

Cross Haplotype Sharing Statistic: Haplotype length based method for whole genome association testing

André R. de Vries^a, Ilja M. Nolte^b, Geert T. Spijker^c, Dumitru Brinza^d, Alexander Zelikovsky^d, Gerard J. te Meerman^a

^aDepartment of Medical Genetics, Medical Center Groningen and University of Groningen, the Netherlands;

^bDepartment of Pathology and Laboratory Medicine, section Medical Biology, University Medical Center Groningen and University of Groningen, the Netherlands;

^cDepartment of of Dermatology, University Medical Center Groningen, the Netherlands;

^dDepartment of Computer Science, Georgia State University, Atlanta, USA;

Abstract

We analyzed a dataset of 2,300 SNPs from a region of 10 Mb of chromosome 18q, which had shown linkage evidence for Rheumatoid Arthritis [1]. To test for disease association, a novel method, called the Cross-test, was used, which is based on differences between direct sharing of a patient and a control haplotype and sharing of two patients or two control haplotypes. We found highly significant association at a position ~4,407 kb from the starting SNP and a second less significant one at position ~4,863 kb. These results are supported by a single marker chi-square test and the Haplotype Sharing Statistic, but the Cross-test outperforms them with respect to significance and signal-to-noise ratio. For determining the significances, a sequential randomization procedure is implemented to render the method fast enough and hence practically useful even for whole genome screens with marker densities of 500,000 SNPs/genome.

Background

With the current advances in genotyping technology, genome wide association studies for detecting genes involved in complex diseases have recently become feasible. However, the computational and statistical methodology for analyzing such studies needs optimization and standardization. Various methods and strategies have been investigated, e.g. reviewed in [2-4].

Historically, in candidate region studies, association was mainly calculated using single locus tests, such as the allele frequency chi-square test or the TDT-test [5]. The advantage of single locus tests is that haplotypes are not needed and that haplotype inference can be avoided. However, so far single locus tests on whole genome data sets seem to be unsuccessful because of large numbers of false positive results. The reason for this is that the true association information is not enclosed in single SNPs, but in haplotypes [6]. Therefore, association analysis should be based on haplotypes.

Complex diseases are thought to be caused by multiple genetic variants, each with a moderate effect, as well as by various environmental factors possibly also interacting with these genes. According to the common variant hypothesis, the genetic variants related to the disease are old mutations and are common in the population (with allele frequencies > 5%). Subsequent mutations and recombinations acting on the ancestral haplotype have led to various haplotype patterns among the present population descending from that ancestral haplotype. The haplotype pattern that is still shared among haplotypes descending from these ancestors is shortened, but is still recognizable. Or more precisely, it can still be distinguished from haplotypes from other ancestors. Therefore, since there is a shared haplotype pattern expected within patients, we hypothesize that there is a difference in haplotype patterns between patients and controls. The length of the reduced ancestral haplotype is variable because of the uneven nature of recombination and mutation processes.

It is expected that there is more sharing within patient haplotypes than within control haplotypes, because control haplotypes are thought to descend from more and older ancestral haplotypes. This is the basic idea of the Haplotype Sharing Statistic (HSS), which we have evaluated before [7,8].

Here, we present a new method based on the differences in haplotype patterns between patients and controls at regions associated with the disease, evaluated as the Cross Haplotype Sharing Statistic (Cross). The method is implemented in a program called GronCross and it is suitable for whole genome association studies.

Methods

Materials

We have used a dataset of 2,300 SNPs, covering 10 Mb at chromosome 18q, of the North American Rheumatoid Arthritis Consortium (NARAC), as distributed by the Genetic Analysis Workshop 15 (GAW15). It was part of problem 2 of GAW15, of which we did not have the answers. The dataset contains 460 cases and 460 controls, the latter being recruited from a New York City population.

Haplotype inference

We phased the 920 genotypes using the phasing program 2SNP [9,10], which is a fast algorithm for haplotype inference based on genotype statistics for pairs of SNPs.

