Published in final edited form as:
Genet Epidemiol. 2009 ; 33(Suppl 1): S105–S110. doi:10.1002/gepi.20481.

Gene- or Region-Based Analysis of Genome-Wide Association Studies

Joseph Beyene[1,2], **David Tritchler**[1,3,4], **Jennifer L. Asimit**[5], and **Jemila S. Hamid**[2]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

[2]Child Health Evaluative Sciences, The Hospital for Sick Children Research Institute, Toronto, Ontario, Canada

[3]Division of Epidemiology and Statistics, Ontario Cancer Institute, Toronto, Ontario, Canada

[4]Department of Biostatistics, State University of New York at Buffalo, Buffalo, NY

[5]Samuel Lunenfeld Research Institute of Mount Sinai Hospital, Toronto, Canada

Abstract

With rapid advances in genotyping technologies in recent years and the growing number of available markers, genome-wide association studies are emerging as promising approaches for the study of complex diseases and traits. However, there are several challenges with analysis and interpretation of such data. First, there is a massive multiple testing problem due to the large number of markers that need to be analyzed, leading to an increased risk of false positives and decreased ability for association studies to detect truly associated markers. In particular, the ability to detect modest genetic effects can be severely compromised. Second, a genetic association of a given single-nucleotide polymorphism as determined by univariate statistical analyses does not typically explain biologically interesting features and often requires subsequent interpretation using a higher unit such as a gene or region, for example as defined by haplotype blocks. Third, missing genotypes in the data set and other data quality issues can pose challenges when comparisons across platforms and replications are planned. Finally, depending on the type of univariate analysis, computational burden can arise as the number of markers continues to grow into the millions. One way to deal with these and related challenges is to consider higher units for the analysis such as genes or regions. This paper summarizes analytical methods and strategies that have been proposed and applied by Group 16 to two genome-wide association data sets made available through the Genetic Analysis Workshop 16.

Keywords

rheumatoid arthritis; case-control data; family-based study

INTRODUCTION

It is believed that genetic factors play an important role in the etiology of common diseases and traits. Over the last several decades, various molecular epidemiologic approaches have been developed and used to dissect genetic contributions for a host of diseases. Most notably, linkage and association studies that are based on candidate genes have been extensively studied. With advances in high-throughput genotyping technologies in recent

Corresponding Author: Joseph Beyene Dalla Lana School of Public Health University of Toronto, Health Sciences Building 155 College St, Toronto, ON M5T 3M7 Canada Telephone: 416-813-7654 x 2302 joseph@utstat.toronto.edu .

years, however, there has been a shift towards a genome-wide interrogation of common polymorphisms in relation to possible links with clinical phenotypes. It is now possible to study one million single-nucleotide polymorphisms (SNPs) on thousands of subjects; the number of markers is expected to grow further.

Two study designs are commonly used in the context of genome-wide association studies (GWAS). The most popular is the case-control design, which includes unrelated individuals. Several case-control GWAS have recently been published reporting disease susceptibility loci for a range of complex and chronic diseases [Hirschhorn and Daly, 2005; McCarthy and Zeggini, 2009]. The other study design is family-based, in which related subjects are included [Wilk et al., 2009]. Each study design has its own strengths and weaknesses. For example, the case-control design is easier to carry out compared with family-based studies, but it is prone to bias arising from population stratification.

Although GWAS are promising to be key tools in the investigation of complex diseases, there are many analytic and interpretation challenges. It is customary to analyze the million or so SNPs one at a time and declare significance at a pre-specified genome-wide *p*-value threshold. Such univariate analyses are less than optimal for many reasons. First, a SNP-by-SNP analysis leads to the well known multiplicity problem, resulting in highly inflated risk of type I error and a decreased ability to detect modest effects. Second, the functional unit of interest is often the gene, not a single SNP. For example, biologists typically think of a gene when designing a replication study. Therefore, interpretation of results from a SNP analysis may lack biological insight. Third, in high-throughput genome-wide data, there may be a significant proportion of missing genotypes. This and other data-related issues can pose difficulty when cross-platform comparisons and replications are planned. Finally, there is computational burden due to the sheer volume of the data, which is expected to grow even further as technology advances.

This paper summarizes contributions to Group 16, the “gene- or region-based association tests” group of the Genetic Analysis Workshop 16 (GAW16). Authors in this group contributed methods and strategies that can help alleviate some of the challenges described above. Contributors were interested in statistical and computational techniques for the analysis of high-dimensional SNP data using a gene or region as the primary unit of analysis.

