WHAP: haplotype-based association analysis

Shaun Purcell a,b, Mark J Daly a,b, Pak C Sham c

aCenter for Human Genetic Research, MGH, Boston, USA, bBroad Institute of Harvard and MIT, Cambridge, USA, cGenome Research Center, University of Hong Kong, Pokfulam, Hong Kong.

ABSTRACT

Summary: We describe a software tool to perform haplotype-based association analysis, for quantitative and qualitative traits, in population and family samples, using single nucleotide polymorphism or multiallelic marker data. A range of tests is offered: omnibus and haplotype-specific tests; prospective and retrospective likelihoods; covariates and moderators; sliding window analyses; permutation p-values. We focus on the ability to flexibly impose constraints on haplotype effects, which allows for a range of conditional haplotype-based likelihood ratio tests: for example, whether an allele has an effect independent of its haplotypic background, or whether a single variant can explain the overall association at a locus. We illustrate using these tests to dissect a multi-locus association.

Availability: WHAP is a C/C++ program, freely available from the author’s website: http://pngu.mgh.harvard.edu/purcell/whap/

Contact: Shaun Purcell, shaun@pngu.mgh.harvard.edu

1 INTRODUCTION

WHAP performs haplotype-based association analysis, using a method similar to other recent methods (Schaid et al, 2002; Zaykin et al, 2002; Dudbridge, 2003). We use a weighted maximum likelihood model to account for the potential ambiguity in individuals’ statistically-inferred haplotypes. For H haplotypes, two types of basic test are available: haplotype-specific tests, i.e. H separate 1 degree of freedom tests of each specific haplotype compared with all others; and a single omnibus H - 1 degree of freedom test, jointly testing all haplotypes. In this Application Note, we introduce WHAP and describe a class of conditional tests that allow multi-locus associations to be dissected by asking “does X have an effect independent of Z?”. Here, X and Z can be single, or multiple, markers or haplotypes.

2 METHODS

WHAP implements an extension of Sham et al (2004), which performed well in an independent and comprehensive comparison of haplotype-based case/control association methods (Cordell, 2006). Briefly, an observed, unphased multi-locus genotype is denoted G; a phased haplotype pair H; a phenotype Y; paternal, maternal and offspring indicators are P,M and O. For individuals without genotyped parents, a modified E-M algorithm (Clayton, 2003) is used to obtain the set of possible phases and their probabilities P(Hh|G). When present, both parents are phased separately; assuming independence between paternal and maternal chromosomes and applying basic Mendelian laws, we calculate the probability of each offspring phase consistent with the observed offspring genotypes and parental phase, giving phase probabilities P(Hh|G,p,Hp,SM). Parental genotypes also allow partitioning of genetic effects in separate between-family

(based on the expected offspring haplotype counts for each possible parental phase) and within-family (the difference between the observed offspring haplotype counts and the between component) components (Abecasis et al, 2000). By default, between and within components are equated, but other models can be specified. The contribution to the likelihood from each individual is \( P(Y|G) = \sum_h P(Y|H_h)P(H_h|G) \) where \( P(Y|H_h) \) terms is parameterized following ordinary linear or logistic regression models and maximum likelihood is used to estimate the regression coefficients as described in Sham et al (2004).

Finally, a retrospective likelihood using \( P(G|Y) \) in place of \( P(Y|G) \) can be used, which can have some advantages in selected samples (Sham et al, 2000) including the case of affected-only offspring designs. Omitting the individual subscript, the conditional likelihood is in the form

\[
P(G|Y) = \sum_{g_1} P(Y|g_1)P(G_{g_1}) \sum_{g_0} P(Y|g_0)P(G_{g_0})
\]

where the numerator sum \( \{g_1\} \) represents all phases consistent with the individual’s or trio’s observed genotypes, whereas the denominator sum \( \{g_0\} \) is over all possible phases, regardless of observed genotypes.

Conditional tests

The framework described above allows for a range of basic haplotype tests. WHAP also allows the user to flexibly equate haplotype effects, to group haplotypes and perform cladiastic tests, or to test null hypotheses of homogeneous effects rather than no effects. As shown below, this also facilitates conditional testing: once an association signal has been detected, conditional tests can be useful in determining which variant, or variants, is most likely to be causal, as opposed to showing only indirect association due to linkage disequilibrium with the causal variant.

