Published in final edited form as:

Genet Epidemiol. 2009 April ; 33(3): 228–236. doi:10.1002/gepi.20373.

More Powerful Haplotype Sharing by Accounting for the Mode of Inheritance

Andreas Ziegler¹, Adel Ewhida^{1,2}, Michael Brendel^{1,3,*}, and André Kleensang^{1,**}

¹Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Ratzeburger Allee 160, Haus 4, 23538 Lübeck, Germany

²Al-Fateh University, Faculty of Science, Department of Statistics, P.O. Box 13219, Tripoli, Libya

³Bernhard-Nocht-Institut für Tropenmedizin, Abteilung für Tropenmedizinische Grundlagenforschung, Bernhard-Nocht-Str. 74, 20359 Hamburg, Germany

Abstract

The concept of haplotype sharing (HS) has received considerable attention recently, and several haplotype association methods have been proposed. Here, we extend the work of Beckmann and colleagues [Hum Hered 2005; 59:67-78] who derived a HS statistic (BHS) as special case of Mantel's space-time clustering approach. The Mantel-type HS statistic correlates genetic similarity with phenotypic similarity across pairs of individuals. While phenotypic similarity is measured as the mean-corrected cross product of phenotypes, we propose to incorporate information of the underlying genetic model in the measurement of the genetic similarity. Specifically, for the recessive and dominant modes of inheritance we suggest the use of the minimum and maximum of shared length of haplotypes around a marker locus for pairs of individuals. If the underlying genetic model is unknown, we propose a model-free HS Mantel statistic using the max-test approach. We compare our novel HS statistics to BHS using simulated case-control data and illustrate its use by re-analyzing data from a candidate region of chromosome 18q from the Rheumatoid Arthritis (RA) Consortium. We demonstrate that our approach is point-wise valid and superior to BHS. In the re-analysis of the RA data, we identified three regions with point-wise p-values < 0.005 containing six known genes (PMIP1, MC4R, PIGN, KIAA1468, TNFRSF11A and ZCCHC2) which might be worth follow-up.

Keywords

Case-control studies; genetic association studies; haplotype sharing; Mantel statistics; rheumatoid arthritis

1. Introduction

Haplotype-based methods can be substantially more powerful than single-locus approaches in the presence of multiple ancestral disease alleles, especially when the linkage disequilibrium (LD) between single nucleotide polymorphisms (SNPs) is weak to moderate [Morris and Kaplan 2002]. Haplotype sharing (HS) is one among these and has originally

Corresponding author: Prof. Dr. Andreas Ziegler Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Ratzeburger Allee 160, Haus 4, 23538 Lübeck, Germany Tel: +49 451 500 2780, Fax: +49 451 500 2999, ziegler@imbs.uni-luebeck.de.

*Present address: Boehringer Ingelheim Pharma GmbH & Co. KG, Medical Data Services, 88397 Biberach, Germany

**Present address: European Commission, Directorate General Joint Research Centre, Institute for Health and Consumer Protection, European Centre for the Validation of Alternative Methods, Via E. Fermi 2749, 21027 Ispra (VA), Italy

been proposed by te Meerman and colleagues [1995]. It is based on the idea that patients share longer stretches of haplotypes in genomic regions of interest compared to controls because control haplotypes are thought to descend from more and older ancestral haplotypes. The fundamental work on HS has been extended in several ways [Allen and Satten 2007a; Beckmann et al., 2005b; Bourgain et al., 2000], and it has been successfully employed recently in several applications [e.g. Diepstra et al., 2005; Foerster et al., 2005].

One of the extensions has been proposed by Beckmann and colleagues, where HS statistics are interpreted as Mantel-type statistics [Beckmann et al., 2005a; Beckmann et al., 2005b; Kleensang et al., 2005; Qian 2005]. Here spatial similarity is defined by the shared length between haplotype pairs and temporal similarity as the phenotypic similarity between pairs. Although not explicitly stated, an underlying additive genetic model is assumed. In this work, we extend this HS idea, abbreviated by BHS, and propose new model-based HS Mantel statistics for the analysis of case-control data. We also introduce a flexible approach for gene mapping by adapting the corresponding mode of inheritance. We investigate its validity in Monte-Carlo simulations and demonstrate its increased power when compared with the approach of Beckmann and colleagues. The advantage of our novel approach thus is its ability to detect HS for dominant and recessive modes of inheritance. We illustrate its use by re-analyzing fine-mapping data from the North American Rheumatoid Arthritis Consortium (NARAC). Specifically, 2,300 SNPs were typed in a 10 mega bases (Mb) region of chromosome 18q which had previously shown evidence for linkage to rheumatoid arthritis [Amos et al., 2007].

2. Material and Methods

2.1 Haplotype sharing based on Mantel statistics

Mantel [1967] proposed a product statistic to investigate space-time clustering between pairs of individuals. In haplotypes sharing analyses, “space” is substituted by haplotypes, and the measure of spatial similarity, i.e., the distance between two points in space, is replaced by a measure of haplotype similarity [Beckmann et al., 2005a; Beckmann et al., 2005b]. Comparably, “time” is replaced by phenotype, and the distance in time by a measure of phenotypic similarity. For a sample of subjects, thus $2n$ haplotypes, the Mantel type HS statistic is given by

$$M(x) = \sum_{i=1}^{2n-1} \sum_{j=i+1}^{2n} L_{i,j}(x) Y_{i,j}. \quad (1)$$

Here, $L_{i,j}(x)$ denotes the genetic similarity between haplotypes i and j at the chromosomal position x , and $Y_{i,j}$ is the phenotypic similarity of subjects with haplotypes i and j .

