BMC Proceedings

BioMed Central

Proceedings

Open Access

Genome-wide gene-based association study

Hsin-Chou Yang*[1], Yu-Jen Liang[1], Chia-Min Chung[2], Jia-Wei Chen[1] and Wen-Harn Pan[2]

Addresses: [1]Institute of Statistical Science, Academia Sinica, Number 128, Section 2, Academia Road, Nankang, Taipei 115, Taiwan, Republic of China and [2]Institute of Biomedical Sciences, Academia Sinica, Number 128, Section 2, Academia Road, Nankang, Taipei 115, Taiwan, Republic of China

E-mail: Hsin-Chou Yang* - hsinchou@stat.sinica.edu.tw; Yu-Jen Liang - lyj626@stat.sinica.edu.tw; Chia-Min Chung - akira@ibms.sinica.edu.tw; Jia-Wei Chen - jiawei@stat.sinica.edu.tw ; Wen-Harn Pan - pan@ibms.sinica.edu.tw
*Corresponding author

from Genetic Analysis Workshop 16
St Louis, MO, USA 17-20 September 2009

Published: 15 December 2009

BMC Proceedings 2009, **3**(Suppl 7):S135 doi: 10.1186/1753-6561-3-S7-S135

This article is available from: http://www.biomedcentral.com/1753-6561/3/S7/S135

Abstract

Genome-wide association studies, which analyzes hundreds of thousands of single-nucleotide polymorphisms to identify disease susceptibility genes, are challenging because the work involves intensive computation and complex modeling. We propose a two-stage genome-wide association scanning procedure, consisting of a single-locus association scan for the first stage and a gene-based association scan for the second stage. Marginal effects of single-nucleotide polymorphisms are examined by using the exact Armitage trend test or logistic regression, and gene effects are examined by using a *p*-value combination method. Compared with some existing single-locus and multilocus methods, the proposed method has the following merits: 1) convenient for definition of biologically meaningful regions, 2) powerful for detection of minor-effect genes, 3) helpful for alleviation of a multiple-testing problem, and 4) convenient for result interpretation. The method was applied to study Genetic Analysis Workshop 16 Problem 1 rheumatoid arthritis data, and strong association signals were found. The results show that the human major histocompatibility complex region is the most important genomic region associated with rheumatoid arthritis. Moreover, previously reported genes including *PTPN22*, *C5*, and *IL2RB* were confirmed; novel genes including *HLA-DRA*, *BTNL2*, *C6orf10*, *NOTCH4*, *TAP2*, and *TNXB* were identified by our analysis.

Introduction

Genome-wide association study (GWAS) has been broadly applied to identify disease susceptibility genes of complex disorders. Single-locus association tests are routinely run to identify causal or associated single-nucleotide polymorphisms (SNPs) having strong marginal effects on disease status; however, their power to detect minor-effect SNPs may not be satisfactory. Multilocus association tests, which incorporate genetic information such as linkage disequilibrium (LD) and genetic distance, are performed to improve test power of single-locus association tests.

In order to analyze a large number of SNPs across the human genome, chromosomal regions on which to apply multilocus association tests should be defined in advance. Two frequently used procedures to define regions in a GWAS are the sliding-window approach and LD-block approach. A sliding-window approach defines regions by assigning a pre-determined window size or selecting a window size subject to an optimization criterion, and then a multilocus association test is performed in each window. It provides a convenient way to define regions and then scan each chromosome sequentially; however, the defined chromosomal segments may not have a biological function. An LD-block approach defines regions by determining LD/haplotype blocks, and then a multilocus association test is performed in each block. It uses a data-driven procedure to define blocks and then focuses on the examination of biologically meaningful blocks; however, use of different LD measures or block identification algorithms may obtain different blocks and hence draw different conclusions.

