The Use of Haplotypes in the Identification of Interaction between SNPs

Gie Ken-Dror    Steve E. Humphries    Fotios Drenos
Centre for Cardiovascular Genetics, BHF Laboratories, Department of Medicine, Royal Free and University College Medical School, London, UK

Key Words
Haplotype · Interaction · Single nucleotide polymorphisms

Abstract
Although haplotypes can provide great insight into the complex relationships between functional polymorphisms at a locus, their use in modern association studies has been limited. This is due to our inability to directly observe haplotypes in studies of unrelated individuals, but also to the extra complexity involved in their analysis and the difficulty in identifying which is the truly informative haplotype. Using a series of simulations, we tested a number of different models of a haplotype carrying two functional single nucleotide polymorphisms (SNPs) to assess the ability of haplotypic analysis to identify functional interactions between SNPs at the same locus. We found that, when phase is known, analysis of the haplotype is more powerful than analysis of the individual SNPs. The difference between the two approaches becomes less either as an increasing number of non-informative SNPs are included, or when the haplotypic phase is unknown, while in both cases the SNP association becomes progressively better at identifying the association. Our results suggest that when novel genotyping and bioinformatics methods are available to reconstruct haplotypic phase, this will permit the emergence of a new wave of haplotypic analysis able to consider interactions between SNPs with increased statistical power.

Introduction
Genome-wide association studies (GWAS) have had great success in identifying single nucleotide polymorphisms (SNPs) associated with a number of common complex diseases [1, 2]. Although most of these newly identified SNPs are non-functional (i.e. they do not alter the function or expression of a protein), they can be used as surrogate markers for the unobserved functional variant due to allelic association (linkage disequilibrium, LD) between adjacent SNPs. The extent of LD across a genomic region dictates the density of SNP markers necessary to ensure the capture of the association between a marker and the causative allele sought. The average distance that LD extends away from a specific SNP varies from 5 to 60 kb, with an upper range able to extend up to hundreds of kb [3–6], usually being larger in populations of European descent, but smaller (e.g. only a few kb) in individuals from African populations [4–7]. Commonly, we see a pattern of high LD across long stretches of DNA punctuated by recombination hot-spots [3, 5, 7]. The traditional GWAS analysis tests the association between the phenotype and SNPs at a locus one at a time [3, 8–10], thus discounting information from their joint distribution. If any interactions between SNPs exist, they need to have a strong marginal effect to be detected through the univariate analysis, something that we cannot a priori assume. This led many to believe that it is much more in-
formative to analyse multiple markers in a region of interest simultaneously. This is especially true when multiple polymorphisms are involved in the phenotype, and more so when there is interaction between the functional SNPs [11, 12].

A biologically sound way to combine the information of multiple SNPs at a locus is with the use of haplotypes, groups of alleles from the same gamete, usually but not necessarily statistically associated with each other [13, 14]. Haplotypes reflect the chromosomal organisation of alleles and hence their pattern of inheritance over evolution [13]. Because of shared inheritance, the total number of actual haplotypes formed by n diallelic SNPs is generally much less than the 2^n that are possible [15, 16]. For a typical gene, the number of common haplotypes, even with variable numbers of SNPs, is of the order of 10–15 [16]. It has been suggested that the analysis of haplotypes can be more powerful, informative and effective than that based on single SNPs [13, 17, 18]. Haplotypes, in contrast to single SNP associations, take local LD structure into account and decrease both the dimensions and number of tests required, which under some circumstances might increase the power of the statistical tests used and help with the problem of multiple testing [15, 19]. Moreover, haplotypes can potentially capture cis-interactions between two or more causal SNPs and define functional units of multiple variants; for example, the Apolipoprotein E protein has three common alleles variants whose different functions strongly influence plasma lipid levels, these are caused by the presence of two coding SNPs which form three common haplotypes [13, 19–21].

An important problem with the use of haplotypes in association tests is that they are not usually observed directly, and precise inference of them from observed genotypes requires genetic information from both parents. Most large-scale association studies include unrelated individuals, making the identification of the true haplotypes difficult. A number of statistical approaches for the reconstruction of the haplotypes are available, either based on the EM-algorithm [22] or alternative methods [23], but their main drawback is the remaining uncertainty of the imputed phase.