Two genome-wide association data sets were analyzed: a case-control study on rheumatoid arthritis (RA) from the North American Rheumatoid Arthritis Consortium (NARAC) and a data set on traits related to cardiovascular conditions using the family-based Framingham Heart Study (FHS). Both data sets were made available through the GAW16.

METHODOLOGICAL APPROACHES AND ISSUES

Study focus

The scope of the analyses varied from candidate regions (e.g., HLA versus non-HLA region) to genome-wide searches. Analyses to detect marginal effects of features were performed, and some analyses included interactions. One study specifically focused on rare variants.

Pre-processing approaches and challenges

Eight groups analyzed the Problem 1 RA data and two groups analyzed the Problem 2 FHS data. Although within each problem a single genotyping platform was used, one group compared their results with the Wellcome Trust data, which raised the issue of combining and comparing data from different platforms with differing coverage.

Quality control approaches applied by the participants included removing SNPs that were not in Hardy-Weinberg equilibrium had low call rates, or very small minor allele frequency (MAF). Example criteria are Hardy-Weinberg equilibrium declared not to hold for $p < 5.7 \times 10^{-7}$, 10^{-10}, or 10×10^{-5}; call rates at least 97% or 90%; and MAF of 0.01 or more.

When SNPs are combined to form region-based features, the handling of missing values must be considered. It is undesirable to fail to define an entire region due to a few missing member SNPs. Imputation of missing SNPs becomes very important. Imputation is also important when combining different platforms.

Software used by members of our group includes R, Stata, C++, SAS, PLINK, GenABEL, Haploview (allele frequency estimation), GADA (copy number variation), FastPhase, MACH (missing value imputation), and Eigenstrat (outlier detection).

Defining regions and analytical strategies

The definition of "region" was very broad, to include a wide variety of multi-allelic analyses. The minimum region was a single SNP analyzed in a model including known effects (e.g., HLA) and interactions. In this case, the region is the selected SNP in conjunction with the known effect. Haplotype was also used as a region. A common theme was assembling SNPs to represent a gene. With multiple causal variants in the same gene, we would expect several SNPs to be associated with disease, and hence to observe a strong gene-based signal of association. For those applications, defining genes and mapping SNPs to genes are important issues. The next level of complexity is the gene set, which also requires biological knowledge to group genes by function or pathway. An alternative is to simply group SNPs annotated with similar function, making mapping to genes unnecessary. Another biologically motivated region definition used was interval of constant copy number.

Some groups defined regions based solely on statistical properties. Regions of significant SNPs were identified by a scan statistic approach, which requires the SNP position and the *p*-value for association at that position. Windows along the chromosome including varying numbers of SNPs were tested for region-level significance, where the regional *p*-value is the probability of observing the same number of significant markers over a distance as short as or shorter than observed. The scan statistic is simply the distance spanned by the group of markers of interest.

Another approach identified clusters of SNPs within a candidate region. The variation within the cluster was summarized by principal components. Other authors clustered individuals according to genotypic similarity. Cluster membership then represented a genotypic profile analogous to a haplotype.

The definition of region has implications for what is found. A single variant with large effect is more likely found by SNP-based analysis, while gene-based analysis is more likely to find multiple causal SNPs. These SNPs might be related by membership in a common gene, pathway, or function. They may share only rarity, and the need to be consolidated for detection.

An important aspect of gene-based analysis is that biological knowledge (e.g., GO and KEGG) can be used to interpret genes. For cross-platform comparisons, gene-based interpretation is easier to generalize. An important consideration is different coverage of the platforms. Imputation may help in some cases. A central issue in these analyses is the mapping of SNPs to genes. A variety of databases are available for mapping, and there is flexibility in defining the extent of genes. For example, up- and down-stream segments may

be included to represent regulatory or other components. The databases used were NCBI, UCSC, Illumina, and the CHIP Bioinformatics tools from the University of Florida.

Reducing the number of features was a common objective. Composites of SNPs can potentially represent biological variation more appropriately for the goals and interpretation of the study, so the true dimension is less than the number of SNPs. The association of the derived feature with outcome may be readily detected and more directly related to existing biological knowledge. The reduced number of features also reduces the adverse effect of multiple testing corrections on power.

