

Proceedings

Open Access

## Robust testing of haplotype/disease association

Andrew S Allen\*<sup>1,2</sup> and Glen A Satten<sup>3</sup>

Address: <sup>1</sup>Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA, <sup>2</sup>Duke Clinical Research Institute, Duke University, Durham, NC, USA and <sup>3</sup>Centers for Disease Control and Prevention, Atlanta, GA, USA

Email: Andrew S Allen\* - andrew.s.allen@duke.edu; Glen A Satten - gas0@cdc.gov

\* Corresponding author

from Genetic Analysis Workshop 14: Microsatellite and single-nucleotide polymorphism Noordwijkerhout, The Netherlands, 7-10 September 2004

Published: 30 December 2005

BMC Genetics 2005, 6(Suppl 1):S69 doi:10.1186/1471-2156-6-S1-S69

### Abstract

Haplotypes, the combination of closely linked alleles that fall on the same chromosome, show great promise for studying the genetic components of complex diseases. However, when only multilocus genotype data are available, statistical approaches need to be employed to resolve haplotype phase ambiguity. Recently, we have proposed an approach to testing and estimating haplotype/disease association that is invariant to any existing genetic structure in the population. Here we evaluate this approach by applying it to the Genetic Analysis Workshop 14 simulated data.

### Background

Single-nucleotide polymorphism (SNP)-based haplotype association studies show great potential for dissecting genetic influences on complex diseases. Haplotype-based methods incorporate linkage disequilibrium information from many markers and, hence, may be more powerful than traditional linkage disequilibrium methods that focus on a single SNP. Haplotype-based methods also have the promise of being able to identify unique segments of DNA containing sequences predisposing individuals to disease. In addition, when multiple alleles at a single disease locus influence disease susceptibility, single marker tests can be under-powered relative to haplotype-based association methods [1]. In the Genetic Analysis Workshop 14 (GAW14) simulated dataset, multilocus genotypes were constructed using haplotypes at 2 of the 4 disease loci.

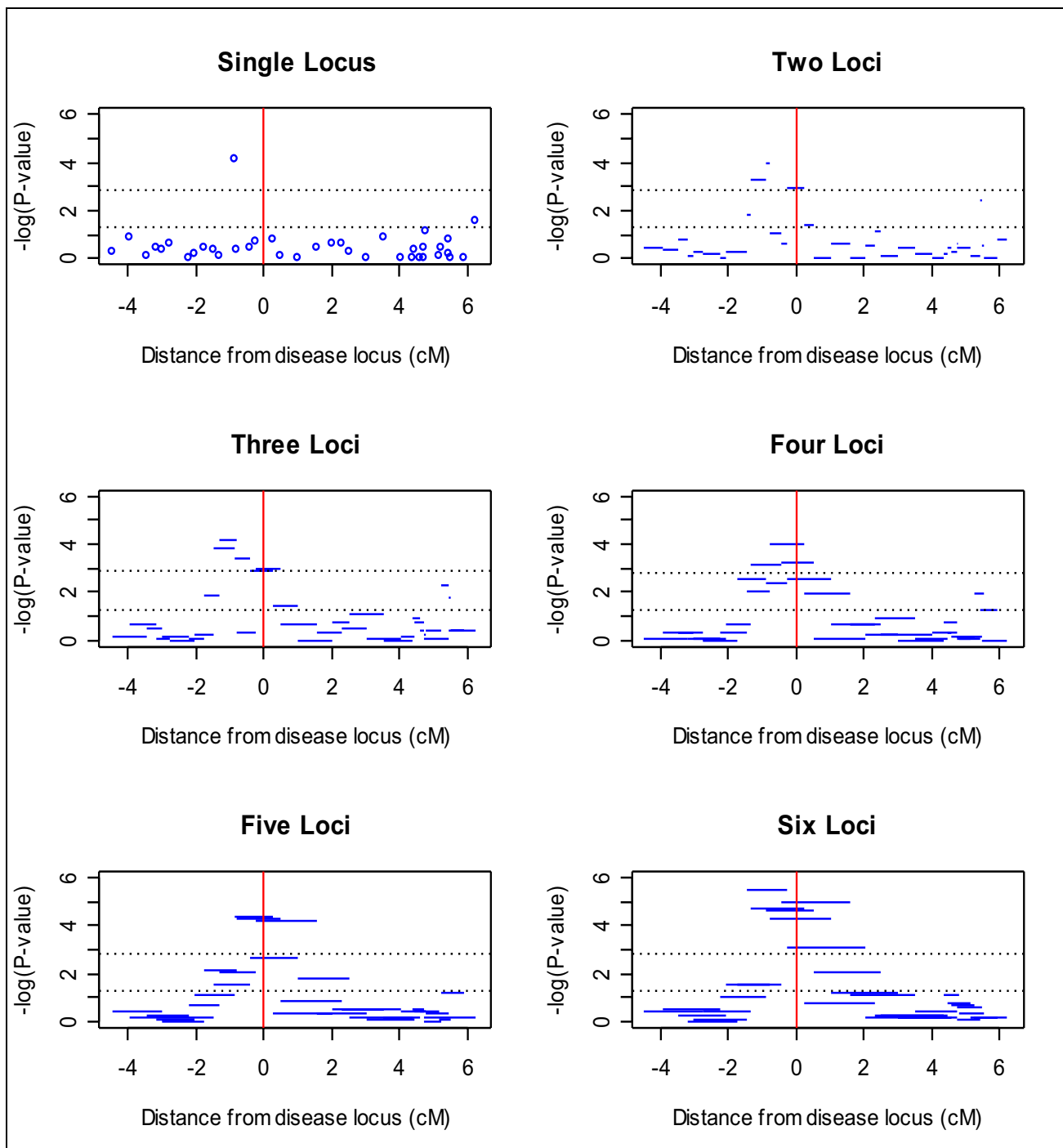
A major problem when conducting haplotype studies is that the available marker genotype data (often SNPs) are unphased, resulting in haplotype ambiguity. This fact has stimulated the development of a number of statistical techniques designed to reconstruct haplotype phase. Most of these methods treat the haplotypes as missing data and apply the expectation-maximization (EM) algorithm [2]