In this paper, we propose a two-stage genome-wide association scan, consisting of a single-locus association scan for the first stage and a gene-based association scan for the second stage. In comparison with a single-locus association test, the proposed method has the following merits: 1) biological information is incorporated into the definition of study regions, 2) tests are more powerful relative to single-locus association tests, 3) the multiple-testing problem is alleviated, and 4) the impact of genes can be evaluated directly and results are easier to interpret and generalize. Compared with a sliding-window approach, a gene-based approach contains richer information in a biological sense; compared with an LD-block approach, the regions analyzed by a gene-based approach are more stable and the analysis involves less intensive computation.

The proposed method was used to identify disease genes susceptible to rheumatoid arthritis (RA). We analyzed Genetic Analysis Workshop 16 Problem 1 RA data. The data consisted of 2,062 Illumina 550 k SNP chips from 868 RA patients and 1,194 normal controls collected by the North American Rheumatoid Arthritis Consortium [1]. Genotype data of 545,080 SNPs, which were probed on an Illumina 550 k SNP chip, were provided. A dichotomous disease status of RA and 530,720 autosomal SNP markers were analyzed in this GWAS.

Methods

We illustrate the flow of the proposed two-stage genome-wide association scan as follows. At the first stage, we quantify trend effects of alleles for autosomal SNPs by calculating *p*-values of the exact version of the Armitage trend test [2], which is a powerful and valid association test even for analyses of rare-allele loci and Hardy-Weinberg-disequilibrium loci. The exact *p*-value of the i^{th} SNP is the sum of probabilities for the permutations with statistics at least as extreme as the observed statistic. A logistic regression can be carried out if genetic and/or environmental effects should be adjusted.

At the second stage, we carry out a genome-wide gene-based association scan. All SNPs are divided into two types: inter-gene SNPs and intra-gene SNPs according to the annotation information. Inter-gene SNPs are treated as singletons, and their *p*-values and the corresponding physical positions are denoted as $\{p_r^{\text{Singleton}}, r = 1, \cdots, R\}$ and $\{\ell_r^{\text{Singleton}}, r = 1, \cdots, R\}$, respectively. Intra-gene SNPs within the same gene are bound as an SNP cluster, and the *p*-value of the s^{th} intra-gene SNP within the t^{th} gene and the corresponding physical position are denoted as $\{p_{s,t}^{\text{Singleton}}, s = 1, \cdots, S_t, t = 1, \cdots, T\}$ and $\{\ell_{s,t}^{\text{Singleton}}, s = 1, \cdots, S_t, t = 1, \cdots, T\}$, respectively. We use physical position of the first SNP within a gene to represent the gene location for a result display in the Results section.

To evaluate total effects of genes (SNP clusters) on RA, we combine *p*-values of intra-gene SNPs within a gene by using the truncated product *p*-value method [3]. The combination is based on multiplication of *p*-values, less than some pre-specified cut-off threshold, from single-locus association tests. The test statistic for the t^{th} gene is defined as:

$$Z_t = \prod\nolimits_{1 \le s \le S_t} (p_{s,t}^{\text{Singleton}})^{I[p_{s,t}^{\text{Singleton}} < \theta]}, t = 1, \cdots, T, \qquad (1)$$

where θ is a threshold of p-value truncation. The cumulative distribution function of Z_t is:

$$F(z_t) = \sum_{s=1}^{S_t} C(S_t, s)(1-\theta)^{S_t - s} \cdot \left\{ z_t \sum\nolimits_{j=0}^{s-1} \frac{(s \ln \theta - \ln z_t)^j}{j!} I[z_t \le \theta^s] + \theta^s I[z_t > \theta^s] \right\}.$$

p-Values of genes and the corresponding physical positions are denoted as $\{p_t^{\text{Cluster}}, t = 1, \cdots, T\}$ and $\{\ell_t^{\text{Cluster}}, t = 1, \cdots, T\} = \{\ell_{1,t}^{\text{Singleton}}, t = 1, \cdots, T\}$. In accordance with the sorted physical positions,

$$\{\ell_{(k)}, k = 1, \cdots, R+T\} = \text{sort}\{(\ell_r^{\text{Singleton}}, r = 1, \cdots, R), (\ell_t^{\text{Cluster}}, t = 1, \cdots, T)\},$$

p-values of inter-gene SNPs and genes are arranged in order. Finally, false-discovery rate (FDR) correction [4] is applied to all *R*+*T* *p*-values to adjust for multiple testing.

