Published in final edited form as:

Genet Epidemiol. 2009 ; 33(Suppl 1): S8–12. doi:10.1002/gepi.20465.

Genome-wide Association Studies for Discrete Traits

Duncan C. Thomas¹

¹Department of Preventive Medicine, University of Southern California, Los Angeles, CA

Abstract

Genome-wide association studies of discrete traits generally use simple methods of analysis based on chi-square tests for contingency tables or logistic regression, at least for an initial scan of the entire genome. Nevertheless, more power might be obtained by using various methods that analyze multiple markers in combination. Methods based on sliding windows, wavelets, Bayesian shrinkage, or penalized likelihood methods, among others, were explored by various participants of Genetic Analysis Workshop 16 Group 1 to combine information across multiple markers within a region, while others used Bayesian variable selection methods for genome-wide multivariate analyses of all markers simultaneously. Imputation can be used to fill in missing markers on individual subjects within a study or in a meta-analysis of studies using different panels. Although multiple imputation theoretically should give more robust tests of association, one participant contribution found little difference between results of single and multiple imputation. Careful control of population stratification is essential, and two contributions found that previously reported associations with two genes disappeared after more precise control. Other issues considered by this group included subgroup analysis, gene-gene interactions, and the use of biomarkers.

Keywords

rheumatoid arthritis; single-nucleotide polymorphisms; multi-marker associations; imputation; population stratification; gene-gene interactions; biomarkers

INTRODUCTION

The Genetic Analysis Workshop 16 (GAW16) Group 1 discussion group on discrete traits investigated a broad range of methodological issues relating to genome-wide association studies (GWAS). All contributions in this group used the Problem 1 data on rheumatoid arthritis (RA) from the North American Rheumatoid Arthritis Consortium [Amos et al., 2009; Plenge et al., 2007]. A theme common to more than half the contributions was whether to analyze single-nucleotide polymorphisms (SNP) one-at-a-time or jointly—either for multiple SNPs in a region or genome-wide—and if jointly, how. Other themes addressed by multiple contributions included methods for imputation, population stratification, subgroup analyses, gene-gene (G×G) interactions, and the use of biomarkers. Some of these issues are common, with several other groups addressing methodological problems in GWAS data for other kinds of endpoints (see especially the reports in this volume from Group 4 for haplotype-based analysis [Hauser, 2009], Group 9 on G×G interactions [An et al., 2009], and Group 13 on population stratification [Hinrichs et al., 2009]). We begin, however, with a brief discussion of a fundamental issue raised by two contributions

concerning what parameter should be used to describe the associations with a given locus and will conclude with a brief summary of some of the substantive results.

Choice of parameter

Most GWAS have relied on a strict frequentist genome-wide significance level (e.g., $p < 10^{-7}$) to select a subset of SNP associations worthy of reporting. The magnitude of the corresponding effect, however, depends upon a combination of sample size and minor allele frequency. It is noteworthy that most of the associations reported from GWAS to date have tended to be small, with relative risks (RR) ranging from 1.2 to 1.5 (see, for example, the Catalog of Published Genome-Wide Association Studies at <http://www.genome.gov/gwastudies/> [Hindorff et al., 2009]). It has also been frequently pointed out that in the aggregate, all the published SNP associations to date still account for only a small proportion of the estimated heritabilities for many complex diseases. The “dark matter” remaining to be discovered is a mystery, possibly comprising some combination of rare variants, structural variants, interactions (gene-environment or G×G), epigenetic effects, or other as yet unknown mechanisms.

In this vein, Lorenzo Bermejo et al. [2009] compared rankings of SNPs by Bayes factors and by attributable sibling relative risks λ_s by simulation and by application to the GAW16 Problem 1 RA data for 12 regions selected on the basis of previously reported associations in the Wellcome Trust Case-Control Consortium data. Bayes factors were computed by first performing single-SNP associations for all SNPs on a given chromosome and computing their mean absolute deviations; these were then used as priors in a Bayesian logistic regression for the SNPs in the candidate regions on that chromosome. Their analyses suggested that such a two-dimensional classification might produce a higher yield of true-positive signals than those based on any single statistic such as a summary p -value.