Phenotypic similarity—A common measure of phenotypic similarity between two subjects is the mean-corrected cross product

$$Y_{i,j} = (Y_i - \mu)(Y_j - \mu), \quad (2)$$

where Y_i and Y_j are the observed phenotypes, μ and denotes the population mean of the phenotype. In applications, estimates of the population mean are either obtained from external sources or estimated by the sample mean [Beckmann et al., 2005a]. The rationale behind this choice is the idea to attribute most of the influence to the most extreme phenotypes [Elston et al., 2000].

For a dichotomous trait, Y_i is equal to 1 (0), if a subject is affected (unaffected). The population mean μ boils down to the frequency of the disease in the population, i.e., the population prevalence of the disease, or the frequency of the disease in the sample, i.e.,

$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n Y_i$. For rare diseases, $\mu \approx 0$, is approximately 1 for concordant affected individuals. $Y_{i,j}$ is approximately 0 for concordant unaffected subjects, and it attains small negative values for discordant pairs.

Measures of genetic similarity—HS follows ideas of coalescence theory: Affected subjects should share longer stretches of haplotypes, i.e., more alleles around a putative disease locus than unaffected subjects. Genetic similarity between haplotypes i and j at locus x is therefore defined as the shared length between these haplotypes. More precisely, $L_{i,j}(x)$ represents the number of intervals surrounding locus x that are flanked by markers with the same allele (Figure 1). If haplotypes differ at both neighboring marker loci even though they are identical at x or if the two haplotypes differ at x , we let $L_{i,j}(x) = 0$. The definition clearly shows that only haplotype sharing instead of genotype sharing is considered, and the idea underlying this definition is that sharing should be for a stretch of DNA.

The BHS approach of using Mantel statistics on the level of haplotypes rather than individuals might be improved by explicitly taking into account different modes of inheritance, as an additive genetic model is implicitly assumed in the BHS statistic. For example, while under a dominant mode of inheritance, we cannot expect that both haplotypes derived from an affected subject carry the causal variant (Figure 2a), under a recessive genetic model this is supposed to be the case (Figure 2b). To formulate the problem more precisely, we are looking for haplotype fragments that are shared on longer stretches within patients than within controls in case of dominant or recessive modes of inheritance. For this purpose the shared length between all haplotypes derived from individual i and all haplotypes derived from individual j is computed. For example, the shared length between the first haplotype derived from individual i and the second haplotype derived from individual j at the putative disease locus x is denoted by $L_{i,j}^{(1,2)}(x)$. For a dominant mode of inheritance we suggest to use

$$L_{i,j}^{\max}(x) = \max \{L_{i,j}^{(1,1)}(x), L_{i,j}^{(1,2)}(x), L_{i,j}^{(2,1)}(x), L_{i,j}^{(2,2)}(x)\} \quad (3)$$

as the shared length between the individuals i and j . The basis of this definition can be explained with help of the following example. Imagine that two affected individuals i and j with two and one haplotype carrying the disease causing mutation at locus x (Figure 2a). Some haplotype pairs of these individuals do not share the disease causing fragment, i.e.,

$L_{i,j}^{(1,2)}(x) = 0$. By taking the maximum of the shared length of the haplotype pairs, it is ensured that the largest haplotype fragment around the locus x is found that both individuals have in common.

Under a recessive mode of inheritance, both haplotypes of an affected individual have to contain the disease causing mutation. This implies that all haplotype fragments around the locus x of two subjects should be identical (Figure 2b). It is thus reasonable to search for the smallest shared fragment across all individual haplotype combinations, i.e.,

$$L_{i,j}^{\min}(x) = \min \{L_{i,j}^{(1,1)}(x), L_{i,j}^{(1,2)}(x), L_{i,j}^{(2,1)}(x), L_{i,j}^{(2,2)}(x)\}. \quad (4)$$

Figures 2c and 2d consider an affected-unaffected pair of individuals. While $L_{i,j}^{max}(x)$ is expected to be 0 at the disease locus because the unaffected subject should not carry any disease allele, $L_{i,j}^{min}(x)$ is expected to be 0 around the disease locus for a recessive mode of inheritance because the unaffected subject should not have more than one disease carrying haplotype.

For a pair of unaffected subjects, only random haplotype sharing is expected, and both $L_{i,j}^{max}(x)$ and $L_{i,j}^{min}(x)$ are expected to be 0 at the disease locus in panmictic populations. An example for the calculation of $L_{i,j}^{max}(x)$ and $L_{i,j}^{min}(x)$ is given in Figure 3.

