Genetic Epidemiology 31 (Supplement 1):S12–S21 (2007)

Genetic Association with Rheumatoid Arthritis—Genetic Analysis Workshop 15: Summary of Contributions from Group 2

Marsha A. Wilcox,[1]* Zhong Li,[2,3] and Will Tapper[4] on behalf of Group 2†

[1]*i3 Drug Safety, Waltham, Massachusetts*
[2]*Department of Computational Genetics, High Throughput Biology Inc., Livingston, New Jersy*
[3]*Center for Computational Biology and Bioinformatics, Columbia University, New York, New York*
[4]*Human Genetics Division, University of Southampton, Southampton General Hospital, Southampton, United Kingdom*

The papers in presentation group 2 of Genetic Analysis Workshop 15 (GAW15) conducted association analyses of rheumatoid arthritis data. The analyses were carried out primarily in the data provided by the North American Rheumatoid Arthritis Consortium (NARAC). One group conducted analyses in the data provided by the Canadian Rheumatoid Arthritis Genetics Study (CRAGS). Analysis strategies included genome-wide scans, the examination of candidate genes, and investigations of a region of interest on chromosome 18q21. Most authors employed relatively new methods, proposed extensions of existing methods, or introduced completely novel methods for aspects of association analysis. There were several common observations; a group of papers using a variety of methods found stronger association, on chromosomes 6 and 18 and in candidate gene *PTPN22* among women with early onset. Generally, models that considered haplotypes or multiple markers showed stronger evidence for association than did single marker analyses. *Genet. Epidemiol.* 31 (Suppl. 1):S12–S21, 2007.

Key words: genome-wide association; chromosome 6; chromosome 18q21; HLA; *PTPN22*; search algorithms

*Correspondence to: Dr. Marsha A. Wilcox, i3 Drug Safety, 950 Winter Street, Waltham, MA 02451.
E-mail: Marsha.Wilcox@i3DrugSafety.com
†A complete list of the Group 2 primary contributing authors is given in the Acknowledgment section.
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20276

INTRODUCTION

The focus of Group 2 was association analyses of the rheumatoid arthritis (RA) data. Most participants analyzed data provided by the North American Rheumatoid Arthritis Consortium (NARAC), either in a genome-wide association context or with candidate genes. Several participants focused their work on a region of interest on chromosome 18q21 [Amos et al., 2007]. The unifying feature of this collection of papers is the application of relatively new methods, the extension of existing methods, or the proposal of completely novel methods for various aspects of association analyses. This summary presents each of the papers in the context in which the new or novel approach was applied. Table I shows the research design, methods, and results of each contribution to the group.

METHODS

PHENOTYPES

The majority of investigators used the binary RA affectation status indicator as provided [Browning and Thomas, 2007; Curtin et al., 2007; Ding et al., 2007; Gagnon et al., 2007; Kuo et al., 2007; Li et al., 2007; Mei et al., 2007; Tapper et al., 2007; Yu and Schaid, 2007b; Wei and Li, 2007]. Several investigators included other phenotypes; Curtin et al. [2007] also examined rheumatoid factor (RF) score and the anti-cyclic citrullinated peptide in their analyses; Gagnon et al. [2007] examined RA stratified by RF level. Wilcox and McAfee [2007] did not use the RA phenotype, but instead empirically identified four subtypes of RA based on clinical and biomarker data.

DATA

The data provided to GAW15 by NARAC were the basis for all but one of the analyses presented [Amos et al., 2007]. The genome scan data, candidate gene data sets, and chromosome 18q21 candidate region were used [Amos et al., 2007]. One group [Zhi et al., 2007] used the Canadian data (Canadian Rheumatoid Arthritis Genetic Study, CRAGS) for chromosome 6.

STATISTICAL METHODS

Phenotype definition. Wilcox and McAfee [2007] used a strategy for the development of qualitative traits that included nonparametric data reduction (multiple correspondence analysis), iterative two-staged clustering on the observed dimensions (k-means and agglomerative hierarchical clustering), and the assignment of

TABLE I. Association in rheumatoid arthritis: genetic analysis workshop 15; summary of contributions from group 2