Rather than focusing on detection of loci, Meyer, Jeffries, and Zheng (unpublished) considered estimation of effect sizes, in particular the absolute penetrances. Starting with the well known observation that absolute penetrances cannot be estimated from case-control estimates of RRs without knowledge of the marginal population rates, they made the surprising observation that under a constraint about the dominance parameter, it was possible to estimate the baseline risk (and hence the absolute penetrances) from the estimated homozygote and heterozygote RRs. Although mathematically correct, the result is of limited utility in practice because the dominance parameter cannot be specified independently from the two RRs.

METHODS FOR ANALYSIS OF SNP ASSOCIATIONS

Single-SNP methods

Two papers described approaches to using the full 2×3 contingency table data for a single-SNP genotype by case-control status to derive various tests of association. Chen et al. [2009] compared allelic tests based on collapsing the counts into a 2×2 table of 2N alleles (where N is number of subjects) and applying the standard chi-square test (equivalent to the score test for logistic regression under a log-additive model) with the Hotelling’s T^2 test (essentially equivalent to the standard 2-df chi-square test for a codominant model, but using an F -distribution for finite samples rather than the asymptotic chi-square distribution). In general, most SNP associations detected by the allelic coding in T^2 test were detected by the genotype coding and vice versa.

Matthews et al. [2009] used a recently developed likelihood ratio test FP introduced by Zhang et al. [2008] that exploits information in the 2×3 table about inbreeding in addition to the standard test of association. This is accomplished by constraining the possible genotype

frequency differences between cases and controls based on possible probabilities F that two alleles are identical by descent and p the frequency of one of the alleles. They found that the FP test generally led to more significant findings than the standard 2-df chi-square test, based on findings for non-HLA-DR β 1 associations (stratified by shared epitope status, as described below), adjusting for multiple comparisons using Bonferroni correction, q -values, or permutation-based p -values.

Multiple-SNP methods

The remaining contributions in this subgroup used a variety of multi-marker methods, several quite novel. Some were applied to specific candidate regions, some genome-wide; some were frequentist, some Bayesian.

Frequentist methods—We begin with four contributions using a range of frequentist approaches, each novel in its own way. Sliding window approaches using either SNPs or haplotypes have been used previously in GWAS by numerous authors. The novel element in the contribution by Sha et al. [2009] was the introduction of an adaptive window size using principal-components (PC) analysis to select the optimal window size at each location. A novel bisection algorithm is used to search for window sizes for which the first few PCs account for the majority of the variation in the SNP data for that window. This is then followed by a standard score test from a logistic regression on these PCs. Significant windows are adjusted for multiple testing using Bonferroni correction.

Zhang et al. [2009] utilized their proposed penalized orthogonal-components regression (POCRE) algorithm to construct sparsely loaded PCs. Unlike traditional PC analysis, POCRE employs a supervised dimension reduction in the PCs construction and enables variable selection through penalization. The penalization is implemented using an empirical Bayes method, which makes the algorithm adaptive to sparseness of phenotype-related SNPs. (We consider this a “frequentist” method because the inference is still likelihood-based, even though the components are constructed to have sparse non-zero loadings via empirical Bayes thresholding estimates.) For application to case-control data, they introduced a variant of the approach replacing linear regression on the PCs by a linear discriminant analysis (POCRE-LDA). They concluded that their adaptive-penalty method performed better than existing penalization methods such as LASSO or ridge regression.

There have been a variety of flexible modeling methods introduced recently, one of which is based on the idea of “wavelets.” Rather than using splines as basis functions, wavelets begin with a wavelet transform of the data into a sequence of wavelet coefficients, and then apply shrinkage to the wavelet coefficients so as to obtain a sparse representation of the data, eliminating noise. The modified wavelet coefficients are transformed back to “genotype” data, which are then used in a score test for association with disease. By exploiting the spatial ordering of the SNPs in this manner, Jiang et al. [2009] showed in a separate simulation study that this approach increased the power of the test.

Bayesian methods—The LASSO penalty mentioned earlier is based on a double exponential penalty function $\exp(-\lambda \sum_i |\beta_i|)$ on the vector β regression coefficients for association with each SNP rather than standard normal prior $\exp(-\lambda \sum_i \beta_i^2)$ used in ridge regression. This has the advantage of not only shrinking all estimates of the regression coefficients towards zero, but also shrinking many of them all the way to zero, yielding a form of variable selection to obtain a more parsimonious model. It can also be highly computationally efficient if only a few terms remain, thereby making GWAS applications feasible with orders of magnitude more variables than subjects being considered jointly. These two alternatives were compared by Wang et al. [2009]. Mathew et al. [2009] explored

a more flexible two-parameter approach using a normal-exponential-gamma prior as a penalty function introduced by Hoggart et al. [2008] that includes both the normal and the LASSO as special cases. (See also Zou and Hastie [2005], who call this the “elastic net.”) Although not explored explicitly by Mathew et al. [2009], this approach has recently been extended to consideration of SNPs within genes and genes within pathways, allowing for SNP selection for the former and shrinkage for the latter (Chen, Peters, and Hsu, unpublished).