2.2 Novel haplotype sharing Mantel test statistics

The genetic similarity measures $L_{i,j}^{max}(x)$ and $L_{i,j}^{min}(x)$ which are adjusted to dominant and recessive modes of inheritances form the foundation of our new model-based HS Mantel test statistics

$$M^{(d)}(x) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n L_{i,j}^{max}(x) Y_{i,j} \quad (5)$$

and

$$M^{(r)}(x) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n L_{i,j}^{min}(x) Y_{i,j}. \quad (6)$$

In this context, the term ‘model-based’ means that the test statistic is adapted to a specific genetic model but not to a distributional assumption for the phenotypes and haplotypes. For simplicity, we drop the position index x in the following. In most applications, the underlying genetic model is unknown, and the use of a model-based test statistic may lead to a substantial loss of power [Ziegler and König 2006, section 6.2.2]. On the basis of our HS Mantel statistics and the BHS Mantel statistic [Beckmann et al., 2005a], we propose a model-free max-test statistic in the sense of Freidlin and colleagues [2002]. To make the three test statistics M , $M^{(d)}$ and $M^{(r)}$ comparable with respect to location and variability, we first estimated the mean and the standard deviation of each HS statistic under the null hypothesis from permutations.

Second, we standardized the test statistics using the estimated mean and standard deviation, yielding M_s , $M_s^{(d)}$ and $M_s^{(r)}$. Specifically, let $E(M)$ denote the empirical expectation and $SD(M)$ the empirical standard deviation of the BHS Mantel statistic, both derived by permutations, the standardized statistic is defined as $(M - E(M))/SD(M)$. Although Mantel [1967] provided exact formulae for the mean and the variance of space-time clustering statistics under permutation distributions, we propose estimation from Monte-Carlo simulations. The max-test statistic selects the maximum of the standardized HS statistics for the additive, dominant and recessive models so that $MAX = \max\{M_s, M_s^{(d)}, M_s^{(r)}\}$.

2.3 Assessment of statistical significance

For n individuals in our sample with haplotypes and phenotypes, the null hypothesis of no association is equivalent to the situation that the individual haplotypes occur independently

with the phenotypes. Because of this and the non-normality of the respective test statistics [Allen and Satten 2007b] we employed a Monte-Carlo permutation approach. First, we calculated the statistics M at marker x based on the observed individual haplotype and phenotype dataset. Second, we generated V replicated datasets by randomly permuting the haplotypes shared length values among individuals and keeping the phenotype values fixed. Third, the permuted test statistics M were calculated for deriving the empirical null distribution of the statistic at marker x . The advantage of this approach was that no assumption about the marginal distribution of the phenotypes and haplotypes had to be made, and this approach was analogous to the one taken by Beckmann and colleagues [2005a]. All tests considered were two-sided.

3. Simulated data

To demonstrate the validity of our novel approaches, we conducted Monte-Carlo simulation studies.

3.1 Case-control assessment

For haplotype simulation, we utilized an Ancestral Recombination Graph (ARG) based software [Hudson 2002]. The haplotypes consist of segregating loci, and each locus represents a diallelic polymorphisms. The software operates with variables such as population size, mutation rate, recombination rate and chromosomal length. Analogously to Beckmann and colleagues [2005a], we assumed random mating in a constant effective population size of 20,000 and simulated a 100,000 base pairs (bp) region. Mutation rate per marker and generation was 5×10^{-9} , and the recombination fraction between two consecutive markers was 10^{-9} . A set of 10,000 haplotypes was simulated, whereas each haplotype consisted of 15 single nucleotide polymorphisms (SNPs). The minor allele frequency was greater than 5%, and the haplotype samples were selected for strong linkage disequilibrium (LD), i.e. $D' > 0.7$ for neighboring SNPs.

The set of 10,000 simulated haplotypes was divided into 5,000 individuals for every single replicate; the individual haplotype pairs were randomly chosen without replacement. We generated the disease status based on the genotype at the putative disease locus depending on the disease models stated in Table I for all individuals. The disease causing marker locus was randomly chosen for each replication. Two disease models were considered according to the genotype penetrances and modes of inheritance (dominant, recessive and additive mode of inheritance). Disease model I reflects a weaker complex disease model with a reduced penetrance for the disease locus and phenocopies, whereas disease model II is closer to a Mendelian disease with high penetrance and phenocopies. The baseline risk allows for locus and / or allelic heterogeneity. Case and control samples were randomly chosen from the data. For investigating power, the sample sizes consisted of 100 to 500 case-control pairs in each replication. Under the null hypothesis, the disease status was randomly chosen with probability 0.5. The sample consisted of 100 cases and 100 controls for each replicate, and we chose $\mu = 0.05$ as population prevalence. The results of this simulation were based on 1,000 independent replicates.

3.2 Impact of haplotype estimation

To study the impact of unphased haplotypes, we analyzed both the type I error and the power using the true simulated haplotypes and the best estimate haplotypes. The true and estimated haplotypes consisted of 15 markers. The best haplotype pairs for unrelated individuals were estimated using fastPHASE [Scheet and Stephens 2006]. For the type I error analysis, the samples consisted of 100 case-control pairs where the disease status was assigned as described before under the null hypothesis of no correlation. For power analysis,

the disease status was assigned as described in Table I under disease model II for a dominant, recessive and additive mode of inheritance. The sample consisted of 300 case-control pairs. For 1,000 replicates, HS Mantel statistic methods were applied to the true simulated haplotypes and the best estimate haplotypes.