Author	Data Used	Family-based or Case-control	Phenotype	Existing Methods/Software	New Methods / Software	Results
Browning, S	NARAC—Chromosome 18q data	Case-control	RA	Fisher's exact test; Permutation tests for significance; Haploview; fastPHASE	Beagle software: localized haplotype cluster test—single and multi-locus tests	New protective haplotype suggested ~500 bases upstream of the CCBE1 gene
Ding, Y	NARAC—genome scan and candidate	Case-control	RA	Permutation tests for significance levels fasePHASE for imputation	Backward Genotype Trait Association —extension of Backward haplotype trait association; two-stage genome scan	Scan: one central network—4 satellite networks—large number of interactions—7 of 8 network hubs are replications of previous findings Candidates: PTPN22, SUMO4, PDA14, DLG5, SLC22A4, CARD15
Curtin, K	NARAC—PTPN22 (20 SNPs)	Both in the same analysis	RA, RF, anti-CCP	Empirical null	PedGenie 2.1 – uses Cochran Mantel-Haenszel approach – meta statistics	Generally replicated published findings in 14 SNP panel
Gagnon, F	NARAC—Caucasians; PTPN22 and HLA-DRB1	Affect sibling pairs—one per nuclear family—with parents; linkage and association	RA, RF	Estimation of penetrances: PTPN22 and HLA-DRB1; Marker Association Segregation Chi-Square Method: simultaneously fit genotype frequency and IBD distributions (MERLIN)		Co-dominant model best fits PTPN22, no results for HLA-DRB1, nor for model stratified by RF (binary)
Kuo, T	NARAC—chromosome 18q	Case-control	RA	PHASE; Haploview	Malecot model; LDMAP; CHROMSCAN cluster	Identified two regions associated with RA; reduced SNP density decreased power and accuracy
Li, Z	NARAC—chromosome 18q	Case-control	RA	MC methods for validation	Pattern examiner: non-parametric data mining for multi-locus and interactions	10 potential multi-locus associations; all contained SNP0177 allele 1; ORs round 2; >50% of cases (common); intergenic regions; most in 2 LD blocks; Conclude 2 with gene-gender interaction
Mei, L	NARAC—candidates and PTPN22	Case-control	RA	GEE (in SAS); Haploview; MDR; empirical P-values		SINGLE SNP analyses: HAVCR1, CTLA4, SUMO4, MAP3K7IP2, PTPN22; associated with RA; two haplotypes for PTPN22 were associated, one protective; GENETIC INTERACTIONS: single locus model for PTPN22—known functional variant; no gene-gene interactions; gene-smoking (ever/current [no results]) interactions for CTLA4, PAD14, MIF, SNPs on chromosome 5, and one haplotype in PTPN22 (marginally protective effect)
Tapper, W	NARAC—chromosome 18q21	Case-control	RA		Malecot; CHROMSCAN - cluster	36 kb interval identified; gender-age-of-onset interaction—females with onset before age 40
Wei, Z	CRAGS (Canadian)—chromosome 6—Illumina for linkage, Affy for association	Linkage and family-based association	RA	LAMP/MERLIN	Test for assessing departure from HWE adjusting for association with disease	No replication of earlier published findings—three linkage peaks over 1.5; 13 SNP associations; new locus proposed PCDH9; small sample size and poor coverage likely contributing factors
Wilcox, M	NARAC—chromosome 6	Family-based	four empirically derived phenotypes	FBAT	Non-parametric data reduction and clustering to identify phenotypes	four phenotypes, three associated with different regions on chromosome 6 (also differently in the HLA region)
Yu, Z	NARAC—PTPN22 candidate and chromosome 18q	Case-control	RA		Sequential haplotype scan; goal—choose appropriate haplotypes	Sequential haplotype method and sequential summary method—stronger association than single SNP associations; Single SNP association not significant for R620W, but highly significant for the two sequential methods Chromosome 18: 2 sequential methods—weak regional evidence; no association with single SNP method

binary subgroup membership in each group for all individuals.

Data quality control. Many groups examined departure from Hardy-Weinberg equilibrium (HWE) in the control data as a preliminary quality control step before association analyses [Kuo et al., 2007; Tapper et al., 2007; Yu and Schaid, 2007b]. Single nucleotide polymorphisms (SNPs) with minor allele frequencies with less than 5% were eliminated by several groups [Kuo et al., 2007; Yu and Schaid, 2007b]. Some groups imputed missing genotypes using fastPhase [Scheet and Stephens, 2006; Ding et al., 2007; Browning and Thomas, 2007].

Wei and Li [2007] used the Canadian Rheumatoid Arthritis Genetic Study (CRAGS) data in which all individuals were affected. The expectation, in the presence of association of a SNP with disease, is a departure from HWE making it difficult to use departure from HWE as a data quality control measure. These authors proposed a new method using a likelihood framework for the assessment of departure from HWE while accounting for potential association with the disease.

Estimation of penetrance. Gagnon et al. [2007] estimated penetrances of *PTPN22* and *HLA-DRB1* and tested a co-dominant genetic model of *PTPN22* (rs2476601 – R620W) conditional on RF status. The sib-pair design used by this group included one sib-pair per nuclear family. The Marker Association Segregation Chi-Square (MASC) method [Clerget-Darpoux et al., 1988] was used to simultaneously fit genotype frequency and identical by descent (IBD) distributions for 3,690 individuals from 604 families. This approach uses both the allelic association information and linkage information from IBD sharing among sibs. MERLIN [Abecasis et al., 2002] was used for the IBD estimation.

Malecot model. Kuo et al. [2007] and Tapper et al. [2007] used the CHROMSCAN-cluster program [Morton et al., 2007] to analyze the 18q21 data. CHROMSCAN-cluster splits the data into non-overlapping regions, which cover at least 10 linkage disequilibrium units (LDUs) and contain a minimum of 30 SNPs without breaking linkage disequilibrium (LD) blocks. Once regions have been formed, Malecot parameters are estimated iteratively by composite likelihood, which evades severe Bonferroni corrections by modeling association across all markers in a region. In order to account for correlation between SNPs as a result of LD, *P*-values are determined empirically by a rank-based permutation test.

Two subhypotheses of the Malecot model are used to test for a causal polymorphism. Model A, which assumes no association between affection status and SNPs, is taken as the null hypothesis and compared with Model D which estimates disease location (S), complex inheritance (M), and residual association (L). The test statistic is determined by the difference in minus 2 natural log composite likelihoods between these two models to give a χ^2 with three degrees of freedom. Kuo et al. [2007] and Tapper et al. [2007] used this method to determine the location and weight of evidence for a gene predisposing to rheumatoid arthritis. Additionally, Tapper et al. [2007] created subsets of the NARAC data, based on gender and age of onset, and used estimates of location and information to combine evidence from association with a meta-analysis of linkage studies. Kuo et al. [2007] used the PHASE [Stephens et al., 2001] program to perform a haplotype analysis of the candidate regions and studied the effect of region length in LDU and SNP density on the accuracy of association mapping.

Haplotypes. There were several applications of methods for association analyses with haplotypes. Browning and Thomas [2007] used an algorithm for localized haplotype cluster identification in the software program Beagle for single and multi-locus association tests in the candidate region on chromosome 18q21 [Browning, 2006]. The localized haplotype cluster test models the LD structure in densely spaced markers to derive haplotype clusters. This approach may be more flexible than the haplotype block approach because no structure is imposed on the data. Association is assessed in a case–control paradigm. The false-positive rate is evaluated using permutation tests.

