Published in final edited form as:

Genet Epidemiol. 2009 ; 33(Suppl 1): S19–S23. doi:10.1002/gepi.20467.

Multistage Analysis Strategies for Genome-Wide Association Studies: Summary of Group 3 Contributions to Genetic Analysis Workshop 16

Rosalind J. Neuman^{1,2} and Yun Ju Sung^{1,3}

¹Department of Psychiatry, Washington University Medical School, St. Louis, MO

²Department of Genetics, Washington University Medical School, St. Louis, MO

³Division of Biostatistics, Washington University Medical School, St. Louis, MO

Abstract

This contribution summarizes the work done by six independent teams of investigators to identify the genetic and non-genetic variants that work together or independently to predispose to disease. The theme addressed in these studies is multistage strategies in the context of genome-wide association studies. The work performed comes from Group 3 of the Genetic Analysis Workshop 16 held in St. Louis, Missouri in September, 2008. These six studies represent a diversity of multistage methods of which five are applied to the North American Rheumatoid Arthritis Consortium rheumatoid arthritis case-control data, and one method is applied to the low-density lipoprotein phenotype in the Framingham Heart Study simulated data. In the first stage of analyses, the majority of studies used a variety of screening techniques to reduce the noise of single-nucleotide polymorphisms purportedly not involved in the phenotype of interest. Three studies analyzed the data using penalized regression models, either LASSO or the elastic net. The main result was a reconfirmation of the involvement of variants in the HLA region on chromosome 6 with rheumatoid arthritis. The hope is that the intense computational methods highlighted in this group of papers will become useful tools in future genome-wide association studies.

Keywords

imputation; elastic net; two-locus models; LASSO; sliding window

INTRODUCTION

The arrival of the age of genome-wide association studies (GWAS) can be metaphorically thought of as a blessing and a curse. The blessing is the greatly anticipated possibility that the genetics community will be successful in the quest to dissect the mysteries of complex human disease, or at least their genetic underpinnings. That is, if we put our collective shoulders to the grindstone and use our common attributes of brains, sweat, and tears, we may then be able to begin a path to treatment and cures for disease phenotypes. On the other hand, we know that we have opened a Pandora's Box of new challenges. One major challenge is too much 'success', that is too many statistically significant hits, giving rise to the 'multiple testing problem'. This problem is further exacerbated when testing for genetic

interactions. How shall we define the statistical boundary lines that are needed to declare success and be relatively sure that our boundaries are not so stringent as to miss true associations? The cost of genotyping for a GWAS study, once a major concern, has been decreasing. However, the time needed for computations to be completed is still an issue for many researchers, albeit, somewhat lessened by the use of powerful computers and computer clusters. We also note that although GWAS have made progress toward finding single-nucleotide polymorphisms (SNPs) associated with many complex traits [Hindorff et al., www.genome.gov/26525384], those SNPs have been shown to explain a very small percentage of the existing genetic variance of most complex traits [Goldstein, 2009; Kraft, 2009]. Because most of these findings are based on single-marker analyses, we can ask whether, with enough computing power and/or new statistical methods, incorporating gene-gene or gene-environmental interactions may be a way to enhance gene discovery. Will multistage analyses, in which variants are strategically included in the genetic model, increase the level of genetic variance uncovered, thereby increasing confidence that we have detected a true association?

To address these challenges, several authors have previously explored multistage strategies to attain efficient methods for large case-control GWAS consisting of many hundreds of thousands of markers, or more likely to come, many millions of SNPs. Thomas et al. [2004] considered a two-stage design in which a subset of subjects is used to select tag SNPs (tSNPs) to be analyzed on the full sample. Stage 2 combined the information from the subset of the full sample and the full study for association with the original polymorphism. Satagopan et al. [2004] suggested that all markers be tested on a portion of subjects in Stage 1, followed by evaluating the most significant markers on the remaining subjects in Stage 2. Skol et al. [2006] extended the design of Satagopan et al. by combining the two samples in Stage 2 and jointly analyzing them using the most significant markers found in Stage 1. The Genetic Analysis Workshop 16 (GAW16) datasets offered the opportunity for examining new or existing multistage methods on complex phenotypes in the context of a GWAS.