Results

We calculated exact *p*-values of the Armitage trend test for 530,720 autosomal SNPs in the study of RA. According to the Illumina 550 k SNP-chip annotation file, all SNPs were partitioned into 285,823 inter-gene SNPs and 244,897 intra-gene SNPs, which were located on 15,635 genes. The truncated product *p*-value statistics with $\theta = 0.05$ in Eq. (1) and the empirical *p*-values were calculated for 15,635 genes.

We removed 1,088 SNPs with a minor allele frequency of zero, resulting in the removal of 1,078 inter-gene SNPs and 10 genes. In total, 300,370 *p*-values of genes and inter-gene SNPs were sorted according to their physical positions. FDR was applied to these *p*-values and FDR-adjusted *p*-values in $-\log_{10}$ scale, $-\log_{10}(P_{FDR})$, were displayed (see Figure 1(A)). A high peak of association signals was observed on chromosome 6 (symbol: red square). Zooming in to chromosome 6, we found a 5-Mb region with strong association signals (see Figure 1(B)). Further zooming in to the 5-Mb region (1,090 Mb-1,095 Mb in cumulative physical position), 55 out of 57 genes/SNPs across the human genome satisfying $-\log_{10}(P_{FDR}) > 40$ were located in this region (see Figure 1(C)). In