Yu and Schaid [2007b] presented a sequential haplotype scan method motivated by the notion that arbitrarily chosen markers may lower the statistical power to detect association due to the presence irrelevant markers and the need to examine all possible combinations making the analysis ''computationally expensive.'' The goal of the method is to choose appropriate markers to include in haplotypes rather than using a traditional sliding window or other similar approach. Specifically, markers close to each other are added sequentially; markers are added if they provide additional information conditional on the existing haplotype. The Mantel-Haenszel test is used to assess association [Yu and Schaid, 2007a]. This approach results in two test statistics, a haplotype χ^2 and summary statistic. Yu and Schaid [2007b] employed this method in the analysis of candidate gene *PTPN22* and the chromosome 18q21 candidate region. The false-positive rate was assessed using permutation tests.

Association: data mining methods. Ding et al. [2007] used the Backward Genotype-Trait Association (BGTA) method [Zheng et al., 2006], an extension of the Backward Haplotype-Trait Association (BHTA) method for case-parent trios [Lo and Zheng, 2002]. The BGTA is a backward greedy

search algorithm removing one marker at a time from the set until the ''optimal'' set of markers is identified. Statistical significance is assessed using permutation tests. Once the set of markers is identified an association network is created. This group conducted both a genome scan of 5,407 SNPs in 642 families and an examination of candidate genes (20 SNPs in 14 genes) in 839 cases and 855 unrelated controls. For the genome scan they applied a two-stage selection process to avoid the computational burden that would have been encountered if all markers had been considered at once. Markers with high-dimensional interactions were assumed to have information in the lower dimensional pair-wise evaluation. Instead of analyzing all 5,407 SNPs this strategy resulted in the consideration of 707 SNPs. The analyses were conducted on 700,000 subsets of eight SNPs each. In the candidate gene study all possible subsets of 20 SNPs were used, resulting in more than 1 million GTD (genotype-trait distortion) scores. Next, 30,000 subsets of size 8 were screened. The family-wise 1% significance threshold was established using a permutation approach in which the affection assignment was randomly assigned. Graphical networks were constructed to show the associations among subsets with more than 1 SNP.

Both Li et al. [2007] and Shaw (personal communication) used data mining approaches to identify unlinked multi-locus genetic effects for complex disease. Li et al. [2007] used a ''pattern'' to represent a collection of markers with particular allele or genotype values shared among a number of individuals. Shaw (personal communication) employed a ''factorset'' to represent combinations of ''factors,'' which in turn is a combination of a feature, such as a marker or gender, and a feature value, such as a marker genotype. In the pattern-based strategy (Pattern Examiner) developed by Li et al. [2007], a pattern-based data mining method was used to exhaustively, yet efficiently, identify all patterns satisfying predefined pattern search criteria. The *support threshold* specified the minimum number of cases a pattern must have. The *locus threshold* specified the extent of locus interaction. The significance of association with disease for each pattern was evaluated through the combination of a χ^2-based test statistic, a multiple-testing correction scheme, and a Monte-Carlo permutation test. In FOCUS Discovery, which is an asynchronous, parallel, nonlinear software developed by Shaw (personal communication), a suite of machine learning inspired algorithms was used to detect epistatic and nonlinear gene-gene interactions. Factorsets containing 1–3 factors and satisfying preset minimal accuracy (the total number of cases containing a factorset over the total number of cases and controls containing that factorset) and support (the total number of cases and controls containing a factorset) threshold were identified and ranked according to their accuracies.

Gene-gene and gene-smoking interactions using MDR. Mei et al. [2007] employed both case-control data sets (20 SNPs from 15 candidates and 14 SNPs from *PTPN22*) to examine gene-gene and gene-smoking interactions. Single gene associations among 15 candidate genes and haplotypes of *PTPN22* were examined using generalized estimating equations (GEE) [Zeger and Liang, 1986] to account for within-family correlation. Haploview [Barrett et al., 2005] was used to examine haplotypes. Haplotypes were reconstructed using PHASE 2.0 [Stephens et al., 2001]; low probability haplotypes were removed. Association analyses used common haplotypes and assumed a dominant model. Gene-gene interactions were assessed using the multifactor dimensionality reduction method (MDR) [Ritchie et al., 2001] and logistic regression.

Candidate gene-smoking interactions were evaluated using the MDR approach in a case-only sample (19 SNPs from 14 candidate genes). Logistic regression was conducted for interactions suggested by the MDR analyses.

MDR is a frequently used data mining approach for detecting multi-locus genetic association. A comparison of MDR, Pattern Examiner, and FOCUS Discovery is presented in Table II to show the similarities and differences among these methods. The three methods share the following traits: they are computationally intensive, designed to detect locus interactions, non-parametric, and designed to be used on data collected using case-control designs.

Meta-analysis. Curtin et al. [2007] conducted a combined analysis of both family-based and case-control subjects across multiple resources in the *PTPN22* candidate gene data. They used the RA phenotype and subtypes (RF and anti-CCP) in 1,285 family-based cases from sib-pair pedigrees and 1,518 matched, independent controls. In addition, 103 unaffected family-based controls were included from the sib-pair pedigrees to use all available data in association testing. For demonstrating meta analysis across multiple studies, the data were separated into two study designs. The first included the families and independent controls reported by Plenge et al. [2005], the second included the remainder of the available NARAC data [Carlton et al., 2005]. Results were presented for the combined analysis of the two association studies using PedGenie version 2.1 software [Allen-Brady et al., 2006]. This strategy accounted for correlations between related individuals. PedGenie summarizes data in three-way contingency tables. The tables are

TABLE II. Feature comparisons among data mining methods; MDR, Pattern Examiner, and FOCUS Discovery