METHODS

The six studies performed by the GAW16 Group 3 participants represent an interesting diversity of multistage methods for increasing the analytic efficiency of large GWAS. Five authors analyzed the rheumatoid arthritis (RA) phenotype in the North American Rheumatoid Arthritis Consortium (NARAC) dataset. Sung et al. [2009] used the simulated Framingham Heart Study dataset to analyze the 200 replicates of the low-density lipoprotein (LDL) phenotype collected at the first visit. Table 1 displays a summary of the GAW16 contributions by Group 3 published authors.

Stage 1 Methods

While the first stage in four of the studies [Cho et al., 2009; Niu et al., 2009; Wang et al., 2009a; Wu et al., 2009] had the same goal, namely to eliminate the noise of non-significant SNPs in preparation for Stage 2, this first step was performed in four different ways. Cho et al. [2009] and Niu et al. [2009] performed, respectively, a series of single-marker logistic regression and log-linear analyses to screen each SNP separately. Cho et al. [2009] included sex as a covariate; Niu et al. [2009] did not include any covariates. Cho et al. [2009] used only an additive model for their analysis while Niu et al. [2009] tested each SNP with three models: dominant, recessive, and additive after using the genomic control method [Devlin and Roeder, 1999] to control for spurious associations induced by population structure. The determination of which SNPs to include at Stage 2 was a practical decision for Cho et al. [2009]: they selected the largest number of SNPs that could be handled in their computationally intensive penalized logistic regression analysis in Stage 2. Out of the 48,336 SNPs with p -values ≤ 0.05 in Stage 1, this group chose the top 1,000, then 2,000,

and finally 3,000 to test in Stage 2. On the other hand, Niu et al. [2009] used a Bonferroni correction to determine each SNP's minimum p -values among the three models, and choose the 500 SNPs with the smallest p -values for inclusion in their Stage 2 analyses.

Wu et al. [2009] also used a sequence of three logistic regression models to screen the SNPs: a simple model regressing only on each SNP and, after removing the top 200 SNPs from the single-SNP analysis, two exhaustive searches using all pairs of SNPs, first without and then with a SNP-SNP interaction term. (The top 200 SNPs were removed to prevent a SNP with an extremely high marginal association from forcing significance on the two-locus model.) And finally, Wang et al. [2009a] used a two-stage sliding window approach with a likelihood-ratio test statistic (LRT_PC), which was developed using principal components derived from the subjects' genotypes. Based on a power analysis, 15% of the sample was allocated to Stage 1 in which the top 1,000 SNP windows were to be selected as determined by the LRT_PC statistic.

The design of the remaining two-stage analyses of Childers et al. [2009] and Sung et al. [2009] did not require the reduction of the number of SNPs in preparation for the Stage 2 analysis. Childers's research project was to compare the performance of BIMBAM [Servin and Stephens, 2007], IMPUTE [Marchini and Howe, 2008], and TUNA [Nicolae, 2006] on the NARAC data. The first step, which is automated within these three software packages, makes use of an outside reference sample such as HAPMAP to determine linkage disequilibrium (LD) patterns and genotype frequencies in the reference sample. Included in this initial stage is the imputation of untyped SNPs, which is determined using the genotypes of the subjects in the observed data and their correlation with the imputed SNPs.

Sung et al. [2009] were the only participants in Group 3 to use all 200 replicates of the Framingham Heart Study simulated family data (Table 1). They first performed a single-marker analysis using Merlin's family-based association software [Chen and Abecasis, 2007]. The benefits of Merlin's association test are that it imputes missing genotypes and ignores families with Mendelian inconsistencies, sparing the investigator time spent finding and removing such problem individuals before to analysis.

Stage 2 Methods

The second stage of the six studies involved computationally intensive methods, most of which are not practical on the entire genome. Three studies used a penalized regression model [Cho et al., 2009; Sung et al., 2009; Wu et al., 2009] either LASSO [Tibshirani, 1996] or the elastic net [Zou and Hastie, 2005]. Briefly, these two computer-intensive procedures are variable selection processes useful for regression procedures where predictor variables may be highly correlated, resulting in multicollinearity. The multicollinearity problem is resolved by the strategic elimination of predictor variables. Thus, both LASSO and elastic net can be thought of as automatic variable selection methods. Also note that both methods can be applied to datasets with more predictors than individuals (the small n , large p problem).