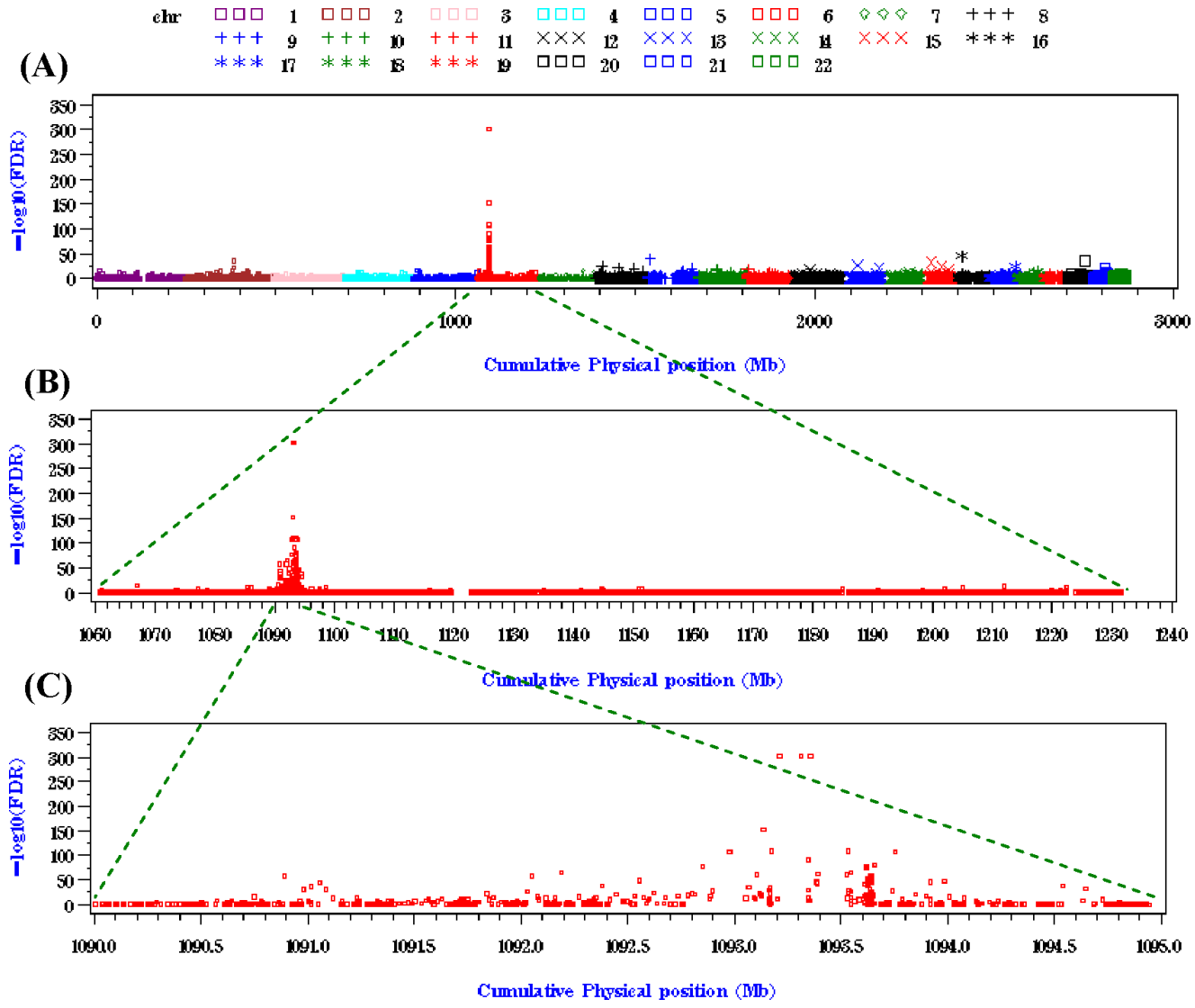

Figure 1

Genome-wide gene-based association mapping of RA data. The horizontal axis denotes the physical position of SNPs (scale in Mb) and the vertical axis denotes the FDR-adjusted *p*-values (scale in $-\log_{10}$). A, FDR-adjusted *p*-values (scale in $-\log_{10}$) of genome-wide gene-set association tests; B, Association signals on chromosome 6; C, Association signals on a chromosomal region between 1,090 Mb and 1,095 Mb (in cumulative physical position).

addition, the top three association signals were found at *C6orf10*, *BTNL2*, and *HLA-DRA*. Their individual $-\log_{10}(P_{FDR})$ values were as high as 302.65.

The $-\log_{10}(P_{FDR})$ values of the top 100 significant loci were all greater than 20 (data not shown). Among the 100 loci, 86 genes/SNPs were located in a region between 30,041,240 bp and 33,797,498 bp on chromosome 6. The region overlapped with the human major histocompatibility complex (MHC) region, which is well known proven to be one of the most important genomic regions related to RA [5]. There were 14 genes on other chromosomes, and some have already been proven to be associated with RA. For example, gene *C5* had a $-\log_{10}(P_{FDR})$ of 20.46 in our study, and this gene has been shown to be an RA-associated gene [1].

We compared the results of single-locus association tests and gene-based association tests under two thresholds of significance, $-\log_{10}(P_{FDR}) = 3$ and $-\log_{10}(P_{FDR}) = 7$. For a threshold of $-\log_{10}(P_{FDR}) = 3$, the exact Armitage trend tests identified 433 SNPs among a total of 529,632 SNPs and gene-based association tests identified 849 genes/SNPs among a total of 300,370 genes/SNPs. In total, 463 genes/SNPs were identified by gene-based association tests but failed to be detected by single-locus association tests. On the other hand, 69 intra-gene SNPs revealed by the exact Armitage trend tests failed to be identified by gene-based association tests, including 65 SNPs located on individual genes and 2 SNP pairs located on two individual genes. For a threshold of $-\log_{10}(P_{FDR}) = 7$, the exact Armitage trend tests identified 141 SNPs among a total of 529,632 SNPs and gene-based association tests identified 308 genes/SNPs among a total of 300,370 genes/SNPs. We found that 157 genes/SNPs found by gene-based association tests failed to be identified by the exact Armitage trend tests; however, only 10 intra-gene SNPs identified by the exact Armitage trend tests but not by gene-based association tests, where all of the 10 SNPs were on individual genes. The intra-gene SNPs missed by gene-based association tests were not in the list of top 100 genes/SNPs.