A rather different approach was developed by Kwon et al. [2009] using the singular value decomposition in a Bayesian manner, assuming a probit model for disease in relation to the PCs of the SNP data. The model is fitted using Markov-chain Monte Carlo methods to sample iteratively from the linear predictor in the probit model and the PCs.

The various methods described above are not an exhaustive list of those that have been proposed for use in GWAS (see [Wu et al., 2009] for a review of other literature). Chapman and Whittaker [2008] had found by simulation that either multiple single-marker tests or an empirical Bayes test [Goelman et al., 2005] were generally the most powerful, and more powerful than either logistic regression assuming a common effect (the L-C test) or the weighted score test of Wang and Elston [2007]. However, in unpublished work prior to GAW16, Wei Pan found by simulation that at least for some simulated models, the L-C test could be more powerful. Zhong and Pan (unpublished) further compared the performance of these methods on several well established RA-associated genes or regions in the GAW16 RA data and concluded that, while no test was uniformly superior to all alternatives, the L-C and empirical Bayes tests generally performed well.

OTHER METHODOLOGICAL ISSUES

Methods of imputation

The Mayo clinic group [Biernacka et al., 2009; Fridley et al., 2009] provided two closely related contributions on the subject of genotype imputation. The first compared several different programs that are widely used for this purpose—IMPUTE, MACH, PLINK, and fastPHASE—under two scenarios, one in which some genotypes were removed for all subjects and imputed based on external linkage disequilibrium (LD) information, the other in which some genotypes were removed from one subset of the data and others from the remaining subjects to mimic the situation of a combined analysis of studies using different platforms. They concluded that, while all methods performed well in regions of high LD, MACH and IMPUTE generally produced lower imputation error rates and more reliable association tests than fastPHASE or PLINK when using only the best genotype calls. Association tests based on MACH imputation provided results closer to tests based on complete data. Some of these programs also yield genotype probabilities that can be used in association tests, as described in the second of these two papers. Specifically, they compared single imputation (using the “best” genotype call) with multiple imputation using posterior genotype probabilities in both scenarios and found little difference between the two approaches. They questioned whether the additional effort and computation burden was worth it, except perhaps in regions where imputation produced more significant results. Although getting MACH to run efficiently on large sets of markers can be challenging, one strategy is to estimate recombination rates first, then use these estimates for imputation; an implementation of this strategy took two days to run using up to 15 GB of memory (Sinnwell et al., unpublished).

Population stratification

In the initial report of an association of *TRAF1-C5* with RA, Plenge et al. [2007] found a variance inflation factor (VIF, commonly denoted λ) for the RA data of 1.44, an unusually

large value that would be expected to produce numerous false-positive associations in a GWAS. Using PLINK to cluster individuals into ancestral groups reduced the VIF to 1.14, after which Plenge et al. [2007] applied genomic control to address residual confounding. Sarasua et al. [2009] contrasted this approach with the use of the first 10 PCs of about 82k SNPs with low LD as predictors in a logistic model for disease; after stratification into five subgroups based on quantiles of the risk score, the VIF was reduced to 1.034. With this fine control of population stratification, no other loci outside the HLA region maintained genome-wide significance, including the *TRAF1-C5* locus reported by other groups, raising the possibility that this could be a false-positive result due to population stratification (or that results had been over-corrected).

The *FP* statistic discussed above [Matthews et al., 2009] can be seen as a form of adjustment for the closely related problem of cryptic relatedness, rather than cryptic stratification, i.e., a situation in which individual cases and controls are not independent (as assumed in the standard chi-square test) due to unknown genetic relationships between them. However, the differences in estimated inbreeding coefficients between cases and controls used in the *FP* test may not actually be measuring relatedness but could arise from non-comparable control groups or differential genotyping error. This calls into question the conventional wisdom of excluding markers that are out of Hardy-Weinberg equilibrium in controls [Zou and Donner, 2006]. Nevertheless, it has been noted that a substantial proportion of the most significant findings in GWAS data are often false positives due to data errors, so it is only after careful quality control that it can be assumed that the most significant results are more likely to be true positives.

