

Genome-wide association studies using single-nucleotide polymorphisms versus haplotypes: an empirical comparison with data from the North American Rheumatoid Arthritis Consortium

Heejung Shim^{†1}, Hyonho Chun^{†1}, Corinne D Engelman²
and Bret A Payseur^{*3}

Addresses: ¹Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, Wisconsin 53706 USA, ²Department of Population Health Sciences, University of Wisconsin-Madison, 707 WARF Building, 610 North Walnut Street, Madison, Wisconsin 53726 USA and ³Laboratory of Genetics, University of Wisconsin-Madison, 425-G Henry Mall, Madison, Wisconsin 53706 USA

E-mail: Heejung Shim - shim@stat.wisc.edu; Hyonho Chun - chun@stat.wisc.edu; Corinne D Engelman - cengelman@wisc.edu; Bret A Payseur* - payseur@wisc.edu

*Corresponding author †Equal contributors

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

Published: 15 December 2009

BMC Proceedings 2009, 3(Suppl 7):S35 doi: 10.1186/1753-6561-3-S7-S35

This article is available from: <http://www.biomedcentral.com/1753-6561/3/S7/S35>

© 2009 Shim et al; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The high genomic density of the single-nucleotide polymorphism (SNP) sets that are typically surveyed in genome-wide association studies (GWAS) now allows the application of haplotype-based methods. Although the choice of haplotype-based vs. individual-SNP approaches is expected to affect the results of association studies, few empirical comparisons of method performance have been reported on the genome-wide scale in the same set of individuals. To measure the relative ability of the two strategies to detect associations, we used a large dataset from the North American Rheumatoid Arthritis Consortium to: 1) partition the genome into haplotype blocks, 2) associate haplotypes with disease, and 3) compare the results with individual-SNP association mapping. Although some associations were shared across methods, each approach uniquely identified several strong candidate regions. Our results suggest that the application of both haplotype-based and individual-SNP testing to GWAS should be adopted as a routine procedure.

Background

Advances in genotyping technology have stimulated genome-wide association studies (GWAS) of common human diseases. The high genomic density of single-nucleotide polymorphisms (SNPs) available on current genotyping platforms raises the prospect of combining neighboring SNPs into haplotypes for association analysis.

Haplotype-based association testing offers several advantages over the standard "one-SNP-at-a-time" approach [1]. First, genome-wide haplotype approaches reduce the dimension of association testing when a single global test for a block is used. Performing fewer tests preserves power and helps to maintain reasonable false-positive rates. Second, haplotype methods facilitate detection of

associations driven by *cis*-interactions among nearby SNPs that might be missed by methods that consider SNPs one at a time. Finally, haplotype approaches 1) recognize that variation in populations is inherently structured into genomic blocks and 2) exploit these correlations among SNPs. For all these reasons, using haplotypes in association testing is expected to increase power relative to single-SNP approaches, and studies based on human haplotype structure have provided support for this claim [2].

Nevertheless, haplotype association mapping faces several challenges. Haplotype block structure and phase is rarely observed in human genotyping data, requiring the application of statistical procedures that introduce additional error. Haplotype reconstruction methods that use different criteria produce different results [3], leaving the choice of the best approach for association mapping unclear. Moreover, when a block contains a large number of haplotypes, the increased degrees of freedom within a block can erode power.

The performance of haplotype-based and single-SNP association mapping have been compared, with mixed results. Power calculations using explicit analytical formula [4] and simulations based on the HapMap [2] suggested that using haplotypes can significantly improve power. Alternatively, Long and Langley [5] argued that single-SNP tests are at least as powerful in their simulation studies. Empirical comparisons have also yielded inconsistent conclusions [4].

The observed inconsistencies suggest that the performance of methods may depend on the nature of the data. Some associations might be detected more readily using individual SNPs, while others might only be discovered using haplotypes. Interestingly, most previous studies have restricted their comparisons of method performance to a small subset of the genome. To address these issues, we conducted GWAS with both haplotype-based and individual-SNP methods. By applying both approaches to the same genome-wide dataset (the North American Rheumatoid Arthritis Consortium (NARAC) provided by the Genetic Analysis Workshop 16 (GAW16)) we could directly compare the performance of these methods in an empirical context.