to infer haplotype frequencies by assuming these frequencies are in Hardy-Weinberg equilibrium (HWE) [3-5]. This assumption results in estimates and tests that are sensitive to the genetic structure of the sampled population, potentially leading to biased estimates and incorrect inferences [6]. Recently, we have developed an approach to testing haplotype/disease association that does not vary based on the distribution of haplotypes in the population [7]. In this work, we give an overview of this approach and apply it to triad datasets extracted from the GAW14 simulated data.

### Methods

#### Model and projection approach

We assume a case-parent sampling design in which individuals with disease ( $D = 1$ ) or trait of interest and their parents are sampled. At a locus of interest, let the proband's genotype be denoted  $G_o$  and let  $G_p$  denote the parents' genotypes. Let  $O = (G_o, G_p)$  denote the observed data (realizations are denoted by  $g_o, g_p, o$ ). Let the proband's haplotypes be denoted  $H_o$  and let  $H_p$  denote the parents' haplotypes (again, realizations are denoted by  $h_o, h_p$ ). As a first step, note that we can write the  $i^{\text{th}}$  family's contribution to the observed-data likelihood as



**Figure 1**  
**Result of overall haplotype score test with windows of 1 to 6 loci.** Upper and lower horizontal lines correspond to Bonferroni-corrected and uncorrected 0.05 significance levels, respectively. Vertical line corresponds to location of disease locus.

$$L_i^o = \sum_{h_o, h_p \in H(o_i)} \Pr(H_{o,i} = h_o | H_{p,i} = h_p, D = 1; \gamma) \Pr(H_{p,i} = h_p | D = 1; \eta),$$

where  $\gamma$  ( $q$ -dimensional) is the haplotype relative risk parameters of interest,  $\eta$  ( $r$ -dimensional) is nuisance parameters describing the distribution of parental haplotypes, and  $H(o)$  is the set of all  $h_o, h_p$  consistent with  $o$ . The model for  $\Pr(H_{o,i} = h_o | H_{p,i} = h_p, D = 1; \gamma)$  can be parameterized in a manner analogous to genotype relative risk models for case-parent designs [8]. We assume this model is correct (see below). This assumption is unnecessary when testing against the simple null hypothesis that no haplotype affects disease, but is required for valid parameter estimation or for tests that have a composite null hypothesis. The model for  $\Pr(H_{p,i} = h_p | D = 1; \eta)$  is problematic since: 1) a nonparametric model (i.e., saturated multinomial) is both computationally challenging and nonidentifiable, and 2) an incorrectly specified parametric model (i.e., HWE) can lead to significant bias in parameter estimates.

Our approach is to find the efficient score ( $U$ ) for the parameter of interest by projecting the observed data score onto the nuisance tangent space [9] assuming a saturated multinomial model for the distribution of parental haplotypes (note: HWE is not assumed). Computation of  $U$  is complicated by the fact that one must estimate  $\Pr(H_{p,i} = h_p | D = 1; \eta)$  and the nuisance parameter,  $\eta$ , is not identifiable in the saturated multinomial model. However, we have shown that  $U$  has mean zero regardless of the model used to estimate  $\Pr(H_{p,i} = h_p | D = 1; \eta)$  in the construction of the projection [7]. Hence, one can use an identifiable (but possibly incorrect) parametric model to estimate  $\Pr(H_{p,i} = h_p | D = 1; \eta)$  and use this estimate to compute  $U$ , secure in the knowledge that  $U$  will still have mean zero. This leads to tests and estimators of haplotype effects that are robust to the misspecification of the distribution of parental haplotypes. Note that both parameter estimates and their estimated variances are robust to misspecification of the distribution of parental haplotypes. Further, if the parental haplotype distribution is correctly specified, these estimators and tests will be optimal, having minimum variance among all estimators that are robust to misspecification of the parental haplotype distribution (for complete details see [7]).