Subgroup analyses, G×G interactions, and use of biomarkers

Matthews et al. [2009] and Suarez et al. [2009] hypothesized that the powerful associations with *HLA DRB1* genotypes could obscure the ability to detect associations with other SNPs outside the HLA region or could modify their effect. Unfortunately, using the conventional “shared epitope” definition, the sample size of the low-risk group is too small to allow for tests of G×G interaction, but Suarez et al. [2009] found 279 SNPs that differed significantly at $p < 0.001$ between the high- and low-risk subgroups of cases (a case-only test of G×G interaction). All 279 of these SNPs were also significantly ($p < 0.05$) different between low-risk cases and low-risk controls, albeit based on small samples. Using the *FP* test discussed above, Matthews et al. [2009] found many SNPs that were significantly associated with rheumatoid factor in the high-risk subgroup, including some on chromosome 9 (e.g., *TRAF1-C5*), other genes in the HLA region, and a novel association with *UBXD2* on chromosome 2. They also applied linear regression to the biomarker measurements of rheumatoid factor IgM (RFUW) and anti-cyclic citrullinated peptide (anti-CCP) for the cases only, which have been found to co-occur with RA, but found no genetic associations with either endpoint.

SUBSTANTIVE FINDINGS

All groups that looked for it found extremely strong associations of RA with the MHC region. Table I summarizes the methods and substantive findings in other regions. Most groups that looked for associations with SNPs in the regions near the previously reported RA-associated genes *PTPN22* on chromosome 1 [Begovich et al., 2004] and *TRAF1-C5* on chromosome 9 [Plenge et al., 2007] generally confirmed these findings, using a variety of methods. However, Sarasua et al. [2009] noted that both of these associations disappeared after more careful control for population stratification using the method of Epstein et al. [2007], and Lorenzo Bermejo et al. [2009] were unable to confirm the *TRAF1-C5* association using Bayes factors. One novel association with a SNP near *UBXD2* on chromosome 2 (which plays a role in the immune system) was reported by Matthews et al.

[2009]. Numerous other regions, some containing plausible candidate genes, were identified by various groups and may merit further fine mapping. Although several chromosomes were identified several times (four hits on chromosome 10, three on chromosomes 4 and 18, and two on chromosomes 2 and 17), differences in reporting styles made comparison difficult; however, it appears that no single SNP, candidate gene, or region was identified by different groups.

CONCLUSIONS

In the summary presentation of Group 1, it was agreed that single-SNP analyses were useful as a first pass (methods are intuitive and well established; software is widely available and relatively quick to run), although multiple-SNP methods better deal with LD and complex diseases; missing data pose a challenge for these methods, however. Both frequentist and Bayesian approaches are promising, including methods based on adaptive sliding windows, wavelets, stochastic search variable selection, and various penalizations; the possibility of combining L_1 -norm SNP selection within genes or regions with L_2 -norm smoothing between regions is particularly appealing. Imputation methods were useful, but multiple imputation may not be needed. The importance of strong control for population stratification was highlighted. Subset analyses, although potentially informative, depend upon good study design; in particular for RA, the potentially interesting stratification by HLA genotypes (notably the shared epitope at HLA-DR β 1) was not particularly informative because of small numbers of those not carrying the shared epitope.

Acknowledgments

The Genetic Analysis Workshop is supported by grant R01 GM031575 from the NIH National Institute of General Medical Sciences. This paper is the result of valuable contributions from all of the participants of the GAW16 Group 1 discussion group. The author is grateful to Dr. Ellen Goode, who co-chaired this discussion group, led the presentation at the workshop, and edited the participant contributions. Dr. Thomas's work was supported by grant U01 ES015090 and P30 ES 07048.