3.3 Impact of missing genotypes

In practice, genotypes are not available for all subjects at all genetic markers. Even in this case, fastPHASE [Scheet and Stephens 2006] may be used to estimate the underlying haplotypes. To evaluate the impact of the imputation of missing data and the estimation of haplotypes on our HS test statistics, we generated missing data in our Monte-Carlo simulation studies as follows: for a random 50% of case-control pairs, no missing data was introduced. For the other half of subjects, three markers (located at map positions 2, 7 and 12) were set to missing. Another 0, 1 or 2 out of 12 remaining markers were also set to missing according to a discrete uniform distribution. The position of the additional missing marker(s) was also chosen randomly.

For investigating the validity of the test statistics, the samples consisted in 100 cases and 100 controls with an assignment of the disease status as described before. For power analysis, the disease status was assigned as described in Table I under disease model for a dominant, recessive and additive mode of inheritance for 300 cases and 300 controls. For each scenario, the number of replicates was 1,000.

3.4 Analysis of a candidate region for rheumatoid arthritis on chromosome 18q

As an application of the presented methods, we re-analyzed the RA dataset of the NARAC provided for the Genetic Analysis Workshop 15 [Amos et al., 2007]. The dataset comprises 460 cases and 460 controls which were genotyped at 2,300 SNPs, covering 10 Mega bases (Mb) on chromosome 18q. Controls were recruited from a New York City population, and cases were ascertained from multiple U.S. centers. The phenotype variable to be analyzed was the American Rheumatoid Arthritis (RA) affection status, and the disease prevalence μ is 1% in the population [Begovich et al., 2004]. All affected subjects met the standard American College of Rheumatology criteria for affection with RA [Jawaheer et al., 2004].

3.5 Software

The haplotypes were generated using the program ms [Hudson 2002] (see, <http://home.uchicago.edu/~rhudson1/>), and fastPHASE was employed for estimating missing genotypes and reconstructing haplotypes from unphased SNP of unrelated individuals [Scheet and Stephens 2006]. All other calculations were performed with software developed within our group.

4. Results

4.1 Case-control study analysis

Table II (upper part) gives empirical type I error fractions for our and the BHS Mantel statistics at a nominal significance level of 5%, and for a Bonferroni corrected nominal significance level of $\alpha = 0.05/15 = 0.003$. At $\alpha = 0.05$, all statistics yielded point-wise valid tests (range: 0.025 – 0.054), and the mean of the type I errors of the 15 SNPs ranged between 0.032 and 0.047 for the investigated test statistics. However, the point-wise significance level of 5% led to an unacceptable increase of the relative frequency of replicates with at least one false-positive (multiple tests) over all statistics. Over all replicates, the multiple false-positive fractions varied between 0.142 and 0.267. To adjust for multiple testing, a Bonferroni correction was applied by dividing the point-wise significance level of 5% by the number of tests performed in each replication. After

Bonferroni correction, all statistics led to multiple type I error fractions below 5%. Empirical power was analyzed at the nominal significance level of 0.05. Figure 4 shows that power increased with sample size, and it was greater for the less complex disease model II. Under the dominant and the recessive mode of inheritance, the HS Mantel statistics based on the genetic similarity measures adapted to the corresponding mode of inheritance, i.e., $M^{(d)}$ and $M^{(r)}$, gave the best performance, validating the ideas underlying these statistics. Under a dominant inheritance pattern the empirical power of the HS Mantel statistic for a dominant model did not exceed 65% and 85.2% for 500 case-control pairs in disease model I and II, respectively. Under a recessive inheritance pattern, the empirical power of the HS Mantel statistic for a recessive model was 75.2% and 88.7% for 500 case-control pairs in disease models I and II, respectively. For an additive mode of inheritance, the model-free *MAX* test statistic turned out to be the best (also see Table III). Even under the dominant and recessive genetic mode of inheritance the max-test showed reasonable power compared to the best performing HS Mantel statistic.

If the model is not specified correctly, these tests adopting the mode of inheritance perform poorly, while the BHS statistic M is stable in these cases (Table III). Our novel max-test HS Mantel statistic outperformed the BHS statistic based on an additive genetic model proposed by Beckmann and colleagues [2005a] in all considered scenarios. Overall, we conclude that the max-test is superior to the single model-based test statistics if the mode of inheritance is unknown.

Table IV shows the estimated Pearson product-moment correlation coefficients of the three standardized HS test statistics M , $M^{(d)}$ and $M^{(r)}$ over all replicates for disease model II and 300 case-control pairs per replicate. The correlation is strong between M , and $M^{(d)}$ and between M and $M^{(r)}$ for all models considered. It is weak between $M^{(d)}$ and $M^{(r)}$ for dominant and additive genetic models.

4.2 True haplotypes versus best estimates

In applications haplotypes determined with laboratory methods are rarely available. Therefore, one often has to rely on haplotypes that were inferred from genotypes. Using best estimate haplotypes instead of the true ones we empirically determined the type I error of the different HS Mantel statistics at a nominal significance level of 5% and for a Bonferroni-corrected nominal significance level of $\alpha = 0.05/15 = 0.003$. At a significance of level 5%, all statistics yielded point-wise valid tests (Table II, middle part). The type I error of the single markers varied between 0.023 and 0.064. The means of the type I errors of the 15 SNPs ranged between 0.034 and 0.049 for the investigated test statistics. Applying a Bonferroni correction the empirical multiple type I error was lower than 5% for all investigated test statistics. $M^{(d)}(x)$, $M(x)$ and $MAX(x)$ revealed slightly higher type I errors compared to the case of true haplotypes (Table II).