Methods

Phenotypes and genotypes

The data set from the NARAC contained 868 case subjects with rheumatoid arthritis (RA) and 1194 matched control subjects. Individuals were genotyped at 545,080 SNPs. Our study focused on the autosomes, where genotypes at 531,689 SNPs were available. After removing 17,754 SNPs that showed 1) a deviation from

Hardy-Weinberg equilibrium in controls or 2) less than 0.001 minor allele frequency in the total sample, 513,935 SNPs were retained for further analysis.

Haplotype block partitioning algorithms

Two methods were used to define haplotype blocks. The first method [6] was focused on D' , a normalized measure of linkage disequilibrium (LD). This method identified informative pairs of SNPs as Category (1): those for which the upper 95% confidence bound on D' was between 0.7 and 0.98 and Category (2): those for which the upper confidence bound on D' was less than 0.9. We defined a haplotype block as a region where at least 95% of pairs among informative SNPs belonged to Category (1). The second method [7] used the four-gamete rule. For each pair of SNPs, the population frequencies of the four possible haplotypes were calculated, and the number of haplotypes with observed frequencies of at least 0.01 were counted. Blocks were constructed by combining consecutive pairs of SNPs for which only three haplotypes were observed. All methods were implemented in the computer program Haploview [8].

To examine the effects of changes in parameters on block definition, we varied the proportion of informative SNP pairs in Category (1) needed to combine blocks (in Gabriel's method) and the minimum frequency needed to call a haplotype "observed" (in the four-gamete rule algorithm). We compared the block sizes that resulted from applying these algorithms to the 8,051 SNPs on chromosome 21 and assuming a range of values for these two parameters. Although variation in block size was observed, block sizes for parameter values near the defaults were fairly stable; consequently, default values (0.95 in Gabriel's method, 0.01 in the four-gamete rule algorithm) were used in all subsequent analyses.

Testing for associations between disease status and genotype

A high variance inflation factor (1.45) [9] suggested that association analyses of these data might be affected by population stratification. To account for these effects, we calculated top eigenvectors of the covariance matrix across the samples [10] using SNPs sampled every fifth position after excluding SNPs on the short arms of chromosomes 6 and 8, as Plenge et al. suggested [11]. Ten outliers, detected from ten eigenvectors, were excluded in all subsequent analyses. Both individual SNP associations and haplotype associations were measured by likelihood ratio tests via logistic regression where three eigenvectors were included as covariates to correct population stratification. These tests for individual SNP associations were implemented in the computer program PLINK [12,13]. For haplotype association tests, we estimated haplotypes in each block by the

standard expectation maximization algorithm, implemented in PLINK, and conducted likelihood ratio tests via logistic regression with haplotypes by using the statistical package R [14]. Because we aimed to detect collective associations between groups of haplotypes and arthritis, we used a single global test of association for haplotypes. *p*-Values were compared to the Bonferroni threshold ($\alpha = 0.05/\#$ tests) to identify statistically significant loci (SNPs not belonging to haplotypes were counted in both sets of analyses).

Results

In what follows, we refer to a haplotype block as a partition containing at least two SNPs and a singleton as a partition containing only one SNP. GAB blocks and GAM blocks represent blocks constructed by the algorithm of Gabriel et al. [6] and the four gamete rule, respectively.

Haplotype block partitioning

GAM partitioned the genome into more blocks (100,121) than GAB (97,881). On average, GAM blocks included more SNPs and were larger in size than GAB blocks (median number of SNPs: 3 in GAB; 4 in GAM), suggesting greater genomic coverage by GAM blocks. We note that a block is defined when a partition consists of at least two SNPs. Because the GAB method produces more singletons, the GAM method has both more blocks and a higher average block size. Block sizes estimated by both methods (GAM median size = 8,639 bp; GAB median size = 7,335 bp) were similar to those observed for other populations of European descent [15]. We also uncovered considerable variation in haplotype block structure across the genome, with block size ranging from 2 bp to 3,547,000 bp for both methods. Similar numbers of SNPs were assigned to haplotype blocks on the different chromosomes (median = 3 for most chromosomes). Because chromosomes vary in physical size, this suggests that variation in block size among chromosomes primarily reflected differences in the density of genotyped SNPs. Most (92%) SNPs were localized to GAB or GAM blocks, indicating that the density of genotyped SNPs was sufficient to conduct haplotype-based association analysis.