	MDR	Pattern Examiner	FOCUS discovery
Data analyzed	Genotype	Allele and genotype	Genotype
Dimension reduction method	Label each multifactor cell in *n*-dimensional space as either "high risk" or "low risk" to create a one-dimensional variable	Identify patterns (equivalent of multifactor cells in MDR) as the one-dimension variable	Identify factorsets (equivalent of multifactor cells in MDR) as the one-dimension variable
Significance Evaluation	Rank multifactor models based on its ability to classify and predict disease status	χ^2 statistics for each pattern	Rank by accuracy
Validation	Cross-validation	Multiple testing correction and Monte Carlo permutation test	None
Result reported	One best multi-factor model to classify and predict disease status	All statistically significant patterns that suggest disease association and gene-gene interaction	Several top factorsets that provide the greatest overall dataset coverage at the highest overall accuracy
Disease model predicted?	No	Yes (depending on the inclusion of allele or homozygote genotypes in a significant pattern)	No

stratified by study (two studies, in this case), and combined in a meta analysis, with disease status cross-classified by genotype. Meta statistics for allele- and genotype-based associations use the generalized Cochran Mantel-Haenszel (CMH) approach. Simulation of an empirical null distribution, based on a Monte-Carlo procedure, is used for significance testing.

RESULTS

PHENOTYPE DEFINITION

Wilcox and McAfee [2007] identified five distinct phenotypic subgroups in the data. The first was composed of unaffected family members. Subgroup 2 was a mix of affected and unaffected in which patients endorsed symptoms not corroborated by physicians. Subgroups 3–5 represented a severity continuum in RA. Subgroup 5 was characterized by early onset severe disease. A binary indicator of subgroup membership was used in the association analyses for the chromosome 6 data. Each individual was identified as belonging to one and only one group.

GENETIC DATA: DEPARTURE FROM HWE

The CRAGS provided two data sets containing 128 individuals from 60 families and 158 affected sibling pairs (ASPs). There were 5,429 SNPs across 22 autosomes in the family-based sample and 113,237 SNPs in the ASP sample. Wei and Li [2007] restricted the data set to those with more than 85% complete genotypes and minor allele frequencies greater than 5%. On the basis of this new method to examine departure from HWE in a highly ascertained sample, 145 SNPs were omitted from the association tests.

GENETIC DATA: ESTIMATION OF PENETRANCES

Gagnon et al. [2007] estimated penetrances for *PTPN22* and *HLA-DRB1*. For *PTPN22*, the co-dominant model fit the data. The estimates of the penetrances for TC and CC genotypes are 0.54 and 0.20 respectively. No differences between these penetrances were found with respect to RF status.

The classification of the *HLA-DRB1* alleles was based upon the amino acid sequence at position 71–74. Individuals with HLA-DRB1∗0101, ∗0405, ∗0408, ∗1001, ∗1402, and ∗16 were assigned E_1; ∗0401 and ∗0409 were assigned E_2; ∗0102, ∗0404, ∗423, ∗12, and ∗1406 were assigned E_3; all other alleles were assigned E_x (non-susceptibility alleles). Because of the lack of subtyping for some alleles, all of the observed frequencies differed significantly from the expected distributions in this study. This precluded an examination of the interaction of *PTPN22* and *HLA-DRB1*.

HAPLOTYPES

Browning and Thomas [2007] used the localized haplotype cluster test to examine ~2,300 SNPS on a 10 Mb region of chromosome 18q21. They conducted single marker and haplotype association analyses in 460 case-control pairs. They derived 4,345 haplotype clusters and conducted 2,295 single marker tests. This approach resulted in the identification of a rare protective haplotype less than 500 bases upstream of the *CCBE1* gene.

Yu and Schaid [2007b] employed the sequential haplotype scan method to the *PTPN22* and chromosome 18q21 data. They reported stronger association using their new method than with single-SNP methods. The single-SNP association was not significant for the putative region, R620W (with R620W removed), but the two sequential methods showed highly significant results. On chromosome 18q21, the 2 sequential methods showed weak regional evidence of association. When the R620W variant was not included in the analyses, the sequential methods showed stronger association than did the single-locus methods.

GENOME-WIDE ASSOCIATION

In a genome scan, Ding et al. [2007] identified one central network with four satellite sub-networks. The network is depicted in Figure 1. A large number of interactions was identified. Seven of the eight identified 'hubs' in the network had prior reports of association with a liability for RA.