Association analysis

A haplotype is a vector of alleles ordered along the chromosome. The amount of sharing between any two haplotypes can be evaluated from the perspective of each locus k as the number of SNPs in a shared haplotype including locus k (figure 1). The shared haplotype is therefore of variable length. At locus k we calculate the average Cross sharing as:

$$SH_{CROSS}(k) = \frac{1}{M * N} \cdot \sum_{i=1}^M \sum_{j=1}^N h(X_i, X_j; k)$$

where M is the number of patients, N the number of controls and $h(X_i, X_j; k)$ the length of sharing between patient haplotype X_i and control haplotype X_j at locus k . Our hypothesis is that at regions associated with the disease, $SH_{CROSS}(k)$ is lower than the mean sharing of all haplotype pairs (regardless of disease status), $SH_{MEAN}(k)$, which is defined as:

$$SH_{MEAN}(k) = \frac{1}{(M + N) * (M + N - 1)} \cdot \sum_{i=1}^{M+N} \sum_{j=1, j \neq i}^{M+N} h(X_i, X_j; k)$$

The prominent statistical problem in evaluating mean haplotype sharing is the way to calculate the distribution of the mean sharing between all pairs of haplotypes. Generally, haplotypes will share alleles in groups and this means that haplotype agreement between haplotype pairs is not independent. We solved this problem for the HSS [7] using the theory of U-statistics. However, an equivalent using U-statistics for the Cross test cannot be determined because of the high correlation between SH_{CROSS} and SH_{MEAN} , in contrast to the independence of haplotype sharing of patients and controls. Therefore, the variance and significance of the Cross test is determined by a randomization test. Since the

	haplo 1	haplo 2	IBS	haplotype sharing
locus 1	1	2		0
locus 2	1	1		3
locus 3	1	1		3
locus 4	1	1		3
locus 5	1	2		0
locus 6	1	2		0
locus 7	1	1		2
locus 8	1	1		2

Figure 1: Haplotype sharing calculation

randomization procedure is performed anyway, for this paper the variance of the HSS is estimated from the same randomization. The randomization procedure is done by permuting the affection status of each haplotype. This will yield an average simulated Cross z-score and HSS t-score and their variances. As evaluation of the z-score is possible after each randomization step, the randomization procedure can be stopped early as soon as it is clear that non-significant z-scores will be found (which is the case for the larger part of the SNPs). The statistical parameters are evaluated as follows:

$$z_{CROSS}(k) = \frac{SH_{CROSS}(k) - SH_{MEAN}(k)}{SD_{CROSS,RANDOMIZATION}(k)}$$

$$t_{HSS}(k) = \frac{SH_{PATIENTS}(k) - SH_{CONTROLS}(k)}{SD_{HSS,RANDOMIZATION}(k)}$$

Normally, z-scores can be converted to p-values assuming normal distribution, but the tails of the z_{CROSS} distribution are not well approximated by a normal distribution, leading to downward biased p values for extreme z-scores. To correct for that, the z-scores were transformed to a chi-square distribution with v degrees of freedom by using the asymptotic distribution for large v :

$$P(\chi^2 | v) \sim P(x) \quad \text{where } x = \frac{\chi^2 - v}{\sqrt{2v}}$$

Therefore:

$$\chi^2_{CROSS} \sim z_{CROSS} * \sqrt{2v} + v$$

With an appropriate chosen v , the distribution resembles the true z-score distribution, especially in the tails, so that realistic p-values are obtained. We found empirically that the best choice for v depends on the sample size. A value of 3 gives a good approximation for the data in this study. It appeared that the Cross and HSS scores were weakly correlated except at the highly significant loci, so a combined score can be evaluated by adding the chi-square statistics.

Results

The running time of the analyses for 2,300 SNPs and 920 individuals was 80 minutes on a single laptop PC (Celeron 1500 Mhz), which is sufficiently fast to be acceptable for whole genome association studies.

The results are presented in figure 2. It shows that the Cross test has a background signal of up to a $-\log(p\text{-value})$ of about 2. A highly significant association is observed in this region with a $-\log(p\text{-value})$ of 6.6 near 4,407 kb from the first SNP. The odds-ratio based on allele frequencies at this location is 1.41 (1.16-1.72, 95% CI). A second peak with $-\log(p\text{-value})$ of 4.1 is found near 4,863 kb from the first SNP, with a corresponding odds-ratio of 1.21 (0.99-1.51, 95% CI). In the same figure, the Cross results are compared with the HSS test results. Although the HSS is significant at the same two locations, the HSS-test by itself is not a good discriminator. The same holds for the single marker chi-square test, shown at the bottom of figure 2. The chi-square test also shows a significant difference near 4,407 kb and this result therefore supports the Cross result, but over the total of 2,300 SNPs, the signal-to-noise ratio is bad, with too many SNPs showing significant association.