We ran Haploview on Intel Xeon 3 GHz dual Quad core system with 32 Gb of RAM. GAB block partitioning and GAM block partitioning on chromosome 22 including 8,205 SNPs required 19 and 14 minutes, respectively.

Haplotype association test vs. individual SNP association test

A large number of tests on chromosome 6 showed strong associations (data not shown). The significance level,

0.05 becomes 9.73×10^{-8} for single SNP, 2.71×10^{-7} for GAB, 3.11×10^{-7} for GAM after Bonferroni correction with total numbers of tests of 513,935 for single-SNP, 184,504 (97,881 GAB blocks plus 86,623 singletons) for GAB, 160,737 (100,121 GAM blocks plus 60,616 singletons) for GAM, respectively. Several associations survived the stringent Bonferroni correction for multiple testing (51 GAB blocks, 50 GAM blocks, and 21 individual SNPs) and some associations were shared among methods. A total of 8 out of 51 significant GAB blocks (Table 1) and 6 out of 50 significant GAM blocks included SNPs that were also significant in individual SNP association tests. Four SNPs showed significant associations using all three methods (Table 1). However, many associations were only detected when certain methods were applied. 43 GAB blocks and 44 GAM blocks that showed significant associations were not detected by individual SNP association tests and 11 of 21 SNPs that showed significant associations in individual tests were not significant in haplotype association tests.

We asked whether reducing the significance threshold ($9.73 \times 10^{-8} \times 2000$) in individual SNP association tests and testing associations involving haplotypes within ± 100 kb of the resulting significant SNPs improved the consistency between individual SNP association tests and haplotype association tests. However, even using these extreme criteria, only 25 GAB blocks and 29 GAM blocks overlapped with the regions containing SNP association tests that were significant.

Most of the SNPs and haplotypes that showed significant associations on chromosome 6 were located near the HLA region. Similar to the pattern for significant associations in other genomic regions, *p*-values from haplotype association tests were smaller than *p*-values from individual SNP association tests (Figure 1).

Haplotype-based association tests of 1,578 blocks on chromosome 22 using R required 4 minutes on Intel Xeon 3 GHz dual Quad core system with 32 Gb of RAM.

Haplotype association test: GAB vs. GAM

Twenty-five significant GAB blocks and 25 significant GAM blocks overlapped with each other. Interestingly, 18 GAB and GAM blocks out of 25 are identical, which is why these 18 haplotype-based associations using GAM and GAB blocks showed consistent results. For the remaining seven overlapped blocks, we hypothesize that signals of association with disease were strong, so that their detection was not sensitive to block partitioning. The fact that three of these blocks included SNPs that were significant in individual SNP association tests supports this hypothesis. In the regions where