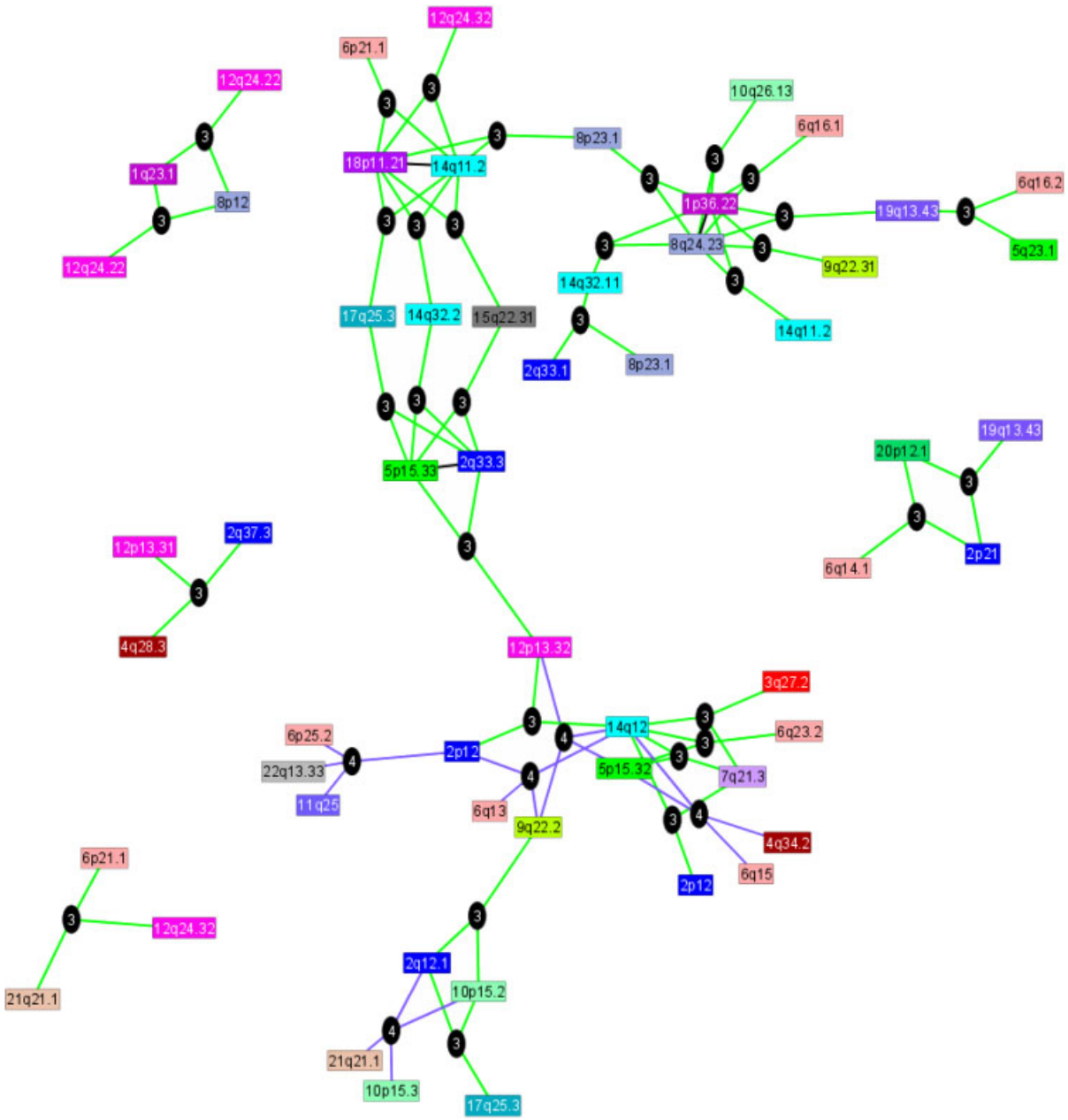

Fig. 1. Association network example: Ding et al., Color of loci node: chromosome; black circle with number: n-way interaction; black edge: 2-way interaction; green edge: 3-way interaction; purple edge: 4-way interaction. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com]

CANDIDATE GENES

PTPN22. Table III summarizes the results for *PTPN22* among the variety of methods applied by this working group. Ding et al. [2007] showed association with *PTPN22* and *SUMO4*. The association network included *PTPN22, PDA14, DLG5*, with three-way interaction for *SLC22A4 SUMO4, CARD15* ($P = 0.0002$). Using meta-statistics, Curtin et al. [2007] generally replicated published findings in the 14-SNP panel for *PTPN22* with RA and RF. However, their findings did not replicate previous findings for anti-CCP antibody status and disease in the functional R620W SNP. The sequential haplotype scan method employed by Yu and Schaid [2007b] resulted in associations with the putative R620W region that were not replicated in the traditional single-SNP analyses when the R620W variant was not included in the analyses.

Mei et al. [2007] showed association with one SNP in each gene of the following genes in the single-SNP analyses: *HAVCR1, CTLA4, SUMO4, MAP3K7IP2*, and *PTPN22*. This group constructed five common haplotypes for *PTPN22*. Two of these were associated with RA and one showed evidence of a protective effect. No epistasis or gene-gene interactions were reported for *PTPN22* by this group. They did report a marginally protective gene-environment interaction with one haplotype in PTPN22 and current smoking.

Gagnon et al. [2007], using the MASC method, reported a co-dominant model for *PTPN22* and RA.

In the Canadian data, Wei and Li [2007] failed to replicate previous findings for *PTPN22*. This was likely due to the restricted sample size and the lack of adequate coverage of this area in the chips used in the study.

CHROMOSOME 6

Wilcox and McAfee [2007] examined SNPs on chromosome 6 in each of the empirically derived subgroups. The severity continuum in subgroups 3–5 showed association with 17 SNPs. There was little overlap in the results across subgroups. In the HLA region, subgroup 3 showed single, 2- and 3-SNP association with the centromeric side of the region in an area of linkage disequilibrium. Subgroup 5 showed both single and 2-SNP association with the telomeric side of the region in a second area of linkage disequilibrium.

Using linkage analysis in the Canadian sample, Wei and Li [2007] reported a maximum LOD score of 1.83 on chromosome 6. However, there was no confirmation of association findings in the HLA region. Gagnon et al. [2007] failed to replicate previous reported findings for *HLA-DRB1* probably due to the lack of subtyping for some alleles.

CHROMOSOME 18Q21

Several papers analyzed the chromosome 18 datasets for disease association. Kuo et al. [2007] and Tapper et al. [2007] used the CHROMSCAN-cluster program to establish the location and weight of evidence for a gene predisposing to rheumatoid arthritis. Both authors removed seven markers due to extreme HWDs in controls ($\chi^2_1 \geq 10$) and partitioned the remaining data into non-overlapping regions which cover a minimum of 10 LDUs and contain at least 30 SNPs. Tapper et al. [2007] used an LDU map generated from HapMap data to split the data into 18 regions, whereas Kuo et al. [2007] used an LDU map constructed from controls to form 14 regions. These investigators used CHROMSCAN cluster to test each of these regions for the presence of a causal locus by using composite likelihood to model association across all markers simultaneously thereby reducing the number of tests. Despite using different LDU maps, Kuo et al. [2007] and Tapper et al. [2007] identified the same causal locus, defined S_1 by Kuo et al. [2007], at 53,308 kb when using the default region size of 10 LDUs.