Sung et al. [2009] used LASSO to analyze 6,857 individuals and 4,589 SNPs with marginal effects identified in Stage 1. LASSO was also used by Wu et al. [2009] with a sample of 2,002 subjects and 914 SNPs, all with either marginal or interaction effects identified in Stage 1. The third group to use a penalized regression method was Cho et al. [2009], who employed the elastic net to analyze approximately 3,000 SNPs genotyped in 2,148 subjects. This group fit two models, one testing the marginal effect of each SNP with sex as a covariate, and a second model testing each SNP's marginal effect with the variable sex in the model plus a term reflecting an interaction between the SNP and sex. We note that neither LASSO nor the elastic net allow missing data. Accordingly, Cho et al. [2009] only

used individuals with complete genotypes; Sung et al. [2009] imputed the partially missing genotypes using Merlin [Chen and Abecasis, 2007], and Wu et al. [2009] imputed missing genotypes using Beagle [Browning and Browning, 2007].

Among the remaining three studies, Niu et al. [2009] analyzed 17 two-locus models for association with RA, forming all possible pairs among the 500 SNPs with the lowest p -values in their Stage 1 analysis. Significance was assessed by 1,000 permutation tests over cases and controls. Wang et al. [2009a], using only the second half of their divided sample, determined the significance of their top 1,000 SNP windows also with a permutation approach. The Stage 2 analyses of Childers et al. [2009] consisted of single-marker association studies computed by BIMBAM, IMPUTE, and TUNA on a dataset enlarged with the imputed genotypes of untyped SNPs. The statistical tests to determine significance differ for each software package: BIMBAM uses a Bayes factors approach; IMPUTE uses the Cochran-Armitage trend test; and TUNA tests the differences of allele frequencies between cases and controls.

RESULTS

The salient observation about the NARAC studies was that most findings, irrespective of method, once again confirmed the important association between variants on chromosomes 6 and RA. The roles of variants in the *PTPN22* gene on chromosome 1 were also confirmed by several of these studies. The three NARAC studies of Cho et al. [2009], Niu et al. [2009], and Wu et al. [2009] used Stage 1 regression-based analysis (e.g., log-linear or logistic-regression models) to reduce the noise of non-significant SNPs, thereby limiting the number of SNPs analyzed in Stage 2. The results from these three Stage 1 studies contained statistically significant *HLA-DRB1* variants, or chromosome 6 variants in regions that are in high LD with *HLA-DRB1* variants. Furthermore, SNPs on *PTPN22* on chromosome 1, a gene known to be an RA susceptibility gene, were also identified by these diverse studies. The first stage of the Wang et al.'s [2009a] sliding window approach was also designed to limit the number of SNPs analyzed in Stage 2. Although Wang et al. [2009a] did not specifically report their Stage 1 chromosomal distribution for their top 1,000 windows, their Stage 2 analyses made it clear that almost 40% of their non-overlapping windows (26 out of 68 non-overlapping windows) were on chromosome 6 in a region in high LD with *HLA-DRB1*. In Stage 1, Niu et al. [2009] found 184 SNPs that were significant, with a Bonferroni corrected p -value at 0.05. One SNP was on the *PTPN22* gene on chromosome 1, and the other 183 SNPs were in high LD with *HLA-DRB1* on chromosome 6. Over all, the four studies that analyzed the NARAC dataset using two distinct stages all found chromosome 6 SNPs with strong effects from their Stage 1 single-marker analysis. In addition, three of these four NARAC Stage 1 analyses identified a significant variant in *PTPN22*. The Stage 1, Stage 2 distinction is a bit fuzzy in the Childers et al. [2009] study of imputation methods. However, their three imputation methods consistently found most of their typed and imputed significant SNPs on chromosome 6.

In Stage 2, the elastic net analysis of Cho et al. [2009] identified 806 SNPs on chromosome 6 from the top 3,000 SNPs identified in Stage 1. The 26 windows were on chromosome 6 and were in high LD with *HLA-DRB1*. Wu et al. [2009] found 395 SNPs with significant marginal effect with Bonferroni p -value < 0.1 , most of which were also on chromosome 6p21.