There are two aspects to defining composite features. The first is how SNPs are grouped and the second is how the SNPs in a group are combined. The SNPs can be grouped by *a priori* biological knowledge (e.g., mapping them to genes) or by statistical properties (e.g., cluster, *p*-value, copy number segment). Types of biological knowledge employed included gene, pathway, candidate gene selection, and selection for low allele frequency. The statistical measures were intended to represent meaningful biological variation or to group effects for greater power.

The effect of a group of SNPs comprising a feature can be represented by combining test statistics for the SNPs (e.g., minimum, maximum, linear combinations of test statistics) or combining *p*-values. Principal components were used to summarize the variation of a group of SNPs. For that approach, the choice of the number of components is an important issue. Which SNPs are captured by a principal component may be of interest if interpretation is at a finer level than the gene. SNPs that are associated with disease as a result of linkage disequilibrium (LD) with a single causal variant would be expected to appear in the same principal component, while independent mutations would be expected to appear in different components. Similarly, clusters of SNPs are expected to capture independent biological effects. In some cases, a tagSNP represented the group.

Features were modeled in univariate analyses or jointly, possibly including interactions and covariates. Statistical models used included logistic regression, penalized logistic regression, and linear regression. Validation was based on replication of known genes and comparison with Wellcome Trust data.

HIGHLIGHTS OF INDIVIDUAL CONTRIBUTIONS

In this section, we briefly highlight specific approaches and findings for the ten contributions in our group. As is typically the case in genome-wide studies, some sort of pre-processing or quality filtering techniques were used by all of the ten contributions. The pre-processing steps included assessment of call rate, Hardy-Weinberg equilibrium, and minor allele frequency. Table I shows a summary of data sets and methods used by various authors. Some of the significant findings are also listed in the table.

Analyses of Problem 1 (case-control RA data)

It has been indicated that the RA data set considered in these studies is likely affected by population stratification, which may lead to misleading association results if this is not taken into account in the analyses. Two groups [Beyene et al., 2009; Morris et al., 2009] adjusted for population stratification. Morris et al. [2009] also performed association studies with and without adjustment for the effects of the *HLA-DRB1* locus.

Unlike the traditional GWAS in which SNPs are the units of analysis, all contributions in this group used higher units, and hence regions had to be defined before the association analyses. Most groups that analyzed the RA data mapped each SNP to a gene according to

the gene annotation attached with the data set. Black and Watanabe [2009], however, used orthoblique principal-component analysis to define sets of SNPs representing the unit of analysis.

After defining the unit, all groups implemented data reduction techniques with the aim of aggregating summary measures over all the SNPs or units for a given gene. Yang et al. [2009], for instance, combined marginal *p*-values from Armitage trend tests or logistic regression models to examine gene effects. They used a truncated product *p*-value approach in which *p*-values less than some pre-specified threshold were combined to evaluate the effects of genes (or SNP clusters). Aporntewan et al. [2009] used rough set theory to determine sets of SNPs that are informative and used this information in clustering individuals. Association analysis between cluster membership and disease outcome was then performed to determine significantly associated genes. Beyene et al. [2009] used the maximum explained variation from a logistic regression as well as the maximum chi-square statistic for evaluating the significance of pre-defined gene sets/path ways using gene set enrichment analysis (GSEA). Black and Watanabe [2009] used logistic and linear models in a likelihood-ratio framework to test whole-region association with RA status and RA-related traits, and also for individual cluster association with these outcomes. This highlights the main difference between their method and traditional PCA regression. Buil et al. [2009] performed association between clusters of individuals, defined by genetic similarity in a gene, and traits. Morris et al. [2009] carried out a logistic regression approach to identify accumulations of rare variants within the same gene associated with disease susceptibility. The log-odds of disease is modeled as a linear function of the proportion of rare SNPs within a gene. On the other hand, Qiao et al. [2009] and Xiao et al. [2009] considered interactions between genes and known RA-associated genes. Xiao et al. [2009] investigated the interaction between a particular gene (*KCNB1*), previously shown to have a moderate association with RA, with *HLA-DRB1*. They selected 15 SNPs from the *KCNB1* gene and fitted a logistic regression including interaction. Qiao et al. [2009] conducted genome-wide searches for RA-associated interactions with two known genes (*HLA-DRB1* and *PTPN22*). They used a gene-based measure of interaction defined by aggregating SNP-level information based on genotypetrait distortion statistics.