CHROMSCAN cluster determines *P*-values empirically using a rank-based permutation test, which accounts for the correlation between SNPs due to LD. However, to determine accurate *P*-values, the number of permutation replicates must approach the actual level of significance so that interpolation of the variance under H_1 is reliable. Tapper et al. [2007] performed preliminary screens with 100 replicates

TABLE III. Summary of PTPN22 findings using a variety of methods

Author	Methods	Results
Curtin et al.	Meta-association (CMH) PedGenie v2.1	Confirmed association with functional SNP (R620W) More power in subset analysis by using both family-based and case-control samples
Ding et al.	Backward Genotype Trait Association	Confirmed previous findings
Gagnon et al.	Marker Association Segregation Chi-Square method	Confirmed previous findings
Mei et al.	Related sib-pairs controlled using SAS GENMOD	Ever-smoking (age, sex, body mass index) gene-smoking interaction quantitative trait
Yu and Schaid	Sequential haplotype scan method	Stronger association in the PTPN22 region than single locus analysis

and re-analyzed significant regions using 1,000 and 5,000 replicates to verify convergence, whereas Kuo et al. [2007] used 1,000 replicates. Estimates of the significance of S_1 were similar ($p = 0.0017$ and $p = 0.0039$ for Kuo et al. [2007] and Tapper et al. [2007] respectively). In contrast, single SNP χ^2 analyses identified 125 SNPs with nominal significance, none of which remain significant after Bonferroni correction.

As shown in Figure 3, the 95% confidence interval of S_1 covered 36 kb, between 53296 kb and 53332 kb which includes the most significant single SNP χ^2 (msSNP rs3745064, SNP1100). Although no described genes are within this region it does include four mRNAs (CR590917, AK021217, AK124558, and BC01314). Of these, CR590917 appears to be the most interesting as it is expressed within T-cells and could therefore conceivably affect risk for rheumatoid arthritis.

King et al. [1992] demonstrated that the risk for rheumatoid arthritis is elevated in females, especially with late onset. Tapper et al. [2007] stratified cases into three subgroups according to gender and age of onset. In their study, association evidence was derived almost entirely from females with an age of onset less than 40.

Kuo et al. [2007] identified a second locus, S_2, at 51,585 kb with nominal significance ($P = 0.04$) when their analysis was repeated with a region length of 5 LDUs giving 26 regions. Using PHASE [Stephens et al., 2001] to perform haplotype analysis in candidate regions containing S_1 and S_2, Kuo et al. [2007] identified two protective haplotypes, one in each region. By reducing SNP density, this group demonstrated that the power and accuracy of association mapping is reduced. These reductions were less severe when SNP selection was based on equal LDU distance compared with equal kb distance or SNP tagging by Tagger in the Haploview software [Carlson et al., 2004]. This was consistent with their findings using the simulated data from GAW15 Problem 3 [Zheng et al., 2007].

Using the sequential haplotype scan approach, Yu and Schaid [2007b] identified several sets of SNPs that showed strong significance: one of them included markers rs3848517 (SNP1230), rs4941196 (SNP1231), and rs2319006 (SNP1234).

Browning and Thomas [2007] employed a localized haplotype cluster test as well as a multi-locus score test and a haplotype block-based test using Haploview. The most significant haplotype contained SNPs rs2195534 (SNP1631) and rs1791320 (SNP1632) that were located less than 500 bp upstream of *CCBE1* gene. This haplotype actually is associated with reduced risk of RA. The significant finding was confirmed with a haplotype test using Haploview.

Li et al. [2007] employed a pattern-based mining strategy to detect local markers in moderate or strong LD and global, unlinked markers, multi-locus genetic associations as well as gene-gene/gene-environment interactions. Ten patterns were found