Figure 3 zooms in into the associated region around 4,407 kb in order to pinpoint the causal variant. The Cross test does not show a clear peak at a particular locus. Instead, the graph has a smooth pattern, which was to be expected as the Cross test is based on haplotype sharing. This means that information from neighboring SNPs is used as well and hence, tests at subsequent loci are highly correlated.

However, as shown in figure 3, the single marker chi-square test is not informative about the exact disease locus either. Therefore, a follow-up study on this small region (4,400-4,420 kb) is needed.

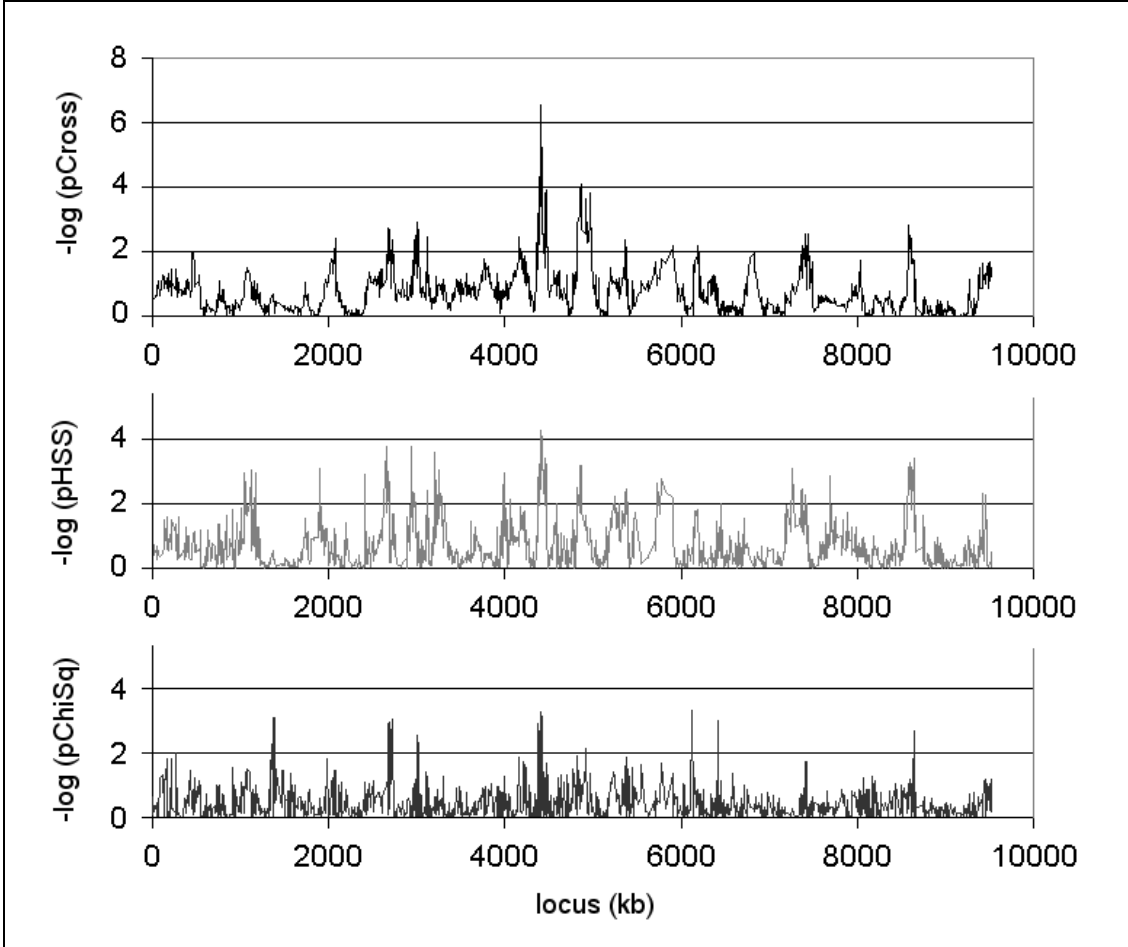


Figure 2: Results of the Cross test (top), HSS test (middle) and single marker chi-squared test (bottom) plotted for all 2,300 SNPs.