We compared our results with other studies. Association of two previously reported genes [1,6], *PTPN22* and *IL2RB*, were confirmed by our method. The adjusted *p*-value $-\log_{10}(P_{FDR})$ of the two important genes were 13.29 and 4.75 in our study, respectively. We also compared our results with the other contributions in GAW16 Group 16 - Gene- or region-based association tests. In spite of the use of various methods and procedures, some consistent results were obtained. The genes included *AGPAT1* (62.04) [7], *HLA-C* (65.47) [8], and *PHF19* (10.21) [9], where the numbers in parentheses were $-\log_{10}(P_{FDR})$ in our study. In addition, we also identified some novel RA-associated genes/SNPs that have not been reported before, for example, *NOTCH4*, *TAP2*, and *TNXB*. The adjusted *p*-value $-\log_{10}(P_{FDR})$ of the three genes were 153.79, 108.33, and 108.09 in our study, respectively. The roles of these genes/SNPs in RA are not clear and merit further study.

To consider strong effects of HLA genes and extensive LD in the human MHC region, we replaced the exact Armitage trend test at the first stage with a logistic regression model adjusting for the status of shared-epitope alleles. After adjusting for the effect of *DRB1* shared-epitope alleles, we found that the results of the top three loci, *C6orf10*, *BTNL2*, and *HLA-DRA*, remained the same, and all of the aforementioned genes were still highly significant. The major difference was that 44% of the top inter-gene SNPs in the human MHC region were not longer significant after the adjustment of shared-epitope alleles.

Discussion

Under the proposed two-stage association mapping framework, there are different methods that can be applied to integrate SNP information within a gene, for example, combination of test statistics, principal-components analysis, and multiple regression analysis. This paper considers a *p*-value combination, which has been broadly used in a GWAS [3,10,11]. SNPs in a disease-gene region are more likely to present association signals compared with SNPs in a disease-gene-free region. Therefore, combination of the *p*-values will strengthen association signals and increase power of association tests in a disease-gene region. However, this method may miss a relatively small number of intra-gene SNPs that can be detected by single-locus association tests. The proposed gene-based association test provides a powerful alternative but is not intended to substitute for a single-locus association test.

Unlike some researchers who have performed *p*-value combination in sliding windows [3,10,11], we combine *p*-values to evaluate a total impact of SNPs within each gene in a GWAS. There are multiple types of *p*-value combination methods. This paper considers a truncated product *p*-value statistic because of its good performance in our previous simulation study [12,13]. However, in the analysis of RA data, we also calculated empirical *p*-values of different combination methods including the minimum *p*-value statistic and Fisher's product *p*-value statistic for the top significant genes. All of the methods obtained similar empirical *p*-values, implicating the strong association of the identified genes.

An extended application of *p*-value combination methods is to study biological pathways or protein networks

of complex diseases. *p*-Values of SNPs within genes involved in a pathway/network can be combined to evaluate the global effect of a pathway/network and then used to identify disease-specific pathways and networks. The applications highlight the potential of *p*-value combination methods in genetic/genomic dissection of complex diseases.