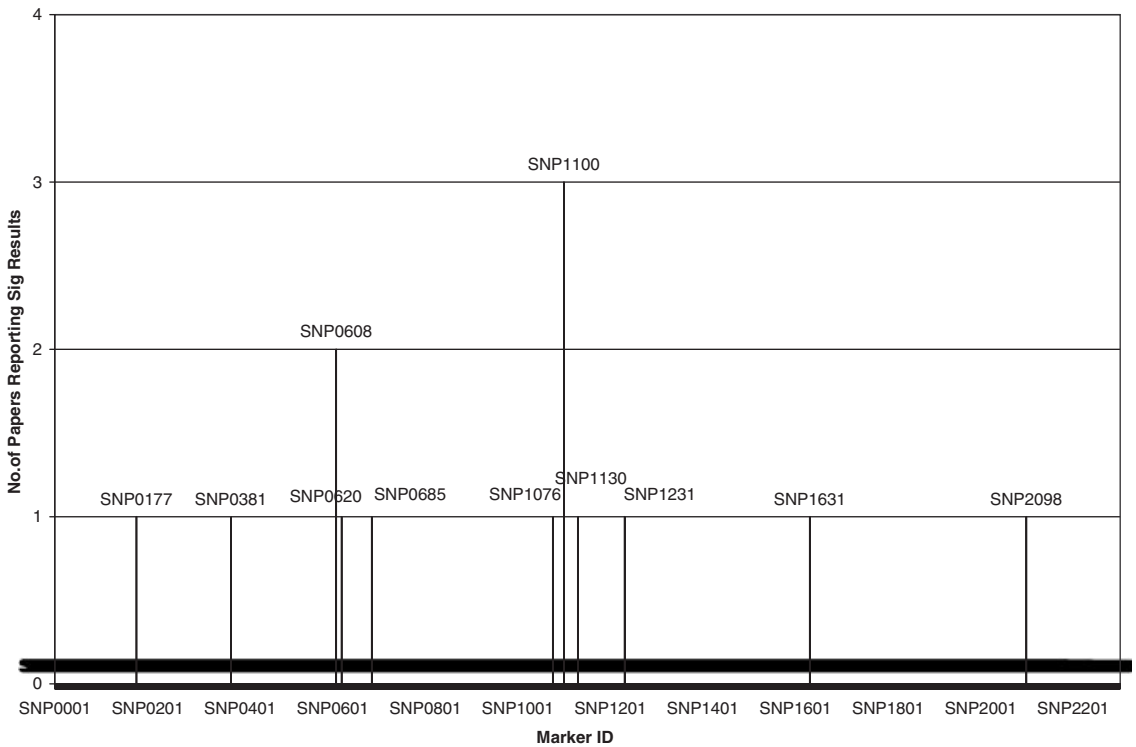

Fig. 2. Summary of findings on chromosome 18q21.

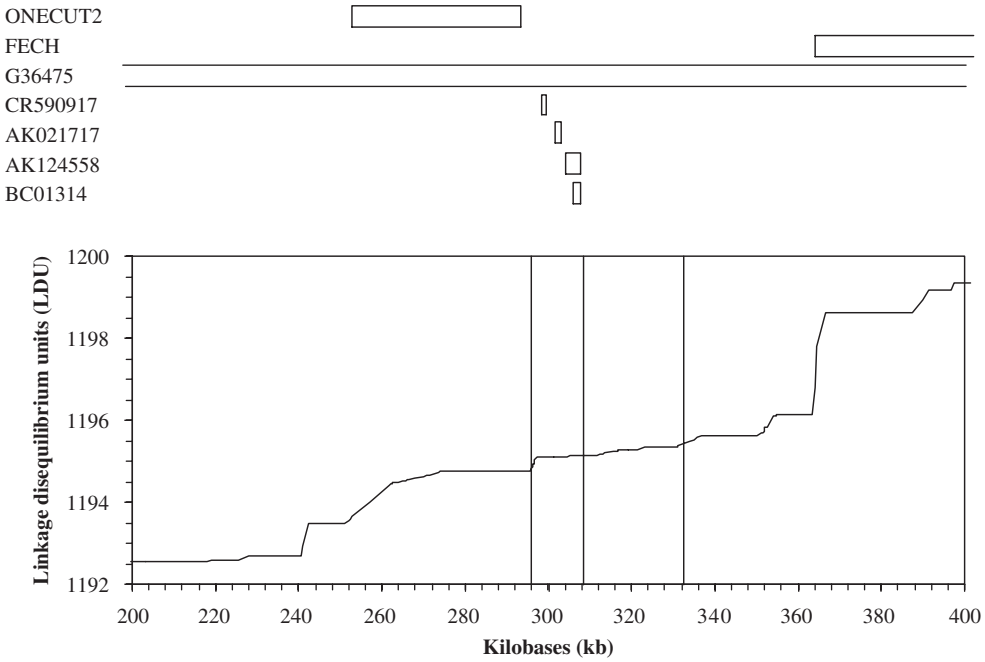

Fig. 3. Malecot Model Results: Candidate loci in 95% confidence interval of region 6 on chromosome 18. Plotting the LDU map shows that fine scale patterns of LD consisted of distinct plateaus and steps which reflected areas of high and low LD that corresponded with regions of low and high recombination. The estimated causal location and its 95% confidence interval are shown by the three vertical lines. Genomic locations are recovered by adding 53000 kb. Candidate loci are represented by oblong boxes in the upper diagram and their position reflects genomic locations in relation to the x-axis of the graph.

to be significantly associated with the RA phenotype after correction for multiple testing. The significance of the patterns was confirmed using Monte-Carlo simulation. This group also identified a potentially significant multi-gene/gender interaction involving two loci: SNP0177 and the LD region containing markers SNP0603–SNP0615.

Figure 2 shows a summary of significant results for the chromosome 18q21 data. The number of papers reporting significant results using the chromosome 18 dataset was tallied according to marker ID. If a haplotype containing several markers in the same LD region was found to be significant, one marker was selected to represent the haplotype.

Meta analysis. Curtin et al. [2007] conducted a meta-analysis of SNPs in *PTPN22*. The approach implemented in PedGenie [Allen-Brady et al., 2006] resulted in the replication of most of the published findings in this gene.

Tapper et al. [2007] conducted a meta-analysis of linkage studies and three subsets of NARAC data using estimates of disease location (S) and information (K) to combine evidence from association. This study provided significant evidence for a determinant of RA in a 36 kb interval and demonstrated that estimates of S and K are more powerful for meta-analysis than estimates of *P*-values alone.

DISCUSSION

The primary unifying feature of this group of papers was association analysis in rheumatoid arthritis. However, most contributions also extended existing methods or proposed novel methods to accomplish this goal. The newer approaches included: an empirical definition of the phenotype; a proposed method for the examination of HWE in an affected-only sample; the application of a method to simultaneously fit genotype frequency and IBD distributions; the extension and introduction of methods to choose optimal haplotypes; data mining methods for pattern discovery; the simultaneous estimation of both location and association of a locus with disease; and the use of meta-statistics, incorporating both family-based and case-control data, to confirm association results.

There were some common findings across studies. Several studies confirmed previous results for *PTPN22*. Sample size and gene chip coverage were likely limitations with the Canadian data that precluded replication here. There was some consensus about a potential locus among the studies that examined the candidate region on chromosome 18. The HLA region on chromosome 6 was not replicated in all of the studies in which it was examined. Many authors proposed novel loci in the etiology of rheumatoid arthritis.

Several contributors using different methods and different data sets observed an apparent age-gender interaction. Tapper et al. [2007] showed that women with an age of onset less than 40 had a large influence on their findings. Curtin [2007] reported a strong age effect. Wilcox and McAfee [2007] reported a subgroup, predominantly women, with early onset severe disease. Li et al. [2007] reported an elegant multi-gene-gender interaction.