Some common genes, including genes in the HLA region, were consistently identified by the different groups. Known RA-associated genes as well as novel genes/regions have also been identified. Morris et al. [2009], for instance, identified novel putative RA susceptibility genes that have not been previously identified in large-scale GWAS. Some of these findings for both RA and FHS data sets are summarized in Table I.

Analyses of Problem 2 (family-based FHS data)

Only two contributions used the FHS data set. A scan statistic was used by Asimit et al. [2009] to identify regions of association with the blood lipid phenotypes high-density lipoprotein, low-density lipoprotein, and triglyceride. Markers from the 500k chip were pruned for LD ($R^2 < 0.5$), and the tagSNP positions and corresponding *p*-values for association were then used as input to the scan statistic to identify regions and test for regional significance. Permutations of the tagSNP *p*-values across positions were used to obtain empirical genome-wide regional *p*-values. Among the genome-wide significant scan statistic regions, there was overlap with a number of previously identified candidate lipid genes. Results were compared with those of several multiple-SNP regression test statistics in gene and inter-gene regions formed using the USCS database, using generalized estimating equations to account for familial correlation. Approximately half of the genome-wide significant scan statistic regions did not overlap with the SNP-database regions, and were considered to be novel.

Shtir et al. [2009] performed a gene-based approach to test association between diabetes and copy number variation (CNV) using the 500k SNP data for the subset of individuals with diabetes. Intensities were generated from .cel files via the Affymetrix Power Tools (APTools) software, and copy number inference was done using two CNV-calling methods: GADA and CNAM. The GADA analysis employs a median normalization step followed by the normalization scheme within the APTools suite, while the CNAM analysis uses a proprietary normalization scheme for which details are not available. Due to familial structure, the points where changes in copy number occur are likely to be correlated across individuals, so they partition the genome into intervals such that the copy number remains constant for any individual in the sample within any interval. They found little evidence of association, with no gene attaining genome-wide significance, which may be due to the Affymetrix 500k product not being designed very well for CNVs. In addition, there is a lack of agreement between the two CNV algorithms, which seems to be due to the different normalization techniques for the SNP intensities.

DISCUSSION

A common thread among the region-based analyses is the issue of how to handle SNPs that are not mapped to genes. One approach is simply to ignore such SNPs. This would be valid when the focus is on coding variation because there would be no loss of information in removing inter-gene SNPs [Buil et al., 2009; Morris et al., 2009; Qiao et al., 2009; Shtir et al., 2009; Xiao et al., 2009]. Another approach is to map each SNP to a gene by an approximation that assigns the gene that is physically the nearest or by LD with markers in nearby genes [Aporntewan et al., 2009; Beyene et al., 2009]. Moreover, SNPs may be stratified and classified as “gene” and “non-gene” [Asimit et al., 2009]. Upon stratification, each “non-gene SNP” may then be analyzed separately [Yang et al., 2009], or inter-genic SNPs may be grouped in conserved blocks to form post-transcription regulators. An alternative approach is to identify non-gene regions statistically, which removes some dependence on a gene database [Asimit et al., 2009; Black et al., 2009].

Groups met various issues in performing region-based analyses. In single-SNP analyses, for instance, there is a clear difficulty in replication studies due to the lack of common SNPs across platforms. Although this problem is reduced when the unit of analysis is a gene, it is still encountered, as was the case for Aporntewan et al. [2009] when comparing their GAW16 RA results with those of Wellcome Trust. In mapping SNPs to genes, there is the possibility of misclassification caused by making an incorrect assignment of a gene. There may also be disagreement between studies due to the way a gene is defined. Many groups used a common gene database, but defined genes differently, stretching the gene boundaries $\pm L$ kb, where L was 0, 5, or 50. Different databases and different versions of databases may assign different genes to a single SNP, adding a source of difficulty in comparing candidate genes from multiple studies. For example, the same SNP effect was detected by Buil et al. [2009] and Aporntwean et al. [2009], but the respective assigned gene was *PHF19* and *TRAF1*, likely due to the use of different databases (NCBI [Buil et al., 2009] versus Illumina [Aporntwean et al., 2009]). Additional sources of ambiguity in gene assignment are that multiple genes may code for a single SNP, SNPs in LD may be from several genes, and regions may include multiple genes. Furthermore, statistically defined regions or SNP sets may or may not overlap with genes. When performing a rare variant analysis, there is the additional issue of platforms not having many rare variants; in particular, the Affymetrix 500k product is not designed very well for rare variants. In the case of the GAW16 RA data, Morris et al. [2009] found a scarcity of variants with MAF < 2% in the chip, and relaxed the definition of rare variants to those with MAF < 5%. The strongest association signal, from the rare variant scan for RA adjusted for sex and *HLA-DRB1*, was in the class III region of

the MHC, in two overlapping genes that share the same three rare variants, suggestive of RA association with rare variants within the MHC, independent of the effect of *HLA-DRB1*.