The statistical power of haplotypes acting as the functional unit is under-studied. The aim of the present study is to evaluate how the statistical power to identify a functional haplotype is influenced by allele frequency, LD between the SNPs and the effects of non-informative SNPs when considered as part of the functional haplotype. We also examine the efficiency of the classical statistical interaction method to detect haplotypic effects between SNPs.

Methods

Genotype and Haplotype Data

Using STATA 11, we developed a simulation able to create random diallelic SNPs (major allele, A; minor allele, B) for 3,000 individuals with varying minor allele frequencies (MAF) ranging from 5 to 50%. For the initial phase of our simulation, no LD between the SNPs was modelled. In this case, the haplotype was considered as a random selection from the SNPs for each individual. We subsequently added pair-wise LD, simulating the r^2 measure, with values ranging from a low (r^2 = 0.1) to a high (r^2 = 0.9) correlation between the SNPs. To better represent association studies of unrelated individuals, we also assumed that the haplotypic phase was unknown, and the PHASE algorithm [23, 24] was used to infer the haplotypes. The results obtained after the use of PHASE were compared to results based on the known phase of the haplotypes.

Phenotype Data

We assumed that the functional unit is not the set of SNPs but their combination in the haplotype, i.e. the SNPs show interaction. If we consider two SNPs, these can give rise to four possible haplotypes: A_1 A_2, A_1 B_2, B_1 A_2, and B_1 B_2, with A and B being the two alleles for SNP 1 and 2, respectively. There are a number of ways in which the effect of a haplotype can be modelled. The additive model assumes that each change from a major to a minor allele increases the phenotype by one unit, resulting in A_1 A_2 = 0, A_1 B_2 = 1, B_1 A_2 = 1, and B_1 B_2 = 2, which is similar to having the two SNPs working independently. The cis-recessive scheme considers that only the haplotype with the minor allele in both SNPs has an effect such that A_1 A_2 = 0, A_1 B_2 = 0, B_1 A_2 = 0, and B_1 B_2 = 1; while the cis-dominant scheme is the opposite with A_1 A_2 = 0, A_1 B_2 = 1, B_1 A_2 = 1, and B_1 B_2 = 1, where any haplotype with a minor allele influences the phenotype. The trans model assumes that only when the major allele of one SNP is combined with the minor allele of the other is an effect evident, and is coded as A_1 A_2 = 0, A_1 B_2 = 0, B_1 A_2 = 1, and B_1 B_2 = 1, where any haplotype with a minor allele influences the phenotype. We simulated a continuous dependent variable Y using the linear regression model Y = βX + ε, where β represents the vector of coefficient effects of haplotypes on the average trait value, X is the matrix of the haplotypes and ε is the error term with a normal distribution of variance σ and a mean of 0. The variance of the error term was set so that the test achieved a statistical power of 90% with a nominal type I error rate of 5% in our reference category for each group of comparisons.

Statistical Analysis

The test of association between each genotype and the dependent variable was done using a general linear model in STATA with no dominance between the two haplotypes. The test of association for the haplotypes used indicator variables for each haplotype category assuming additive effects between the two haplotypes in the same individual throughout.

Results

The statistical properties of a single SNP test are well known. Generally, the statistical power of an association test between a SNP and a trait increases with increasing
MAF. This is true for almost all of the different genetic models commonly used, i.e. additive, recessive for the minor allele, and the hypothesis-free model (fig. 1; 5,000 replications). The exception is the dominant model which reaches a plateau around a MAF of 25%. As expected, both the dominant and recessive models follow very different trajectories to the same point when the MAF is 50% and the distinction between minor and major allele is lost. It is also evident that when the additive and the hypothesis-free model are applied on the same data, the extra degree of freedom required in the general model causes a pronounced drop of statistical power. This difference between the models is smaller when the MAF is <10% and remains almost constant for higher frequencies.