The main purpose of using more sophisticated methods in Stage 2 was to identify additional genes that may harbor significant variants that cannot be found by simpler methods. In this regard, all five studies that analyzed the NARAC dataset searched for significant SNPs outside of chromosome 6. Cho et al. [2009] observed that of the 398 significant SNPs elastic

net found among the top 3,000 SNPs identified in Stage 1, only 66 SNPs were on chromosome 6. This contrasted with their single-marker analysis in which most of the significant SNPs resided on chromosome 6. Childers et al. [2009] found 44 significant SNPs not located on chromosome 6 using BIMBAM, 55 SNPs using IMPUTE, and 96 SNPs using TUNA. In particular, 53 of these SNPs from IMPUTE and TUNA were in common. The Stage 2 analysis of Niu et al. [2009] identified a gene-gene interaction between a SNP pair on chromosomes 6 and 15, significant at the 0.047 level after correcting for multiple tests. Wang et al. [2009a] found 42 out of 68 significant windows outside chromosome 6. Wu et al. [2009] found 71,693 SNP pairs of gene-gene interaction that were significant with Bonferroni p -value < 0.1 : these pairs included 18,391 unique SNPs, 89% of which were outside of chromosome 6.

Wu et al. [2009] provided three important findings by performing genome-wide gene-gene interaction in Stage 1. First, many of the individual SNPs among the top 208 SNP pairs with strong gene-gene interaction effects showed relatively low marginal effects. Second, 77% of these SNP pairs were either on different chromosomes or outside of LD bins (with $D' < 0.2$). Third, top SNP pairs with strong additive effects had relatively strong LD (61% of the top 494 SNPs had $D' > 0.2$). This suggests that, at least for the NARAC dataset, haplotype analysis may be useful for the SNP pairs with strong additive effects but not for purely interaction effects. In addition to their Stage 2 analysis, Wu et al. [2009] used LASSO to replicate some of their GAW16 findings in the Wellcome Trust Case Control Consortium (WTCCC) SNP dataset.

Finally, Sung et al. [2009] found that the major causal SNP, rs2294207, in the simulated Framingham Heart Study dataset was easily identified as the most significant SNP in almost all 200 replicates by the relatively straightforward Merlin association analysis. Although rs2294207 entered first into the LASSO solution path in 57% ($N=114$) of the replicated samples, in almost every other replicate, a highly correlated, but non-causal SNP would be selected by LASSO before rs2294207, thereby excluding this major SNP from the solution space. The correlation between the ranks of Merlin's p -values and the rank of entry into LASSO's solution path was about 0.07 over all 200 replicates.

DISCUSSION

Various multistage analytic designs of GWA data have been proposed as a practical solution to the cost of genotyping large numbers of SNPs in a large case-control or family-based dataset. No single such design has been declared to be the best way to analyze datasets that may contain millions of SNPs together with many thousands of subjects. As observed among the diverse two-stage designs presented by participants in Group 3, it is difficult to pinpoint any one of these methods as superior to any other, particularly since, with the exception of Sung et al. [2009], these analyses focused on real data for which the true associations are unknown.

The continued use of multistage design for GWAS can be questioned on several levels. First, the cost of genotyping has plummeted over the years and continues to do so. Coupled with the financial aspect is the increased power of computers, their lower cost, and the use of clusters of computers for analysis. Second, many have expressed the concern that by eliminating marginally non-significant SNPs in the first stage, we may be missing variants that are guiding the disease process when examined in the larger context of additional genetic and/or environmental risk factors.

The issue of correcting for multiple testing also plays a role in evaluating SNPs used in a multistage design, namely, how should the corrections be applied? The four manuscripts in

Group 3 that presented traditional multistage analyses [Cho et al., 2009; Niu et al., 2009; Wang et al., 2009a; Wu et al., 2009] used either a 5% Bonferroni or false-discovery rate correction for multiple testing in Stage 1 to determine which SNPs would be used in Stage 2. Similar statistical corrections were then independently employed in Stage 2. Although this seems to be the mode of operation in most two-stage analyses, it is unclear whether this is the optimal way to proceed or even if this is methodologically valid because Stage 2 results may not be independent of Stage 1 results. For example, if a SNP is significant in Stage 1 and therefore included in Stage 2, should the ‘prior’ significance of that SNP somehow be taken into consideration in Stage 2? Clearly, multistage analysis can be exploratory and hypothesis generating. However, we need to be aware of statistical problems such as multiple testing issues that may be invalid.