The real value of the methods proposed and extended by this group will be demonstrated as they are applied to other diseases and other types of data. The extent to which the proposed new loci are in fact part of the etiology of RA will depend on confirmation in other settings.

ACKNOWLEDGMENTS

Sharon Browning, Karen Curtin, Jie Ding, Yuejing Ding, France Gagnon, Qimei He, Tai-Yue Kuo, Mingyao Li, George Matthew, Ling Mei, Shaoqi Rao, Jim Shaw, Zhi Wei, Zhaoxia Yu, Weihua Zhang, Tian Zheng, and Guohua Zhu.

REFERENCES

Abecasis G, Cherny S, Cookson W, Cardon L. 2002. Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101.

Allen-Brady K, Wong J, Camp N. 2006. PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. BMC Bioinformatics 7:209.

Amos CI, Chen WV, Remmers E, Siminovitch KA, Seldin MF, Criswell LA, Lee AT, John S, Shephard ND, Worthington J, Cornelis F, Plenge RM, Begovich AB, Dyer TD, Kastner DL, Gregersen PK. 2007. Data for Genetic Analysis Workshop (GAW) 15 problem 2, genetic causes of rheumatoid arthritis and associated traits. BMC Proceedings 1(Suppl 1):S3.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265.

Browning SR. 2007. Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 78:903–913.

Browning SR, Thomas J. 2007. Multilocus analysis of GAW15 NARAC chromosome 18 case-control data. BMC Proceedings 1(Suppl 1):S11.

Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74:106–120.

Carlton VE, Hu X, Chokkalingam AP, Schrodi SJ, Brandon R, Alexander HC, Chang M, Catanese JJ, Leong DU, Ardlie KG, Kastner DL, Seldin MF, Criswell LA, Gregersen PK, Beasley E, Thomson G, Amos CI, Begovich AB. 2005. PTPN22 genetic variation: evidence for multiple variants associated with rheumatoid arthritis. Am J Hum Genet 77:567–581.

Clerget-Darpoux F, Babron MC, Prum B, Lathrop GM, Deschamps I, Hors J. 1988. A new method to test genetic models in HLA associated diseases: the MASC method. Ann Hum Genet 52 (Pt 3):247–258.

Curtin K, Wong J, Allen-Brady K, Camp N. 2007. Meta genetic association of rheumatoid arthritis and PDPN22 using PedGenie 2.1. BMC Proceedings 1(Suppl 1):S12.

Ding Y, Cong L, Ionita-Laza I, Lo S-H, Zheng T. 2007. Constructing gene association networks for rheumatoid arthritis using the backward genotype-trait association (BGTA) program. BMC Proceedings 1(Suppl 1):S13.

Gagnon F, Hajage D, Plancoulaine S, Tezenas du Montcel S. 2007. Modeling of PTPN22 and HLA-DRB1 susceptibility to rheumatoid arthritis. BMC Proceedings 1(Suppl 1):S14.

King RA, Rotter JI, Motulsky AG. 1992. The Genetic Basis of Common Disease. New York: Oxford University Press. p 598–599.

Kuo T-Y, Lau W, Hu C, Zhang W. 2007. Association mapping of susceptibility loci for rheumatoid arthritis. BMC Proceedings 1(Suppl 1):S15.

Li Z, Zheng T, Califano A, Floratos A. 2007. Pattern-based mining strategy to detect multi-locus association and gene-environment interaction. BMC Proceedings 1(Suppl 1):S16.

Lo S, Zheng T. 2002. Backward haplotype transmission association (BHTA) algorithm—A fast multiple-marker screening method. Hum Hered 53:197–215.

Mei L, Li X, Yang K, Cui J, Fang B, Guo X. 2007. Evaluating gene-gene and gene-smoking interaction in rheumatoid arthritis using candidate genes in GAW 15. BMC Proceedings 1(Suppl 1):S18.

Morton N, Maniatis N, Zhang W, Ennis S, Collins A. 2007. Genome scanning by composite likelihood. Am J Hum Genet 80:19–28.

Plenge RM, Padyukov L, Remmers EF, Purcell S, Lee AT, Karlson EW, Wolfe F, Kastner DL, Alfredsson L, Altshuler D, Gregersen PK, Klareskog L, Rioux JD. 2005. Replication of putative candidate-gene associations with rheumatoid arthritis in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. Am J Hum Genet 77:1044–1060.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629–644.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989.

Tapper W, Collins A, Morton N. 2007. Mapping a gene for rheumatoid arthritis on chromosome 18q21. BMC Proceedings 1(Suppl 1):S18.

Wei Z, Li M. 2007. Cenome-wide linkage and association analysis of rheumatoid arthritis in a Canadian population. BMC Proceedings 1(Suppl 1):S19.

Wilcox M, McAfee A. 2007. Empirically derived subgroups in rheumatoid arthritis: Association with SNPs on chromosome 6. BMC Proceedings 1(Suppl 1):S20.

Yu Z, Schaid D. 2007a. Sequential haplotype scan methods for association analysis. Genet Epidemiol 31:553–564.

Yu Z, Schaid D. 2007b. Application of sequential haplotype scan methods to case-control data. BMC Proceedings 1(Suppl 1):S21.

Zeger S, Liang K. 1986. Longitudinal data analysis for discrete and continuous outcomes. Biometrias 42:121–130.

Zhang W, Lau WWS, Hu C, Kuo T-Y. 2007. Impact of marker density on the accuracy of association mapping. BMC Proceedings 1(Suppl 1):S166.

Zheng T, Wang H, Lo SH. 2006. Backward genotype-trait association (BGTA)-based dissection of complex traits in case-control design. Human Hered 62:196–212.