In the simplest scenario when no LD is modelled between the two SNPs making up the functional haplotype, with the first SNP kept at a MAF of 30% throughout, the statistical power to identify an association between the haplotype and the trait increases with increasing MAF of the second SNP for the additive, complex, cis-recessive and trans models. As in the single SNP case, the cis-dominant model does not follow this pattern, with the power of the test increasing initially until a MAF of 25–30% and then decreasing as the MAF reaches 50%. In all models and with all frequencies tested, the statistical power to identify the association was higher when the haplotype was analysed compared to either of the SNPs. The first SNP with a MAF of 30% in all models remains at the same levels of statistical power for the additive and complex models irrespective of the MAF of the second SNP. In the cis-dominant and trans models, an increase of the MAF of the second SNP, and thus changes in the frequencies of the different haplotypic classes, causes a decrease of power for the univariate test of the first SNP, while in the cis-recessive model there is a slight increase. The power to detect the effect of the second SNP increases with increasing MAF for all models except the cis-recessive, where it remains at the significance level of 5%. The tests for interaction between the two SNPs did not provide statistically significant evidence for interaction in the additive and complex models, while very weak power was seen for the two cis models. The best evidence for interaction between the SNPs was observed in the trans model, particularly for higher SNP frequencies. All results are summarised in figure 2 and online supplementary figure S1 (for all online suppl. material, see www.karger.com/doi/10.1159/000350964).

In most cases, we cannot directly identify the specific SNPs that comprise the functional haplotype. To explore the effects of this, we included up to seven non-functional additional SNPs in our previous example of the two SNP haplotype under the complex model. The extension of the haplotype with an increasing number of non-functional SNPs showed a marked decrease of statistical power with each additional SNP. Even the inclusion of a single non-functional polymorphism in the haplotype caused a decrease of statistical power, such that the association of the dependent variable with the haplotype had lower power compared to the test of the individual SNP. When five additional SNPs were considered, the univariate analysis performed better for both of the functional SNPs compared to the haplotype. Interestingly, including SNPs of low frequency (approx. 5%) caused a smaller drop of statistical power compared to the addition of more common SNPs (>10%; see fig. 3).

The models presented so far have not included LD between the SNPs tested, but this is rather unrealistic when SNPs close together are considered. Using the example of the two-SNP haplotypes and keeping the first SNP at a MAF of 30%, we permitted the second SNP to vary its LD with the first SNP ranging from $r^2 = 0.1$ to $r^2 = 0.9$. In this case, increasing LD between the two SNPs also increased the statistical power for the association between the haplotype and the trait in the additive, cis-recessive and complex models, while it decreased power in the cis-dominant and trans models. The observed increase can be attributed to the decrease of haplotypic classes in the model and thus a decrease in the degrees of freedom considered in each case. The steep decrease of statistical power seen in the trans model can be explained by the decreasing frequency of the trans haplotype as the LD between the two SNPs is increasing. In terms of individual SNPs, the results follow the patterns seen for the independent SNP example, ex-

![Fig. 1. The statistical power of single SNPs among different MAF and genetic models.](image-url)
cept that now the power estimates for the two SNPs are converging as the LD between them increases. Interaction between the SNPs is again very weak except for the trans model for which the statistical power is decreasing with increasing LD due to the lower frequency of individuals having a trans configuration of the two SNPs as the LD increases (fig. 2, online suppl. fig. S2).

Considering again the example when additional non-functional SNPs are included in the haplotype but modeling the LD between them, the results are similar to those seen in the simpler example earlier. As expected, each additional non-functional SNP decreases the statistical power of the haplotype-trait association. When a small number of SNPs is added, the effect of increasing LD on the overall change of power is negligible. By contrast, when the number of non-functional SNPs is more than the number of functional SNPs, increasing LD between the SNPs leads to an increase of statistical power, probably due to the decrease of haplotypic classes (fig. 3).