Among the five authors who analyzed the NARAC data, three tested for stratification in these data: Cho et al. [2009], Wu et al. [2009], and Niu et al. [2009]. However, the inflation factor in these data is 1.4 if stratification is ignored [Hinrichs et al., this volume]. As is well known, ignoring population stratification in association analysis can create false-positive results and mask true associations [Devlin and Roeder, 1999]. It is worth noting that any of the penalized regression methods can include the derived principal components as a covariate to control for population stratification, although this may not have been done among these authors.

Overall, this group of six papers offers an interesting mix of diverse methods for multistage genetic analysis. In particular, penalized regression techniques such as the LASSO [Tibshirani, 1996] or the elastic net [Zou and Hastie, 2005], considered as a generalization of the LASSO, look promising for projects of the magnitude of GWAS. These methodologies can be computer intensive, and there are many technical details not discussed in this summary that one should understand before using either of these regression methods. Given the growing evidence that genetic interactions are ubiquitous in determining susceptibility to common human diseases, the use of these penalized regression methods in future analyses of GWAS should be of interest.

Acknowledgments

We thank to the members of Group 3 for their enthusiasm in our group discussions and their thoughtful contributions to our presentation at the GAW meeting. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. YJ Sung was sponsored by GM28719 from the National Institute of General Medical Sciences.

REFERENCES

- Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007; 8:1084–1097. [PubMed: 17924348]
- Chen WM, Abecasis GR. Family-based association tests for genomewide association scans. *Am J Hum Genet.* 2007; 81:913–926. [PubMed: 17924335]
- Childers DK, Kang G, Liu N, Gao G, Zhang K. Application of imputation methods to the analysis of rheumatoid arthritis data in genome-wide association studies. *BMC Proc.* 2009; 3 Suppl 7:S24. [PubMed: 20018014]
- Cho S, Kim H, Oh S, Kim K, Park T. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc.* 2009; 3 Suppl 7:S25. [PubMed: 20018015]
- Devlin B, Roeder Kathryn. Genomic control for association studies. *Biometrics.* 1999; 55:997–1004. [PubMed: 11315092]
- Goldstein, David B. Common genetic variation and human traits. *The New England Journal of Medicine.* 2009; 360:1696–1698. [PubMed: 19369660]

- Hindorff, LA.; Junkins, HA.; Mehta, JP.; Manolio, TA. A catalog of published genome-wide association studies. Available at: www.genome.gov/26525384.
- Hinrichs AL, Larkin EK, Suarez BK. Population stratification and patterns of linkage disequilibrium. *Genet Epidemiol*. 2009 this volume.
- Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, Thomas G, Hoover R, Hunter DF, Chanock S. Beyond odds ratios – communicating disease risk based on genetic profiles. *Nature Reviews Genetics*. 2009; 10:264–269.
- Marchini J, Howie B. Comparing algorithms for genotype imputation. *Am J Hum Genet*. 2008; 83:535–539. [PubMed: 18940314]
- Nicolae DL. Testing untyped alleles (TUNA)-applications to genome-wide association studies. *Genet Epidemiol*. 2006; 30:718–727. [PubMed: 16986160]
- Niu A, Zhang Z, Sha Q. Application of seventeen two-locus models in genome-wide association studies by two-stage strategy. *BMC Proc*. 2009; 3 Suppl 7:S26. [PubMed: 20018016]
- Satagopan JM, Venkatraman ES, Begg CB. Two-stage designs for gene-disease association studies with sample size constraints. *Biometrics*. 2004; 60:589–597. [PubMed: 15339280]
- Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet*. 2007; 3:e114. [PubMed: 17676998]
- Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*. 2006; 38:209–213. [PubMed: 16415888]
- Sung YJ, Rice TK, Shi G, Gu CC, Rao DC. Comparison between single-marker analysis using Merlin and multi-marker analysis using LASSO for Framingham simulated data. *BMC Proc*. 2009; 3 Suppl 7:S27. [PubMed: 20018017]
- Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B*. 1996; 58:267–288.
- Thomas D, Xie R, Gebregziabher M. Two-stage sampling designs for gene association studies. *Genet Epidemiol*. 2004; 27:401–414. [PubMed: 15543639]
- Wang X, Qin H, Sha Q. Incorporating multiple-marker information to detect risk loci for rheumatoid arthritis. *BMC Proc*. 2009a; 3 Suppl 7:S28. [PubMed: 20018018]
- Wang X, Zhang S, Sha Q. A new association test to test multiple-marker association. *Genet Epidemiol*. 2009b; 33:164–171. [PubMed: 18720476]
- Wu Z, Aporntewan C, Ballard DH, Lee JY, Lee JS, Zhao H. Two-stage joint selection method to identify candidate markers from genome-wide association studies. *BMC Proc*. 2009; 3 Suppl 7:S29. [PubMed: 20018019]
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B*. 2005; 67:301–320.