Strong effects of HLA genes on RA and extensive LD in the human MHC region, where the genes are located, are issues that should be taken into consideration in the analysis of the RA data. An analysis that does not consider the issues may overstate genetic association in this region. To circumvent the issues, an alternative approach may be to replace the exact Armitage trend test at the first stage with a logistic regression model containing covariates of the HLA loci and/or SNPs in LD with the HLA loci. Marginal effects of tested genes/SNPs can be evaluated independently after conditioning out the effects of LD and HLA genes. We only adjust the status of *DRB1* shared-epitope alleles in this paper and the analysis can be further enhanced by considering additional information on LD structure and HLA genes in the future.

Conclusion

This study introduces a two-stage genome-wide gene-based association scanning procedure. Compared with some existing single-locus and multilocus methods, this method has practical merits in aspects of biology, computation, and statistics. We applied this method to analyze GAW16 Problem 1 RA data. Compared with the results from other RA association studies, our analysis not only successfully confirmed association of previously reported genes but also identified novel RA-associated genes/SNPs.

List of abbreviations used

FDR: False-discovery rate; GWAS: Genome-wide association study; LD: Linkage disequilibrium; MHC: Major histocompatibility complex; RA: Rheumatoid arthritis; SNP: Single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

H-CY conceived the experimental design and statistical methods and prepared the manuscript. Y-JL performed data analysis. C-MC prepared gene annotation. J-WC and W-HP contributed to the discussion and preparation of the final manuscript with H-CY. All authors have approved the final manuscript.

Acknowledgements

We thank two anonymous reviewers and Dr. Beyene for their constructive suggestions, which have improved this manuscript. This work was partially supported by a National Research Program of Taiwan for Genomic Medicine grant (NSC 97-3112-B-001-027) and a National Science Council of Taiwan grant (NSC 97-2314-B-001-006-MY3). The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

This article has been published as part of *BMC Proceedings* Volume 3 Supplement 7, 2009: Genetic Analysis Workshop 16. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/3?issue=S7.

References

1. 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 and Gregersen PK: ***TRAF1-C5* as a risk locus for rheumatoid arthritis - a genomewide study.** *N Engl J Med* 2007, **357:**1199–1209.
2. Armitage P: **Tests for linear trends in proportions and frequencies.** *Biometrics* 1955, **11:**375–386.
3. Zaykin DV, Zhivotovsky LA, Westfall PH and Weir BS: **Truncated product method for combing *p*-values.** *Genet Epidemiol* 2002, **22:**170–185.
4. Benjamini Y and Hochberg Y: **Controlling the false discovery rate: A practical and powerful approach to multiple testing.** *J Roy Stat Soc Ser B* 1995, **57:**289–300.
5. Stastny P: **Association of the B-cell alloantigen DRw4 with rheumatoid arthritis.** *N Engl J Med* 1978, **298:**869–871.
6. The Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661–678.
7. Aporntewan C, Ballard DH, Lee JY, Lee JS, Wu Z and Zhao H: **Gene hunting of the Genetic Analysis Workshop 16 rheumatoid arthritis data using rough set theory.** *BMC Proc* 2009, **3(suppl 7):**S126.
8. Black MH and 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.
9. Buil A, Martinez-Perez A, Perera-Lluna A, Rib L, Caminal P and Soria JM: **A new gene-based association test for genome-wide association studies.** *BMC Proc* 2009, **3(suppl 7):**S130.
10. Dudbridge F and Koeleman BPC: **Rank truncated product of p values, with application to genomewide association scans.** *Genet Epidemiol* 2003, **25:**360–366.
11. Sun YV, Levin AM, Boerwinkle E, Robertson H and Lardia SLR: **A scan statistic for identifying chromosomal patterns of SNP association.** *Genet Epidemiol* 2006, **30:**627–635.
12. Yang HC, Lin CY and Fann CSJ: **A sliding-window weighted linkage disequilibrium test.** *Genet Epidemiol* 2006, **30:**531–545.
13. Yang HC, Hsieh HY and Fann CSJ: **Kernel-based association test.** *Genetics* 2008, **179:**1057–1068.