In choosing an approach for analysis, the type of data needs to be considered, as well as the feasibility of meeting certain assumptions. Originally, principal-components analysis was developed for and applied mainly to quantitative variables, but it may also be used to analyze discrete SNP data. Distributional assumptions are required in the variance-components models and for the scan statistic. The high dimensionality of the SNP data leads to the problems of intensive computation and multiple testing, both of which may be alleviated by reducing the dimension, as done by using region-based methods. Additional issues arise depending on whether the focus of the study is exploratory in nature and involves model fitting. When model fitting, in addition to meeting various assumptions, decisions need to be made regarding choice of marginal and/or interaction effects and associations of single or multiple SNPs. Likewise, among the groups using the RA data in exploratory analysis, the choice of investigating the MHC region and/or regions outside of the MHC was an issue.

Biological knowledge is required at some stage, and is utilized either at the stages of defining regions, modeling and interpretation, or for later follow-up upon identification of candidate genes. Statistical and computational methods were combined with biological domain knowledge (e.g., GO and KEGG) in the pathway-based analysis of Beyene et al. [2009]. As indicated by several groups, after a list of candidate genes is obtained, further exploration is needed to understand variant effects on issues such as transcriptional regulation and protein production.

In comparisons between SNP-based tests and gene-based tests, it is apparent that the two tests use genetic information in different ways, and will not necessarily produce compatible results. Buil et al. [2009] found that SNP-based tests and gene-based tests gave different results outside the HLA region; the *P2PN22* gene was found by the SNP-based test, while the *PHF19* gene was found by the gene-based test. They conclude that the genetic information is used differently by the two types of tests, and that the tests are complementary. Different underlying genetic architectures may more easily be captured by one strategy than by the other, and in choosing the most appropriate approach, the hypothesis of interest must be considered. A SNP-based approach would be appropriate when the focus is on genes with a single common functional variant, but for genes with several common functional variants, a gene-based approach is a better choice. A specific test for rare variants is best when the primary interest is in genes with several rare functional variants. Clearly, all situations cannot be covered by a single type of test. However, one strategy may give better results, depending on the genetic architecture of the trait under consideration. The complication is that the underlying genetic architecture is often unknown.

Acknowledgments

We thank all members of GAW16 Group 16 for interesting discussions and helpful comments. We particularly acknowledge Shelley Bull, Laura Almasy, and Paul Marjoram for their useful feedback and comments. This work was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Mathematics of Information Technology and Complex Systems (MITACS), Canadian Institutes of Health Research (CIHR, grant 84392), the Canadian Breast Cancer Foundation (CBCF), and Genome Canada through the Ontario Genomics Institute. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

REFERENCES

Aporntewan C, Ballard DH, Lee JY, Lee JS, Wu Z, Zhao H. Gene hunting of the Genetic Analysis Workshop16 rheumatoid arthritis data using rough set theory. BMC Proc. 2009; 3(Suppl 7):S126. [PubMed: 20017992]

Asimit JL, Yoo YJ, Waggott D, Sun L, Bull SB. Region-based analysis in genome-wide association study of Framingham Heart Study blood lipid phenotypes. BMC Proc. 2009; 3(Suppl 7):S127. [PubMed: 20017993]

Beyene J, Hu P, Hamid JS, Parkhomenko E, Paterson AD, Tritchler D. Pathway-based analysis of a genome wide case-control association study of rheumatoid arthritis. BMC Proc. 2009; 3(Suppl 7):S128. [PubMed: 20017994]

Black MH, Watanabe RM. A principal components-based clustering method to identify multiple variants associated with rheumatoid arthritis and arthritis-related autoantibodies. BMC Proc. 2009; 3(Suppl 7):S129. [PubMed: 20017995]