All statistics showed slight loss of power when best estimate haplotypes were used compared to the results for the case of true haplotypes (Table III). Our model-based and model-free HS Mantel statistics clearly had higher power for both true and estimated haplotypes when compared to the BHS Mantel statistic.

4.3 Missing data

Using the best estimate of incomplete individual haplotypes, we empirically determined the type I error for the HS Mantel statistics at a nominal significance level of 5% and for a Bonferroni-corrected nominal significance level of $\alpha = 0.05/15 = 0.003$. At $\alpha = 0.05$, all statistics yielded point-wise valid tests (Table II, lower part). The type I error of the single markers ranged between 0.025 and 0.064, and the means of the type I errors of the 15 SNPs

ranged from 0.036 to 0.050 for the investigated test statistics. Applying a Bonferroni correction, the empirical multiple type I error was lower than 5% for all investigated test statistics. All HS Mantel statistics revealed slightly higher type I errors when compared with the case of known true haplotypes or haplotypes estimated from the complete data (Table II). As expected, there was a slight loss of power in the case of missing genotypes compared with the case of perfectly known and the best estimated haplotypes (Table III).

4.4 Analysis of the chromosome 18q candidate region for rheumatoid arthritis

The analyses of the RA case-control data were performed in two steps. We first analyzed all SNPs jointly, estimated empirical p-values by 1,000 permutations and defined interesting regions if more than three neighboring SNPs showed empirical p-values < 0.05 (Figure 5). For these regions obtained in the first step of analysis, we increased the number of permutations to 10,000. In this second step, we discarded isolated markers with $p < 0.005$. This led to the identification of three regions containing six known genes according to NCBI built 36.2, namely PMIP1, MC4R, PIGN, KIAA1468, TNFRSF11A and ZCCHC2 (Table V).

5. Discussion

Haplotype-based methods including HS can be very powerful for detecting disease associations. In this paper, we proposed novel model-based HS Mantel statistics. Specifically, we measure genetic similarity by the maximum shared length of haplotypes around a disease locus between individuals for a dominant mode of inheritance and by the minimum shared length of haplotypes around a disease locus between individuals for a recessive mode of inheritance. For an unknown genetic model we suggested to use the maximum of the standardized version of the dominant, recessive HS Mantel statistics and the BHS Mantel statistic as a model-free HS Mantel statistic.

In simulation studies, we showed that the novel HS Mantel statistics based on the genetic similarity measures adapted to the underlying mode of inheritance are more powerful than the BHS statistic of Beckmann and colleagues. Further, if the underlying genetic model is unknown, our novel model-free max-test HS Mantel statistic outperforms BHS.

In addition to the max-test, Freidlin and colleagues [2002] also investigated the performance of a linear combination of the tests for a dominant and a recessive mode of inheritance. We did not include the linear combination approach in our investigation because Freidlin and colleagues [2002] concluded from their simulations that the max-test approach has greater power.

We investigated the impact of unphased haplotypes. To this end, we employed the HS Mantel statistics to samples of haplotypes with known phase and to samples of most likely haplotypes estimated by fastPHASE [Scheet and Stephens 2006]. The results showed a slightly loss of power for all investigated statistics, which coincides with previous findings [Beckmann et al., 2005a; Fischer et al., 2003]. In the simulation situations where some individuals haplotypes were concerned with missing data, we showed that estimating incomplete haplotypes using fastPHASE [Scheet and Stephens 2006] also gives a slightly loss of power for all investigated HS Mantel statistics.

Through the analysis of RA case-control data, we detected three regions with clusters of markers achieving empirical p-values < 0.005 using our and Beckmann HS Mantel statistics, containing six known genes (PMIP1, MC4R, PIGN, KIAA1468, TNFRSF11A and ZCCHC2). The three regions do not overlap with the five regions identified by Huang and colleagues [2007] in their sliding window haplotype analysis. The results from all SNPs are

provided in the Supplementary Table. Region 1 was detected by our recessive HS Mantel statistic only. One possible candidate gene is TNFRSF11A which belongs to the TNF receptor superfamily and encodes for the receptor activator of nuclear factor- κ B (RANK). It is well known that the RANK/RANKL/OPG system plays a central role in bone remodeling. Imbalances in the RANK/RANKL/OPG system could result in several disorders of mineral metabolism [for an overview, see Vega et al., 2007]. For example, elevated serum levels of soluble RANKL and OPG in patients with RA have been found [Vega et al., 2007]. The discovery of the RANK/RANKL/OPG system has led to the discovery of three activating mutations within the TNFRSF11A gene, result in three different rare genetic disorders of mineral metabolism which are namely PDB2, expansile skeletal hyperphosphatasia and familial expansile osteolysis [Hughes et al., 2000; Nakatsuka et al., 2003; Whyte and Hughes 2002]. The results suggest that genes in these three regions may play an important role in the risk for RA disease and they should be subject of further investigation.

The HS Mantel statistics approaches presented here can easily be extended to quantitative traits. However, all existing varieties of HS Mantel statistic methods assume that haplotypes are known or have been inferred. To decrease computational burden, an extension to approaches that avoid the haplotyping step and rely on genotypes only would be welcome. This method could be applied in large scale mapping or in fine mapping, depending on the scope of the study.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by a grant from the Medical Faculty of the University at Lübeck to AK. The work of AE is supported by the Libyan Ministry of Science.