Unless family data are available, it is unlikely that the true haplotypes can be determined unambiguously. The common procedure for unrelated individuals is to infer the probable haplotypic phases using a statistical approach. Here we chose to use PHASE to reconstruct the
haplotypes and test the effect of the extra uncertainty in the statistical tests of association. Using our previous example of a two-SNP functional haplotype, with LD between them varying between $r^2 = 0.1$ and $r^2 = 0.9$, it is obvious that as the LD between the two SNPs decreases, the ability of the algorithm to reconstruct the correct haplotypes also decreases. Although this is evident in all the models considered, the effect was more pronounced in the trans and the two cis models. The power of the statistical test for the individual SNPs and the interaction between them was, as expected, not affected by the inference of the haplotypes (fig. 4, online suppl. fig. S3). Also as expected, when additional non-functional SNPs are included in the inferred haplotype, the overall power of the haplotypic association decreases further, to the point that the univariate analysis is the best approach to identify the effect on the trait of the polymorphisms if more than the two functional SNPs are tested (fig. 4).

**Discussion**

Using simulated data, we tested the statistical performance of haplotype analysis compared to that of univariate SNP analysis plus a test of statistical interaction. Different models of haplotypic effect were tested including simple models, similar to an additive SNP effect, as well as cis and trans interactions between the SNPs. Progressively more realistic conditions were simulated to better represent common practice in the analysis of haplotypes in unrelated individuals. Our simulations suggest that the statistical power to identify a complex association between a pair of SNPs and a trait is higher for haplotypes than the univariate analysis or the interaction term. The inclusion of non-informative SNPs and the uncertainty over the true haplotypic phase are the main limitations for the efficient use of haplotypes in current, large-scale, genetic epidemiology data.

The Example of APOE

The APOE gene, situated on human chromosome 19, has three major isoforms, characterised by two amino acid substitutions at residues 112 and 158 [25]. Two SNPs, rs429358 (encoding amino acid 112) and rs7412 (encoding amino acid 158), define the APOE status. The LD between the two SNPs is low when $r^2$ is considered, but higher when the D’ measure is taken into account [26].

We used six SNPs (rs405509, rs429358, rs7412, rs439401, rs5167, and rs10413089) in the area of APOE and assessed their association with LDL-C in the Whitehall II study [27–29]. As with the simulation above, we tested each SNP in a univariate analysis as well as the interaction terms and haplotypes of all pairwise SNP combinations, and a summary of the results is presented in table 1. Three of the six SNPs tested were significantly associated with LDL-C in the univariate analysis. There was no evidence for the presence of interaction between the SNPs when the interaction term was fitted in the model. Most of the haplotypes tested were significant, as expected in an area of strong association, but the pair with the lowest p values and highest variance explained was the pair of the two known functional SNPs.

**Fig. 4.** The power of two-SNP functional haplotypes with LD and inferred haplotypes (a) and with non-informative SNPs (b).
Table 1. Association between six SNPs (rs405509, rs429358, rs7412, rs439401, rs5167, and rs10413089) in the area of APOE and LDL-C in the Whitehall II study

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>p value</th>
<th>r²</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_1_2</td>
<td>1.4E-16</td>
<td>0.71</td>
<td>105040</td>
<td>105069</td>
</tr>
<tr>
<td>H_1_3</td>
<td>5.1E-54</td>
<td>2.37</td>
<td>104866</td>
<td>104895</td>
</tr>
<tr>
<td>H_1_4</td>
<td>2.0E-09</td>
<td>0.39</td>
<td>105073</td>
<td>105102</td>
</tr>
<tr>
<td>H_1_5</td>
<td>0.010</td>
<td>0.08</td>
<td>105105</td>
<td>105134</td>
</tr>
<tr>
<td>H_1_6</td>
<td>2.2E-04</td>
<td>0.16</td>
<td>105097</td>
<td>105126</td>
</tr>
<tr>
<td>H_2_3</td>
<td>2.4E-65</td>
<td>2.82</td>
<td>104817</td>
<td>104839</td>
</tr>
<tr>
<td>H_2_4</td>
<td>1.2E-16</td>
<td>0.72</td>
<td>105039</td>
<td>105068</td>
</tr>
<tr>
<td>H_2_5</td>
<td>5.3E-19</td>
<td>0.82</td>
<td>105028</td>
<td>105057</td>
</tr>
<tr>
<td>H_2_6</td>
<td>2.4E-18</td>
<td>0.79</td>
<td>105032</td>
<td>105061</td>
</tr>
<tr>
<td>H_3_3</td>
<td>3.8E-58</td>
<td>2.55</td>
<td>104847</td>
<td>104876</td>
</tr>
<tr>
<td>H_3_4</td>
<td>1.1E-53</td>
<td>2.35</td>
<td>104868</td>
<td>104897</td>
</tr>
<tr>
<td>H_3_5</td>
<td>1.3E-53</td>
<td>2.35</td>
<td>104868</td>
<td>104897</td>
</tr>
<tr>
<td>H_4_4</td>
<td>0.046</td>
<td>0.05</td>
<td>105109</td>
<td>105138</td>
</tr>
<tr>
<td>H_4_6</td>
<td>0.195</td>
<td>0.02</td>
<td>105112</td>
<td>105141</td>
</tr>
<tr>
<td>H_5_6</td>
<td>0.048</td>
<td>0.05</td>
<td>105109</td>
<td>105138</td>
</tr>
</tbody>
</table>