Table 1

Summary of Group 3 Data and Methods¹

1 st Author	Regions analyzed	Stage 1	Stage 2	Methods of analysis	Basic results
Childers	515,050 SNPs on 22 autosomes	The software determines SNP LD patterns in the data and imputes untyped SNPs	Association tests are performed on imputed and genotyped SNPs	Compared imputation and analyses methods in BMBAM, IMPUTE, and TUNA software [Marchini et al., 2008; Servin et al., 2007; Nicolae, 2006]	The same SNPs were not identified in all 3 methods. All 3 methods found significant SNPs, most were on chr 6. Most significant untyped SNPs were near significant typed SNPs, especially for chr 6
Cho	474,499 SNPs on 22 autosomes	Screened all SNPs with logistic regression, additive model with sex as covariate	M1: Top 1K, 2K, and 3K Stage 1 SNPs were analyzed with the elastic net; M2: SNP-sex interaction terms included	Elastic net regularization [Zou and Hastie, 2005]	M1: 81 SNPs were commonly selected within the top 1K, 2K, and 3K sets of SNPs; 23 were also identified in Stage 1, and 20 of these are on chr 6. M2: 10 SNP-sex terms selected
Niu	531,689 SNPs on 22 autosomes	Screened all SNPs with 3 log linear models: dominant, recessive, additive	500 stages 1 SNPs with <i>p</i> -values ≤ 0.05 on any Stage 1 model analyzed for 17 2-focus models	Log linear models used for testing 17 2-focus models [Agresti, 2002]	One pair of SNPs, rs693589 (chr 6) and rs11634 (chr 15), were significant in a multiplicative model (Rec>Rec).
Sung²	5,011 SNPs on chromosome 22 in the known region of susceptibility SNPs	Screen all SNPs with MERLIN's family-based association test	Multi-marker analysis using the LASSO	Compared the LASSO vs. Merlin's family-based association test [Chen and Abecasis, 2007; Tibshirani, 1996]	rs2294207, the major SNP, significant by Merlin and LASSO. 12 polygenic SNPs not significant by either method
Wang	531,689 SNPs on 22 autosomes	Analyze 15% of the total sample using the LRT_PC statistic; chose top 1000 windows for Stage 2	Used the remaining 85% of the sample as a replication sample to analyze the top 1000 windows with LRT_PC	The likelihood ratio principal component (LRT-PC) statistic [Wang et al., 2009]	68 non-overlapping windows with nominal <i>p</i> -values (FDR adjusted). 26 windows were on chr 6 in high LD with <i>HLA-DRB1</i> . Additional significant windows on 17 other chromosomes
Wu	500,884 SNPs on 22 autosomes	Screen all SNPs with 3 logistic regression models: M1. single SNPs; M2. all pairs of SNPs; M3. interaction term added to M2	The LASSO model applied to the 914 highest ranked Stage 1 SNPs. WTCCC data used for gene validation	LASSO using the R software, glmnet [Tibshirani, 1996; Friedman et al., 1998]	63 nominally significant SNPs were identified, none from model 3, M3. The SNPs resided in genes <i>P7PN22</i> (chr 1), <i>HLA</i> genes (chr 6), and <i>C5</i> (chr 9)

¹ With one exception all authors analyzed the rheumatoid arthritis phenotype in the NARAC dataset. Sung et al. [2009] analyzed the 200 replicates of the LDL phenotype provided in the simulated Framingham data collected at Exam 1.

² Answers were known.