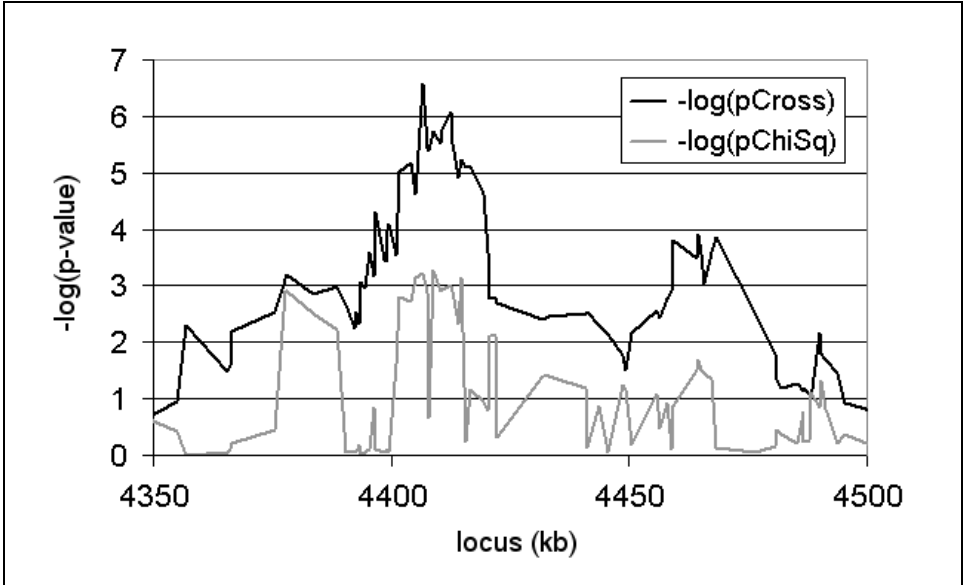


Figure 3: Cross $-\log(p\text{-value})$ and single marker chi-squared $-\log(p\text{-value})$ for a 150 kb region of association.

Discussion and Conclusion

The results of our analyses show that the Cross Haplotype Sharing Statistic is a powerful test statistic for association screening analysis. We found a strong significant association ($-\log(p\text{-value}) = 6.6$) with odds ratio of 1.41. In addition, a second smaller peak is observed ($-\log(p\text{-value}) = 4.1$). As the background signals seemed to range up to a maximum $-\log(p\text{-value})$ of 2, this result possibly indicates a causal variant as well.

The Cross results were compared with the HSS and the single marker chi-square test. Both gave supporting results, but by themselves were not informative enough to declare a significant association because the signal does not really stand out. The Cross test does not point precisely to a particular disease SNP, inherent to the way the test was developed. However, a region of about 20 kb can be identified as the candidate region. This implies that the Cross test is especially useful for initial whole genome association screens. Such a screening will presumably yield several, but a limited amount of significant results, which will have to be analyzed in detail in a follow-up study.

References

1. Jawaheer D et.al.: **Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families.** *Arthritis Rheum.* 2003, **48**: 906-916.
2. Hirschhorn JN and Daly MJ: **Genome-wide association studies for common diseases and complex traits.** *Nature reviews* 2005, **6**: 95-108.
3. Gordon D, Finch, SJ: **Factors affecting statistical power in the detection of genetic association.** *J.Clin.Invest.* 2005, **115**: 1408-1418.
4. Marchini J, Donnelly P and Cardon LR: **Genome-wide strategies for detecting multiple loci that influence complex disease.** *Nat. Genet.* 2005, **37**: 413-417.
5. Spielman RS and Ewens WJ: **The TDT and other family-based tests for linkage disequilibrium and association.** *Am.J.Hum.Genet.* 1996, **59**: 983-989.
6. Te Meerman GJ, Van der Meulen MA, Sandkuijl LA: **Perspectives of identity by descent (IBD) mapping in founder populations.** *Clin Exp Allergy* 1995, **25**: 97-102.
7. In I.M. Nolte: **Statistics and population genetics of haplotype sharing as a tool for fine-mapping of disease gene loci.** PhD thesis 2003, Groningen.
8. Boon M, Nolte IM, Bruinenberg M, Spijker GT, Terpstra P, Raelson J, De Keyser J, Zwanikken CP, Hulsbeek M, Hofstra RM, Buys CH, te Meerman GJ: **Mapping of a susceptibility gene for multiple sclerosis to the 51 kb interval between G511525 and D6S1666 using a new method of haplotype sharing analysis.** *Neurogenetics* 2001, **3**:221-230.
9. Brinza D and Zelikovsky A: **2SNP: Scalable Phasing Based on 2-SNP Haplotypes.** *Bioinformatics* 2006, **22**: 371-373.
10. Brinza D and Zelikovsky A: **Phasing of 2-SNP Genotypes based on Non-Random Mating Model.** *International Workshop on Bioinformatics Research and Applications (IWBRA'06), Proc. of ICCS 2006*, LNCS 3992: 767-774.