We are grateful to Peter K. Gregersen and the investigators in the North American Rheumatoid Arthritis Consortium (NARAC) for allowing us to use their genotyping data for this study. The original data collection was supported by a grant from the National Institutes of Health, RO1 AR44422 and NO1-AR-2-2263.

AZ initiated and supervised this study and wrote the final manuscript, AE did the programming, the analysis and wrote the draft manuscript, MB suggested to investigate the MAX statistic and critically reviewed the content, and AK evaluated the biological relevance of the findings of the RA data analysis and reviewed the content.

Notation

$M(x)$	HS Mantel statistic of Beckmann and colleagues at marker locus x
$L_{i,j}(x)$	Shared length between haplotypes i and j at marker locus x
$Y_{i,j}$	Measure of the phenotypic similarity between the subjects i and j
Y_i	Observed phenotype of subjects i
μ	Disease prevalence in the population
x	Marker locus
$L_{i,j}^{(1,2)}(x)$	Shared length between the first haplotype derived from individual i and the second haplotype derived from individual j at marker locus x
$L_{i,j}^{max}(x)$	$=\max\{L_{i,j}^{(1,1)}(x), L_{i,j}^{(1,2)}(x), L_{i,j}^{(2,1)}(x), L_{i,j}^{(2,2)}(x)\}$
$L_{i,j}^{min}(x)$	$=\min\{L_{i,j}^{(1,1)}(x), L_{i,j}^{(1,2)}(x), L_{i,j}^{(2,1)}(x), L_{i,j}^{(2,2)}(x)\}$

$M^{(d)}(x)$	HS Mantel statistic for a dominant mode of inheritance at marker locus x
$M^{(r)}(x)$	HS Mantel statistic for a recessive mode of inheritance at marker locus x
M_s	Standardized version of HS Mantel statistic of Beckmann and colleagues
$M_s^{(d)}$	Standardized version of the HS Mantel statistic for a dominant mode of inheritance
$M_s^{(r)}$	Standardized version of the HS Mantel statistic for a recessive mode of inheritance
MAX	$=\max\{M_s, M_s^{(d)}, M_s^{(r)}\}$
D'	Lewontin's measure of linkage disequilibrium
α	Nominal significance level

References

- Allen AS, Satten GA. Inference on haplotype/disease association using parent-affected-child data: the projection conditional on parental haplotypes method. *Genet Epidemiol.* 2007a; 31(3):211–23. [PubMed: 17266114]
- Allen AS, Satten GA. Statistical models for haplotype sharing in case-parent trio data. *Hum Hered.* 2007b; 64(1):35–44. [PubMed: 17483595]
- Amos C, Chen W, Remmers E, Siminovitch K, Seldin M, Criswell L, Lee A, John S, Shephard N, Worthington J. Data for Genetic Analysis Workshop (GAW) 15 Problem 2, genetic causes of rheumatoid arthritis and associated traits. *BMC Proceedings.* 2007; 1(Suppl 1):S3. others. [PubMed: 18466527]
- Beckmann L, Thomas DC, Fischer C, Chang-Claude J. Haplotype sharing analysis using Mantel statistics. *Hum Hered.* 2005a; 59(2):67–78. [PubMed: 15838176]
- Beckmann L, Ziegler A, Duggal P, Bailey-Wilson JE. Haplotypes and haplotype-tagging single-nucleotide polymorphism: Presentation group 8 of Genetic Analysis Workshop 14. *Genet Epidemiol.* 2005b; 29(Suppl 1):S59–71. [PubMed: 16342175]
- Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, Ardlie KG, Huang Q, Smith AM, Spoeke JM. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet.* 2004; 75(2):330–37. others. [PubMed: 15208781]
- Bourgain C, Genin E, Quesneville H, Clerget-Darpoux F. Search for multifactorial disease susceptibility genes in founder populations. *Ann Hum Genet.* 2000; 64:255–65. [PubMed: 11409410]
- Diepstra A, Niens M, Vellenga E, van Imhoff GW, Nolte IM, Schaapveld M, van der Steege G, van den Berg A, Kibbelaar RE, te Meerman GJ. Association with HLA class I in Epstein-Barr-virus-positive and with HLA class III in Epstein-Barr-virus-negative Hodgkin's lymphoma. *Lancet.* 2005; 365(9478):2216–24. others. [PubMed: 15978930]
- Elston RC, Buxbaum S, Jacobs KB, Olson JM. Haseman and Elston revisited. *Genet Epidemiol.* 2000; 19(1):1–17. [PubMed: 10861893]
- Fischer C, Beckmann L, Majoram P, te Meerman G, Chang-Claude J. Haplotype sharing analysis with SNPs in candidate genes: The Genetic Analysis Workshop 12 example. *Genet Epidemiol.* 2003; 24:68–73. [PubMed: 12508257]
- Foerster J, Nolte I, Junge J, Bruinenberg M, Schweiger S, Spaar K, van der Steege G, Ehlert C, Mulder M, Kalscheuer V. Haplotype sharing analysis identifies a retroviral dUTPase as candidate susceptibility gene for psoriasis. *J Invest Dermatol.* 2005; 124(1):99–102. others. [PubMed: 15654959]
- Freidlin B, Zheng G, Li Z, Gastwirth JL. Trend tests for case-control studies of genetic markers: power, sample size and robustness. *Hum Hered.* 2002; 53(3):146–52. [PubMed: 12145550]

