sets out of 825 gene sets defined based on GO

Method	Permutation/strategy	Including HLA	Excluding HLA
GSEA	SNP	6	90
	Gene	38	81
	Phenotype	87	26
FET 5 ^I	Gene	98	94
	Phenotype	22	19
FET 18 ^I	Gene	8	8
	Phenotype	203	121
SUMSQ	Gene	38	60
	Phenotype	673	645
SUMSTAT	Gene	39	107
	Phenotype	704	680

^I Different cutoffs (5 and 18) for FET were used. See Tintle et al. [2009] for details.