REFERENCES

- Amos CI, Chen WV, Seldin MF, Remmers E, Taylor KE, Criswell LA, Lee AT, Plenge RM, Kastner DL, Gregersen PK. Data for Genetic Analysis Workshop 16 Problem 1, association analysis of rheumatoid arthritis data. *BMC Proc.* 2009; 3 Suppl 7:S2. [PubMed: 20018009]
- An P, Mukherjee O, Chanda P, Yao L, Engelman CD, Huang C-H, Zheng T, Kovac IP, Dubé M-P, Liang X, Li J, de Andrade M, Culverhouse R, Malzahn D, Manning AK, Clarke GM, Jung J, Province MA. The challenge of detecting epistasis (G×G interactions): Genetic Analysis Workshop 16. *Genet Epidemiol.* 2009 this volume.
- Begovich AB, Carlton VE, Honigberg LA, Schrod SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoerke JM, Conn MT, Chang M, Chang SY, Saiki RK, Catanese JJ, Leong DU, Garcia VE, McAllister LB, Jeffery DA, Lee AT, Batliwalla F, Remmers E, Criswell LA, Seldin MF, Kastner DL, Amos CI, Sninsky JJ, Gregersen PK. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (*PTPN22*) is associated with rheumatoid arthritis. *Am J Hum Genet.* 2004; 75:330–337. [PubMed: 15208781]
- Biernacka JM, Tang R, Li J, McDonnell SK, Rabe KG, Sinnwell JP, Rider DN, de Andrade M, Goode EL, Fridley BL. Assessment of genotype imputation methods. *BMC Proc.* 2009; 3 Suppl 7:S5. [PubMed: 20018042]
- Chapman J, Whittaker J. Analysis of multiple SNPs in a candidate gene or region. *Genet Epidemiol.* 2008; 32:560–566. [PubMed: 18428428]
- Chen L, Zhong M, Chen WV, Amos CI, Fan R. A genome-wide association scan for rheumatoid arthritis data by Hotelling's T^2 tests. *BMC Proc.* 2009; 3 Suppl 7:S6. [PubMed: 20018053]
- Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet.* 2007; 80:921–930. [PubMed: 17436246]

- Fridley BL, McDonnell SK, Rabe KG, Tang R, Biernacka JM, Sinnwell JP, Rider DN, Goode EL. Single versus multiple imputation for genotypic data. *BMC Proc.* 2009; 3 Suppl 7:S7. [PubMed: 20018064]
- Goelman JJ, van den Geer SA, van Houwelingen HC. Testing against a high dimensional alternative. *J R Statist Soc B.* 2005; 68:477–493.
- Hauser E. Haplotype-based analysis: A summary of GAW16 Group 4 analysis. *Genet Epidemiol.* 2009 this volume.
- Hindorf LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA.* 2009; 106:9362–9367. [PubMed: 19474294]
- Hinrichs AL, Larkin EK, Suarez BK. Population stratification and patterns of linkage disequilibrium. *Genet Epidemiol.* 2009 this volume.
- Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.* 2008; 4:e1000130. [PubMed: 18654633]
- Jiang R, Dong J, Dai Y. Genome-wide association study of rheumatoid arthritis by a score test based on wavelet transformation. *BMC Proc.* 2009; 3 Suppl 7:S8.
- Kwon S, Cui J, Rhodes SL, Tsiang D, Rotter JJ, Guo X. Application of Bayesian classification with singular value decomposition method in genome-wide association studies. *BMC Proc.* 2009; 3 Suppl 7:S9. [PubMed: 20018086]
- Lorenzo Bermejo J, Fischer C, Schulz A, Cremer N, Hein R, Beckmann L, Chang-Claude J, Hemminki K. Representation of genetic association via attributable familial relative risks in order to identify polymorphisms functionally relevant to rheumatoid arthritis. *BMC Proc.* 2009; 3 Suppl 7:S10.
- Mathew G, Xu H, George V. Simultaneous analysis of all single-nucleotide polymorphisms in genome-wide association study of rheumatoid arthritis. *BMC Proc.* 2009; 3 Suppl 7:S11. [PubMed: 20017974]
- Matthews AG, Li J, He C, Ott J, de Andrade M. Adjusting for HLA-DR β 1 in a genome-wide association analysis of rheumatoid arthritis and related biomarkers. *BMC Proc.* 2009; 3 Suppl 7:S12. [PubMed: 20017985]
- 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; 357:1199–1209. [PubMed: 17804836]
- Sarasua SM, Collins JS, Williamson DM, Satten GA, Allen AS. Effect of population stratification on the identification of significant single-nucleotide polymorphisms in genome-wide association studies. *BMC Proc.* 2009; 3 Suppl 7:S13. [PubMed: 20017996]
- Sha Q, Tang R. Detecting susceptibility genes for rheumatoid arthritis based on a novel sliding-window approach. *BMC Proc.* 2009; 3 Suppl 7:S14. [PubMed: 20018003]
- Suarez BK, Culverhouse R, Jin CH, Hinrichs AL. A search for non-chromosome 6 susceptibility loci contributing to rheumatoid arthritis. *BMC Proc.* 2009; 3 Suppl 7:S15. [PubMed: 20018004]
- Wang T, Elston RC. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet.* 2007; 80:353–360. [PubMed: 17236140]
- Wang Y, Sha N, Fang Y. Analysis of genome-wide association data by large-scale Bayesian logistic regression. *BMC Proc.* 2009; 3 Suppl 7:S16. [PubMed: 20018005]
- Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics.* 2009; 25:714–721. [PubMed: 19176549]
- Zhang M, Lin Y, Wang L, Pungpapong V, Fleet JC, Zhang D. Case-control genome-wide association study of rheumatoid arthritis from Genetic Analysis Workshop 16 using penalized orthogonal-components regression-linear discriminant analysis. *BMC Proc.* 2009; 3 Suppl 7:S17. [PubMed: 20018006]
- Zhang Q, Wang S, Ott J. Combining identity by descent and association in genetic case-control studies. *BMC Genet.* 2008; 9:42. [PubMed: 18601744]