- Huang BE, Amos CI, Lin DY. Detecting haplotype effects in genomewide association studies. *Genet Epidemiol.* 2007; 31(8):803–12. [PubMed: 17549762]
- Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics.* 2002; 18(2):337–38. [PubMed: 11847089]
- Hughes AE, Ralston SH, Marken J, Bell C, MacPherson H, Wallace RG, van Hul W, Whyte MP, Nakatsuka K, Hovy L. Mutations in TNFRSF11A, affecting the signal peptide of RANK, cause familial expansile osteolysis. *Nat Genet.* 2000; 24(1):45–8. others. [PubMed: 10615125]
- Jawaheer D, Lum RF, Amos CI, Gregersen PK, Criswell LA. Clustering of disease features within 512 multicase rheumatoid arthritis families. *Arthritis Rheum.* 2004; 50(3):736–41. [PubMed: 15022313]
- Kleensang A, Franke D, König IR, Ziegler A. Haplotype sharing analysis for alcohol dependence based on quantitative traits and the Mantel statistic. *BMC Genet.* 2005; 6(Suppl 1):S75. [PubMed: 16451689]
- Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res.* 1967; 27(2):209–20. [PubMed: 6018555]
- Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol.* 2002; 23(3):221–33. [PubMed: 12384975]
- Nakatsuka K, Nishizawa Y, Ralston SH. Phenotypic characterization of early onset Paget's disease of bone caused by a 27-bp duplication in the TNFRSF11A gene. *J Bone Miner Res.* 2003; 18(8): 1381–5. [PubMed: 12929927]
- Qian D. Haplotype sharing correlation of alcohol dependence on chromosomes 1-6 in 93 nuclear families. *BMC Genet.* 2005; 6(Suppl 1):S79. [PubMed: 16451693]
- Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haploypic phase. *Am J Hum Genet.* 2006; 78(4): 629–44. [PubMed: 16532393]
- te Meerman GJ, van der Meulen MA, Sandkuijl LA. Perspectives of identity by descent (IBD) mapping in founder populations. *Clin Exp Allergy.* 1995; 25(Suppl 2):97–102. [PubMed: 8590355]
- Vega D, Maalouf NM, Sakhaee K. CLINICAL Review #: the role of receptor activator of nuclear factor-kappaB (RANK)/RANK ligand/osteoprotegerin: clinical implications. *J Clin Endocrinol Metab.* 2007; 92(12):4514–21. [PubMed: 17895323]
- Whyte MP, Hughes AE. Expansile skeletal hyperphosphatasia is caused by a 15-base pair tandem duplication in TNFRSF11A encoding RANK and is allelic to familial expansile osteolysis. *J Bone Miner Res.* 2002; 17(1):26–9. [PubMed: 11771666]
- Ziegler, A.; König, IR. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications.* Wiley-VCH; Weinheim: 2006.

Figure 1.

The shared length $L_{ij}(x)$ between the haplotype i and j is given by the number of intervals surrounding the locus x that are flanked by markers with the same allele.

Figure 2.

Genetic similarity between two individuals i and j at a disease locus x expressed as shared length of haplotypes using $L_{i,j}^{max}(x)$ for a dominant mode of inheritance (left) and $L_{i,j}^{min}(x)$ for a recessive mode of inheritance. $L_{i,j}^{max}(x)$ is defined as the maximum shared length of haplotypes around a disease locus between individuals i and j , and $L_{i,j}^{min}(x)$ is the minimum shared length of haplotypes around a disease locus between individuals i and j . The upper part of the figure (a) and (b) represents haplotypes for a pair of affected subjects; a) shows that one copy of the disease allele is sufficient to expect excess haplotype sharing, while haplotype sharing is expected to be observed on both copies of the disease allele in b). An affected-unaffected pair of individuals is considered in parts c) and d). It is seen that $L_{i,j}^{max}(x)$ is expected to be 0 at the disease locus because no disease background should be present on the haplotypes of the unaffected subject. Analogously, $L_{i,j}^{min}(x)$ is expected to be 0 for a recessive mode of inheritance because the unaffected subject should not carry more than one disease haplotype.

Figure 3.

The shared length $L_{i,j}^{min}(x)$ for a pair of individuals *i* and *j* is given by the minimum length over all four haplotypes surrounding the disease locus *x*. The dark grey blocks depict the haplotype with the disease causing variant. For illustrative purposes, the disease is assumed to be autosomal recessive. $L_{i,j}^{min}(x)$ is represented by the grey shaded area and equals 2. $L_{i,j}^{max}(x)$ is represented by the light grey area and equals 4.

Figure 4.

Empirical power of the HS Mantel statistic adapted to a dominant and recessive mode of inheritance $M^{(d)}[\times]$ and $[\diamond]$ as well as the HS Mantel statistic suggested by Beckmann M $[+]$ and the model-free $[\bullet]$ statistic based on 1000 replicates for different disease models. The results under a dominant, recessive and additive mode of inheritance are given, respectively, in figures a and b, c and d, and e and f. All statistics were evaluated two-sided at the nominal 5% test level.

Figure 5.