Table 1: Significant associations detected by GAB algorithm

Chr	Start position	End position	p-value	Include significant SNPs ^a	Overlap with significant GAM blocks ^b
1	1,045,729	1,058,627	1.67×10^{-14}	Yes	Yes
1	55,680,975	55,682,288	2.42×10^{-8}	Yes	No
1	150,433,159	150,471,553	1.26×10^{-8}	No	Yes
1	22,244,528	22,251,559	4.01×10^{-10}	No	No
3	58,921,915	59,057,595	6.27×10^{-11}	Yes	Yes
3	108,502,403	108,548,227	2.28×10^{-11}	No	Yes
3	156,354,381	156,381,109	1.26×10^{-7}	No	Yes
4	36,082,845	36,105,879	5.76×10^{-8}	No	No
5	38,767,613	38,785,799	6.62×10^{-8}	No	No
5	127,305,116	127,355,548	2.69×10^{-7}	No	No
5	137,669,365	137,669,929	2.67×10^{-8}	Yes	No
5	168,125,453	168,128,612	1.54×10^{-7}	No	Yes
7	26,122,453	26,143,430	6.70×10^{-9}	No	No
7	129,556,365	129,578,739	4.97×10^{-12}	Yes	No
7	137,469,416	137,478,601	4.64×10^{-8}	No	No
8	20,619,747	20,619,773	4.41×10^{-8}	No	Yes
8	22,942,075	22,948,219	7.39×10^{-10}	No	No
8	144,734,804	144,746,191	1.31×10^{-8}	No	Yes
9	81,662,684	81,666,969	4.29×10^{-8}	Yes	Yes
10	48,046,897	48,050,873	2.60×10^{-11}	No	No
10	71,263,263	71,264,307	9.28×10^{-11}	No	Yes
10	92,582,936	92,663,587	1.05×10^{-37}	No	Yes
10	96,750,802	96,788,739	2.66×10^{-11}	No	No
10	99,590,357	99,601,703	2.41×10^{-7}	No	No
11	44,243,624	44,243,868	2.12×10^{-10}	No	No
11	1,274,888	1,313,639	2.20×10^{-9}	No	No
12	11,438,799	11,477,909	2.44×10^{-7}	No	Yes
13	113,781,019	113,839,177	3.73×10^{-12}	Yes	No
14	80,904,425	80,935,179	4.93×10^{-9}	No	Yes
14	104,491,260	104,507,267	1.15×10^{-9}	No	No
14	29,207,158	29,246,917	2.97×10^{-10}	No	No
15	65,048,226	65,074,484	1.42×10^{-12}	No	Yes
15	87,225,585	87,243,610	1.71×10^{-12}	No	Yes
16	1,478,364	1,484,303	2.01×10^{-8}	No	Yes
16	11,827,538	11,924,420	4.24×10^{-10}	No	No
16	12,089,203	12,101,855	6.25×10^{-10}	No	No
17	67,936,240	67,937,413	6.62×10^{-9}	No	Yes
17	17,997,997	18,077,866	2.87×10^{-9}	No	No
19	39,820,387	39,827,349	2.35×10^{-8}	No	No
19	46,151,081	46,152,461	3.26×10^{-8}	No	No
19	51,106,074	51,111,613	8.85×10^{-8}	No	No
19	61,351,554	61,371,582	5.88×10^{-11}	No	Yes
19	13,193,010	13,196,432	9.92×10^{-11}	No	Yes
19	18,085,729	18,093,754	1.928×10^{-7}	No	Yes
19	2,051,346	2,074,516	2.47×10^{-7}	No	Yes
19	2,957,274	2,966,481	1.90×10^{-11}	No	Yes
22	38,301,711	38,309,783	2.62×10^{-7}	No	Yes
22	42,547,708	42,605,295	1.11×10^{-13}	No	Yes
22	49,370,904	49,400,136	2.93×10^{-8}	Yes	Yes
22	28,201,971	28,202,179	8.14×10^{-12}	No	No
22	35,868,745	35,875,810	5.28×10^{-8}	No	No

^aInclude SNPs which are significant in individual SNP association test.

^bOverlap with blocks which are constructed by four gamete rule and significant in a haplotype association test.

association tests in GAB and GAM blocks showed inconsistent results, we observed that differences between GAB block partitions and GAM block partitions were not unusual. In many cases, one block contained the other block with an additional one or two SNPs, but results from haplotype-based association tests in the two blocks were still substantially different.

Conclusion

GWAS now commonly survey SNPs at a genomic density similar to this study. Consequently, our observation that most of the genome could be organized into multi-SNP haplotypes indicates that available resources are sufficient to conduct haplotype-based mapping on the genomic scale.

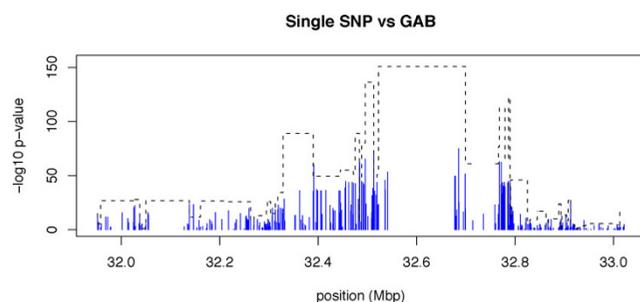


Figure 1
A comparison of haplotype and individual association tests on HLA region. $-\log_{10}$ transformed p -values are plotted across chromosomal location 32 Mbp to 33 Mbp. The results from individual association tests are represented by solid blue sticks; the results from haplotype association tests with GAB blocks are represented by dotted black lines.

Those regions that were significantly associated with RA in both individual-SNP tests and haplotype-based tests represent promising candidates for further study. For example, the HLA region on chromosome 6 remained significant for all three genome-wide association tests, strongly suggesting that genes in this region contribute to disease risk.