Buil A, Martinez-Perez A, Perera-Lluna A, Rib L, Caminal P, Soria JM. A new gene-based association test for genome wide association studies. BMC Proc. 2009; 3(Suppl 7):S130. [PubMed: 20017997]

Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005; 6:95–108. [PubMed: 15716906]

McCarthy MI, Zeggini E. Genome-wide association studies in type 2 diabetes. Curr Diab Rep. 2009; 9:164–71. [PubMed: 19323962]

Morris AP, Zeggini E, Lindgren CM. Identification of novel putative rheumatoid arthritis susceptibility genes via analysis of rare variants. BMC Proc. 2009; 3(Suppl 7):S131. [PubMed: 20017998]

Qiao B, Huang CH, Cong L, Xie J, Lo S-H, Zheng T. Genome-wide gene-based analysis of rheumatoid arthritis-associated interaction with *PTPN22* and *HLA-DRB1*. BMC Proc. 2009; 3(Suppl 7):S132. [PubMed: 20017999]

Shtir C, Pique-Regi R, Siegmund K, Morrison J, Schumacher F, Marjoram P. Copy number variation in the Framingham Heart Study. BMC Proc. 2009; 3(Suppl 7):S133. [PubMed: 20018000]

Wilk JB, Chen TH, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, Myers RH, Borecki IB, Silverman EK, Weiss ST, O'Connor GT. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. PLoS Genet. 2009; 5:e1000429. [PubMed: 19300500]

Xiao X, Zhang Y, Wang K. Association of *KCNB1* to rheumatoid arthritis via interaction with *HLA-DRB1*. BMC Proc. 2009; 3(Suppl 7):S134. [PubMed: 20018001]

Yang H-C, Liang Y-J, Chung C-M, Chen J-W, Pan W-H. Genome-wide gene-based association study. BMC Proc. 2009; 3(Suppl 7):S135. [PubMed: 20018002]

Table I

Highlights of methods used and selected significant association results

Dataset	Contribution	Methods	Significant associations
Rheumatoid arthritis	**Aporntewan et al.**	Rough set theory to determine sets of SNPs representing a gene, clustering of individuals, association between cluster membership and disease status	*TRAF1, PTPN22, ADAM15, AGPAT2*
	Beyene et al.	Measures of explained variation from logistic regression are used, the maximum summary measure over all SNPs for a given gene is used for association, gene set enrichment analysis is used to determine significance of pre defined sets of genes	Ten gene sets were identified, the MHC region was included in one of these pathways
	Black et al.	Principal-components analysis with orthoblique rotation is used to identify subsets of SNPs in the MHC region; principal-components-based clusters are then tested for association studies	*HLA-C, HLA-A, HLA DQB2, HLA-DR* (*DRA, DRB1* and *DRB5*)
	Buil et al.	Clustering of individuals based on genetic similarity, association tests between the clusters and the traits are performed	*PHF19*
	Morris et al.	Tests of association of RA with accumulations of rare variants within a gene are performed using logistic regression in which the phenotype is modeled as a function of the proportion of rare SNPs at which an individual carries minor alleles	*FRY, PRPSlL1, ARNTL1*, and *TRIM58* and *HINT1* when adjusted for *HLA-DRB1*
	Qiao et al.	A statistic for measuring gene-gene interaction is defined based on SNP-level interactions using genotype-trait distortion statistic; significance was determined using permutation	*MGC13017, HSPCAL3, MIA, PTPNSIL, IGLVI-70*
	Xiao et al.	SNPs are selected from the *KCNB1* gene, logistic regression with interaction is then used where an interaction term between *HLA-DRB1* and *KCNB1* is included in the model	Moderate interaction between *KCNB1* and *HLA-DRB1*
	Yang et al.	*p*-Values from single-SNP analyses are combined for a gene-based association analysis	*PTPN22, C5, IL2RB, HLA-DRA, BTNL2, C6orf10, NOTCH4, TNXB*
Framingham Heart Study	**Asimit et al.**	Based on the positions and association *p*-values of tagSNPs, a scan statistic is used to identify regions of significant SNPs. Empirical genome-wide *p*-values are obtained by permutation of the tagSNP *p*-values across positions	*ABCA1, CETP, LPL, LIPG, ACAA2, HERPUD1, LDLR, APOB, BCAM*
	Shtir et al.	A gene-based analysis is used to test for association between copy number variation and diabetes status. The Wilcoxon rank-sum test is used to obtain a *p*-value for each gene.	No gene attained genome-wide significance