Parameter estimates using our model depend on correct specification of the haplotype relative risk model  $\Pr(H_{o,i} = h_o | H_{p,i} = h_p, D = 1; \gamma)$ . However, the availability of unbiased tests and estimators makes it possible to select a model that fits the observed data without having to specify correctly the parental haplotype distribution. In one special situation, our tests (but not estimators) are robust to misspecification of the relative risk model: when testing the simple "global" null hypothesis that no haplotype

affects disease risk. In this case, correct specification of the risk model only affects the power of the test, not its size.

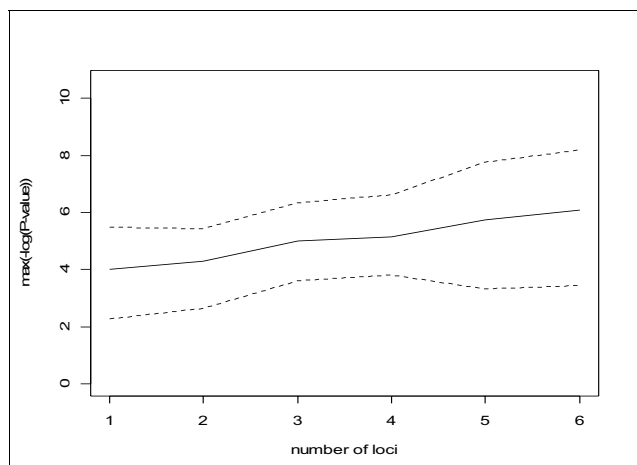
#### Application to GAW14 data

We applied this method to a hypothetical candidate gene study in which triads were ascertained from replicates 1–10 by the presence of subclinical phenotype  $c$  in an offspring. Only triads in the Aipotu, Karangar, and Danacaa populations were sampled (one triad per nuclear family). The genomic region and phenotype of interest were guided by consulting the answers throughout our analyses. For each replicate a region of the genome spanned by markers B09T8321–B09T8360 (known to contain trait locus D4 as well as linkage disequilibrium) was analyzed. Our goal was to study the effects of varying haplotype size on localization of a disease locus as we scanned across a region known to contain a disease locus. Because the D2 locus (the only other trait locus in a region of linkage disequilibrium) was at the end of a chromosome beyond the last visible SNP, we did not analyze on this region. We fit a multiplicative haplotype relative risk model in which we included a parameter for all haplotypes with frequencies greater than 5%. The haplotype with the highest frequency was used as a reference. The resulting score test is invariant to this reference choice, but parameter interpretation would be affected. We report  $-\log_{10}(p\text{-values})$  derived by an overall score test formed by evaluating the efficient score at  $\gamma = 0$ . Note the degrees of freedom for our test varied both with the number of SNPs per haplotype and the specific loci included, depending on how many haplotypes had frequency greater than 5%.

#### Results

Figure 1 presents the results of haplotype analyses of data replicate 1 using haplotype windows of 1 through 6 loci. Analyses with longer haplotypes seem to have higher  $-\log_{10}$ -transformed  $p$ -values with the 6 loci analysis having an order of magnitude smaller  $p$ -values than the single locus analysis. Interestingly, this increase in  $-\log_{10}$ -transformed  $p$ -values in the region of the disease loci was also accompanied with a suppression of a secondary (false) peak in the region of markers 30–40. It also appears that the analyses of longer haplotypes localize the disease locus better than the single locus analysis.

These are observations on a single simulated dataset. To determine a general pattern, we analyzed 9 other replicate datasets (datasets 2–10). Figure 2 presents a summary of the  $-\log_{10}$ -transformed  $p$ -values by size of analysis window for all 10 datasets. The solid line denotes the mean value of the maximum  $-\log_{10}(P\text{-value})$  (maximum is taken over the region). The upper and lower dotted lines represent the 80<sup>th</sup> and 20<sup>th</sup> percentiles of maximum  $-\log_{10}$ -transformed  $p$ -values, respectively. Larger analysis haplotype size does seem to lead to increased  $-\log_{10}$ -transformed  $p$ -



**Figure 2**  
**Maximum (across region within replicate) of  $-\log_{10}$ -transformed  $p$ -values.** Solid line: Mean value (over replicates). Upper and lower dotted lines: 80<sup>th</sup> percentile and 20<sup>th</sup> percentile of the maximum  $-\log_{10}$ -transformed  $p$ -values, respectively.