Standard association analysis is concerned with detection, i.e. whether X is associated with the phenotype, where X is an allele, genotype, haplotype or set of haplotypes. In contrast, conditional tests are concerned with dissection, or how to best characterize multiple-marker associations, by asking whether X is associated with the phenotype independent of something else, i.e. whether X is consistent with the variant being causal (i.e. nothing else has an independent effect) although this does not of course prove this (as the effect might represent indirect association to local ungenotyped variation).

To illustrate the conditional tests, we simulated a dataset (also available from the author’s website) of 200 cases and 200 controls containing 5 SNPs and six common haplotypes. Table 1 shows the estimated frequencies and association results for both conditional and unconditional SNP and haplotype tests. Haplotype H5 (HACTA, not uniquely tagged by any single SNP) was set to be the single risk-increasing variant for disease. Importantly, the significant omnibus result (\( \chi^2 = 19.079, p = 0.00186 \)) suggests that it is appropriate to proceed to conditional testing.
Tests can be represented by the models applied under the alternate and null hypotheses. Each model is described by the coefficient for any six haplotypes in order (1 is the reference category; equality constraints only apply when they are coded to represent allelic dosage and performing a 1 df test as a sixth SNP (i.e. a single SNP unique to SNP-based sole-variant tests (result not shown in Table 1), generally characterized by SNP passes the SNP-based sole-variant test (result). It is also possible to test for independent hypotheses. Each model is described by the coefficients of the six haplotype-specific results: as one would expect, haplotype-specific associations:

**Table 1. SNP and haplotype main effect and conditional effect tests**

<table>
<thead>
<tr>
<th>H</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>P(H)</th>
<th>ORLM</th>
<th>ORHS</th>
<th>HS</th>
<th>SVH</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>A</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>0.264</td>
<td>ref</td>
<td>0.62</td>
<td>0.00346</td>
<td>0.0323</td>
</tr>
<tr>
<td>H2</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>0.237</td>
<td>1.501</td>
<td>1.11</td>
<td>0.513</td>
<td>0.00092</td>
</tr>
<tr>
<td>H3</td>
<td>C</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>0.212</td>
<td>1.309</td>
<td>0.91</td>
<td>0.607</td>
<td>0.000855</td>
</tr>
<tr>
<td>H4</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>0.169</td>
<td>1.467</td>
<td>1.23</td>
<td>0.537</td>
<td>0.000901</td>
</tr>
<tr>
<td>H5</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>A</td>
<td>0.067</td>
<td>3.811</td>
<td>3.08</td>
<td>0.00019</td>
<td>0.272</td>
</tr>
<tr>
<td>H6</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>C</td>
<td>0.050</td>
<td>1.528</td>
<td>1.09</td>
<td>0.787</td>
<td>0.000784</td>
</tr>
</tbody>
</table>

H is haplotype h; S is SNP α; P(H) is haplotype frequency; MAF is SNP minor allele frequency; SS is single SNP allelic p-value; IE is independent effect test p-value; SVH is SNP-based sole-variant test p-value; ORLM is haplotypic odds ratio under the omnibus test; ORHS is haplotypic odds ratio under haplotype-specific c tests; S is haplotype-specific c test p-value; SVH is haplotype-based sole-variant test. Omnibus $\chi^2 = 19.079, p = 0.00186$.

**3 CONCLUSION**

WHAP offers a range of methods for haplotype-based association analysis, with a focus on conditional tests to dissect genetic effects. Although such tests will often be under-powered and inconclusive, other times they will help resolve strong multi-locus association signals as in the example above. Other features of WHAP such as dominant and recessive genetic models, multi-allelic markers and covariates (having main and interacting effects) are described in the on-line documentation. Finally, it should be noted that WHAP is designed for candidate gene studies, or studies of small to moderately-sized chromosomal regions: different tools will be needed for whole genome association studies.

**ACKNOWLEDGMENT**

SP acknowledges UK MRC grant G9901258 and National Eye Institute grant EY-12562.

**REFERENCES**