Although some associations were observed consistently across methods, some associations were only detected using haplotype-based tests. Several factors might explain these differences. Haplotype-based methods required approximately 65% fewer tests than the individual-SNP approach. As a result, the multiple testing correction was less severe for haplotype-based methods. Haplotype-based methods can also detect *cis*-interactions among several causal variants [16]. Furthermore, because the power to detect associations is maximized when marker and causal variant frequencies are similar, analyses using haplotypes could find associations with rare alleles that analyses using individual SNPs may miss. We also discovered associations using individual SNPs that were not seen in haplotype tests. Perhaps these represented cases in which only a single SNP exhibited strong LD with a causal variant, so that forming haplotypes with several adjacent SNPs diluted the strength of association. Regardless of the explanation for observed differences among methods, our results indicate that the application of both individual-SNP and haplotype-based approaches to GWAS will maximize the potential for finding biologically important associations.

Although some regions show consistent significant associations in different block partitions (GAB and GAM), in most regions, haplotype-based association

tests are really sensitive to changes in block partitions. This result suggests that the effects of other block partitioning algorithms on GWAS should be compared. For example, haplotype-based association testing using a sliding window of fixed physical or genetic size would be an alternative approach. Although this strategy is easily implemented, it ignores information about haplotype block structure. The variation in block structure across the genome suggests that methods that use this structure (such as those applied in this paper) should be more powerful for GWAS, but this issue needs to be examined.

Our study also suggests several avenues for future research. Additional measurements of the effects of different haplotype partitioning algorithms on the power of downstream association tests - in both simulations and empirical data - would be useful. For example, the error inherent in haplotype block estimation needs to be incorporated in association analysis. Furthermore, the likelihood ratio tests used here ignored the evolutionary relationships among haplotypes. An improved analysis that uses this information (e.g., a cladistic analysis [17]) would be worthwhile.

List of abbreviations used

GAB: Blocks constructed by the algorithm of Gabriel et al. [6]; GAM: Blocks constructed by the four-gamete rule; GAW: Genetic Analysis Workshop; GWAS: Genome-wide association studies; LD: Linkage disequilibrium; NARAC: North American Rheumatoid Arthritis Consortium; RA: Rheumatoid arthritis; SNP: Single-nucleotide polymorphism.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

HS and HC participated in the design of the study, performed the statistical analyses, and drafted the manuscript. CDE obtained IRB approval for the study, gained access to the dataset, participated in the design of the study, and helped revise the manuscript. BAP conceived of the study, participated in the design of the study, and helped draft and revise the manuscript. All authors read and approved the final manuscript.

Acknowledgements

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. Clark AG: **The role of haplotypes in candidate gene studies.** *Genet Epidemiol* 2004, **27**:321–333.
2. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ and Altshuler D: **Efficiency and power in genetic association studies.** *Nat Genet* 2005, **37**:1217–1223.
3. Indap AR, Marth GT, Struble CA, Tonellato P and Olivier M: **Analysis of concordance of different haplotype block partitioning algorithms.** *BMC Bioinformatics* 2005, **6**:303.
4. Akey J, Jin L and Xiong M: **Haplotypes vs single marker linkage disequilibrium tests: what do we gain?.** *Eur J Hum Genet* 2001, **9**:291–300.
5. Long AD and Langley CH: **The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits.** *Genome Res* 1999, **9**:720–731.
6. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ and Altshuler D: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296**:2225–2229.
7. Wang N, Akey JM, Zang K, Chakraborty R and Jin L: **Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation.** *Am J Hum Genet* 2002, **71**:1227–1234.
8. Barrett JC, Fry B, Maller J and Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21**:263–265.
9. Delvin B and Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55**:997–1004.
10. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D: **Principal components analysis corrects for stratification in genome-wide association studies.** *Nat Genet* 2006, **38**:904–909.
11. 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.
12. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ and Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559–575.
13. **PLINK...Whole-genome association analysis toolset.** <http://pngu.mgh.harvard.edu/purcell/plink/>.
14. R Development Core Team: **R: A Language and Environment for Statistical Computing.** 2008 <http://www.R-project.org>, [ISBN 3-900051-07-0].
15. International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437**:1299–1320.
16. Liu N, Zhang K and Zhao H: **Haplotype-association analysis.** *Adv Genet* 2008, **60**:335–405.
17. Templeton AR, Boerwinkle E and Sing CF: **A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*.** *Genetics* 1987, **117**:343–351.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