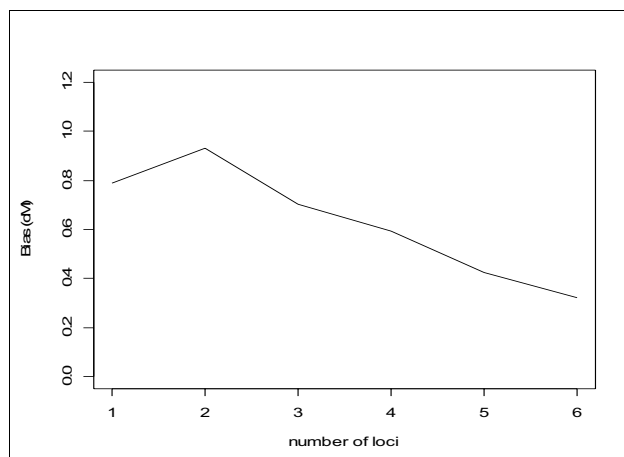
values. We confirmed this trend statistically by using generalized estimating equations (GEE) to regress maximum  $-\log_{10}$ -transformed  $p$ -values on the size of the haplotype window (GEE was used to account for correlation between analyses on the same dataset). We found that the trend for larger haplotype windows to have larger maximum  $-\log_{10}$ -transformed  $p$ -values, shown in Figure 2, was significant ( $p < 0.0001$ ).

We also investigated the effect of haplotype size on disease locus localization. For each replicate we averaged the locations of significant marker loci (or midpoint of significant haplotypes) and then computed distance (cM) between this average and the disease locus. We then averaged this bias over the replicated datasets. The results of this exercise are presented in Figure 3.

In the simulated datasets we analyzed, use of larger haplotypes does seem to better localize the disease locus. Again, we confirmed this trend by using GEEs to regress the average bias of significant loci on the size of the haplotype window. We found that the trend for larger haplotype windows to exhibit smaller bias, illustrated in Figure 3, was significant ( $p = 0.0141$ ).

## Conclusion

Though we attempted to approximate candidate gene studies of the various phenotypes, it is difficult to interpret the results because the data were certainly not generated with this intent in mind. For example, the haplotypes structure was only used to generate linkage disequilibrium and was not specifically related to phenotype. Neverthe-



**Figure 3**  
**Bias in disease locus localization.** Average distance to disease locus of average (over region) of significant (Bonferroni-corrected) loci.

less, the analyses we conducted do suggest a few patterns that merit further exploration. Analyses of longer (more SNPs) haplotypes seem to both result in smaller  $p$ -values in the disease region relative to single locus analyses and result in better localization of the disease loci. In addition, the longer haplotype analyses seemed to suppress secondary (false) peaks, resulting in fewer false positive signals.

## Abbreviations

EM: Expectation maximization

GAW14: Genetic Analysis Workshop 14

GEE: Generalized estimating equations

HWE: Hardy-Weinberg equilibrium

SNP: Single-nucleotide polymorphism

## Acknowledgements

ASA was supported by a research career award (K25 HL077663) from the National Heart, Lung, and Blood Institute, National Institutes of Health.

## References

1. Morris RW, Kaplan NL: **On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles.** *Genet Epidemiol* 2002, **23**:221-233.
2. Dempster AP, Laird NM, Rubin DB: **Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion).** *J R Stat Soc B* 1977, **39**:1-38.
3. Hawley ME, Kidd KK: **Haplo – a program using the EM algorithm to estimate the frequencies of multisite haplotypes.** *J Hered* 1995, **86**:409-411.
4. Excoffier L, Slatkin M: **Maximum-likelihood-estimation of molecular haplotype frequencies in a diploid population.** *Mol Biol Evol* 1995, **12**:921-927.

5. Long JC, Williams RC, Urbanek M: **An E-M algorithm and testing strategy for multiple-locus haplotypes.** *Am J Hum Genet* 1995, **56**:799-810.
6. Satten GA, Epstein MP: **Comparison of prospective and retrospective methods for haplotype inference in case-control studies.** *Genet Epidemiol* 2004, **27**:192-201.
7. Allen AS, Satten GA, Tsiatis AA: **Locally-efficient robust estimation of haplotype-disease association in family-based studies.** *Biometrika* 2005, **92**:559-571.
8. Schaid DJ: **Likelihoods and TDT for the case-parents design.** *Genet Epidemiol* 1999, **16**:250-260.
9. Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA: *Efficient and Adaptive Estimation for Semiparametric Models* New York: Springer; 1993.

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](http://www.biomedcentral.com/info/publishing_adv.asp)