AIC = Akaike’s information criterion; BIC = Bayesian information criterion.

Although the direct determination of haplotypes currently is expensive and labour-intensive in unrelated individuals, recent advances in sequencing and bioinformatics suggest that reconstruction of a single individual haplotyping is feasible [30, 31]. In this case, overlapping DNA sequence fragments, originating from one of the two chromosomes, are phased with the help of heterozygous SNPs. A comparison between the haplotypes obtained from sequencing and trio-based haplotyping showed that current algorithms work well, phasing more than 90% of the SNPs with less than 2% switch error rates [32]. The resulting haplotypic fragments are of limited length and the observation of entire chromosome haplotypes is still unattainable through these tools, although experimental methods based on the microfluidic separation of each chromosome have succeeded in producing chromosome-long haplotypes [33]. Despite these advances, the observation of haplotypes is still more expensive than of the unphased whole-genome data, something that is expected to change with new developments in sequencing technologies [34].

Even if the true haplotypes are known, since only a subset of SNPs contains information for the haplotypic effect, inclusion of non-informative SNPs will effectively divide the sample into multiple haplotypic groups, consequently decreasing the power of the study, while increasing the degrees of freedom of the test as seen in our simulation results. Identifying the most parsimonious haplotype responsible for the association with the phenotype can be done in two different levels. Variable selection techniques can be used to decrease the number of SNPs we consider to a set of markers that can independently contribute to the association [24, 26]. Although these methods are unlikely to capture complex haplotypic effects where many SNPs are involved, each non-informative SNP excluded will halve the number of possible haplotypes to be considered. Alternatively, haplotypic trees, as commonly used in phylogeny, can identify the changes in the haplotype that are most likely to be associated with a change in the phenotype/genotype association [24, 35]. The main issue here is that the complexity of the tree increases very fast with an increasing number of SNPs. A combination of the two levels of selection is also possible and easy to implement [24].

Our study has a number of limitations. To avoid unnecessary complexity, we limited our simulation to two-SNP functional haplotypes and used the APOE gene as an example of such a case. More functional SNPs in larger haplotypes with more complex interactions between them are also possible. Multi-marker tests have not been considered, as this would overlap with the work by Roeder et al. [36] and Rakovski et al. [37]. Generally, as was also apparent in the use of phased haplotypes, these methods become less useful as the number of SNPs considered increases [38].
In summary, haplotypes can be more informative than individual SNPs or interaction terms when interaction is present. The inclusion of non-informative SNPs and the uncertainty of phase in unrelated individuals are the two main problems hindering their use. Technology development will soon permit the direct observation of haplotypes, possibly through entire chromosomes and genome. Statistical methods of SNP selection can help to focus on only the most relevant SNPs, but the problem will persist as the number of SNPs and the length of the haplotypes considered increases. Novel statistical methods will probably be required to reach a point where analyses of a phased diploid genome can be achieved.

Acknowledgements

The British Heart Foundation supports F.D. (PG2005/008). S.E.H. is a British Heart Foundation Chair holder.

References

The Use of Haplotypes in the Identification of Interaction between SNPs