- Zou GY, Donner A. The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: a cautionary note. *Ann Hum Genet.* 2006; 70:923–933. [PubMed: 17044866]
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol.* 2005; 67:301–320.

Table I

Summary of substantive findings by different methods for individual contributions to Group 1 of the Genetic Analysis Workshop 16

Authors	HLA regions identified	Other loci	Methods
Biernacka et al.	Not studied	<i>PADI4</i> <i>PTPN22</i>	Comparison of IMPUTE, MACH, PLINK, and fastPHASE for imputation. Focus on error rates and <i>p</i> -value differences, not substantive results, using four regions on chromosome 1, two associated, two not, with varying LD.
Chen et al.	<i>DRB1</i>	<i>TRAF1-C5</i> <i>PTPN22</i>	Comparison of allelic vs genotype coding of Hotelling's T^2
Fridley et al.	Not studied	<i>PADI4</i> <i>PTPN22</i>	Single vs. multiple imputation (using MACH). See comments on Biernacka et al. above
Jiang et al.	<i>DRB1</i>	Three other non-HLA chr 6 windows; chr 9 near <i>TRAF1-C5</i>	Multi-locus score test based on wavelets
Kwon et al.	Yes (NOS)*	Chr 1, 4, 5, 9, 10, 17, 18, 20	Bayesian classification with singular value decomposition
Lorenzo Bermejo et al.	Yes (rs7765379 and rs2395175)	Rankings differed by two indices. Hits on chr 1 (near <i>PTPN22</i>), 4, 7, 10, 13, 21, 22, but not <i>TRAF1-C5</i>	Joint use of Bayes factors and λ_s
Mathew et al.	<i>DRB1</i>	2627 SNPs significant across all autosomes at $p < 0.05$; largest effect sizes on chr 3, 4, 7, 9, 11, 17, 18, including <i>TRAF1-C5</i> and <i>PTPN22</i>	HLasso and Haploview
Matthews et al.	<i>DRB1</i> shared epitope (stratifying variable)	Many in carriers of shared epitope; low power in non-carriers. No association with biomarkers RFUW or anti-CCP. Novel locus <i>UBXD2</i> (chr 2); others on chr 1, 8, 9, 10	G×G interactions; biomarker associations
Sarasua et al.	Yes (NOS)	None (including <i>TRAF1-C5</i> and <i>PTPN22</i>) after control of stratification	Population stratification adjustment using PLINK and stratification on 10 PCs
Sha et al.	<i>TNF, A, B, C</i>	76 significant windows, including regions near recognized RA genes on chr 2 and 9	Adaptive sliding window approach
Wang et al.	Not examined	Multivariate analysis of 2705 SNPs on chr 9 included 10 SNPs with Z-scores > 5, two near <i>TRAF1-C5</i>	Bayesian penalized logistic regression with L_1 and L_2 norms
Zhang et al.	<i>DRA, DRB1, 5, 9, DQB1, DOB</i>	chr 9q33.1, 10q11.22, 18q12.1	POCRE-LDA

* Yes (NOS) indicates an association reported in the HLA region of chromosome 6, not otherwise specified.