Results of the haplotype sharing Mantel test statistic methods for all 2,300 SNPs applied to chromosome 18q after the initial analysis (1,000 permutations). The horizontal line indicates 5% significance level. Shaded areas indicate regions identified in the second part of the analysis, also see Table V.

Table 1

Penetrances for different disease models used in the analysis.

Genotype ^a	Disease model I			Disease model II		
	Dominant	Recessive	Additive	Dominant	Recessive	Additive
(1, 2)	0.170	0.170	0.017	0.170	0.170	0.017
(1, 2)	0.580	0.170	0.169	0.902	0.170	0.424
(2, 2)	0.580	0.580	0.338	0.902	0.902	0.848

^a 1 = Normal allele; 2 = Disease allele.

Empirical type I error of the HS Mantel statistics at nominal significance levels of 5% and 0.3% based on 1,000 replicates for true haplotypes (true), best estimate haplotypes (complete estimated), and best estimate haplotypes from incomplete genotype data (incomplete estimated). The samples consisted of 100 cases and 100 controls.

Table II

Haplotypes	Method	$\alpha = 0.05$			$\alpha = 0.003^c$		
		Pointwise ^a	Multiple ^b	Multiple ^b	Pointwise ^a	Multiple ^b	Multiple ^b
True	$M^{(d)}(x)$	0.039 (0.032-0.050)	0.142	0.002 (0.001-0.004)	0.013		
	$M^{(c)}(x)$	0.047 (0.041-0.054)	0.267	0.002 (0.000-0.004)	0.014		
	$M(x)$	0.040 (0.032-0.045)	0.198	0.002 (0.000-0.006)	0.017		
	MAX	0.032 (0.025-0.038)	0.162	0.002 (0.000-0.005)	0.017		
Complete estimated	$M^{(d)}(x)$	0.048 (0.041-0.064)	0.178	0.002 (0.001-0.006)	0.010		
	$M^{(c)}(x)$	0.045 (0.035-0.054)	0.263	0.002 (0.000-0.007)	0.020		
	$M(x)$	0.049 (0.037-0.063)	0.236	0.002 (0.000-0.006)	0.020		
	MAX	0.034 (0.023-0.047)	0.186	0.002 (0.000-0.006)	0.015		
Incomplete estimated	$M^{(d)}(x)$	0.049 (0.040-0.064)	0.196	0.001 (0.000-0.006)	0.015		
	$M^{(c)}(x)$	0.047 (0.036-0.057)	0.278	0.002 (0.001-0.007)	0.020		
	$M(x)$	0.050 (0.040-0.062)	0.233	0.002 (0.000-0.007)	0.021		
	MAX	0.036 (0.025-0.048)	0.231	0.002 (0.001-0.005)	0.017		

^aMean type I error for 15 markers, minimum and maximum are given in parentheses.

^bRelative frequency of replicates with at least one false-positive.

^cBonferroni correction for multiple testing: $p < 0.05/15 = 0.003$.

Empirical power of the HS Mantel type test statistics at a nominal significance level of 5% based on 1,000 replicates for true haplotypes (true), best estimate haplotypes (complete estimated), and best estimate haplotypes from incomplete genotypes (incomplete estimated) under disease model II. The samples consisted of 300 cases and 300 controls.

Table III

Genetic model	Haplotypes	$M^{(d)}(x)$	$M^{(r)}(x)$	$M(x)$	MAX
Dominant	True	0.78	0.17	0.60	0.75
	Complete estimated	0.75	0.15	0.58	0.72
Recessive	Incomplete estimated	0.73	0.21	0.53	0.69
	True	0.65	0.82	0.77	0.79
	Complete estimated	0.64	0.81	0.76	0.78
	Incomplete estimated	0.51	0.68	0.64	0.65
Additive	True	0.95	0.73	0.82	0.98
	Complete estimated	0.94	0.72	0.81	0.97
	Incomplete estimated	0.90	0.64	0.78	0.92

Table IV

Pearson product-moment correlation coefficients over all replicates for true haplotypes of the standardized HS Mantel type test statistics and disease model II. Each replicate consisted of 300 cases and 300 controls.

Genetic model		$M^{(d)}(x)$	$M^{(r)}(x)$
Dominant	$M^{(d)}(x)$	1.00	
	$M^{(r)}(x)$	0.01	1.00
	$M(x)$	0.85	0.52
Recessive	$M^{(d)}(x)$	1.00	
	$M^{(r)}(x)$	0.69	1.00
	$M(x)$	0.91	0.91
Additive	$M^{(d)}(x)$	1.00	
	$M^{(r)}(x)$	0.09	1.00
	$M(x)$	0.70	0.76

Table V

Selected regions on chromosome 18q after the second step of analysis (10,000 permutations).

Regions ^a	SNPs	Locations (bp)	Known genes ^b	Test statistics
1	rs1975145-rs4940711	55,716,153-55,920,566	PMIP1	$M^{(r)}(x)$
2	rs4940736-rs3760559	56,130,405-56,727,645	MC4R	$M^{(d)}(x)$, $M^{(r)}(x)$, $M(x)$, MAX
3	rs1943232-rs1497965	57,879,383-58,374,101	PIGN, KIAA1468, TNFRSF11A, ZCCHC2	$M^{(d)}(x)$, $M(x)$, MAX

^aRegions which includes SNPs with p-values < 0.005.

^bKnown genes from NCBI Build 36.2.