Haplotype Based Association Tests

Biostatistics 666

Last Lecture

- Statistical Haplotyping Methods
  - Clark's *greedy* algorithm
  - The E-M algorithm
  - Stephens et al's "coalescent-based" algorithm

Hypothesis Testing

- Often, haplotype frequencies are not final outcome.
- For example, we may wish to compare two groups of individuals…
  - Are haplotypes similar in two populations?
  - Are haplotypes similar in patients and healthy controls?

Today …

- Association tests for haplotype data
- When do you think these will out-perform single marker tests?
- When do you think these will be out-performed by single marker tests?

Introduction:
A Single Marker Association Test

- Simplest strategy to detect genetic association
- Compare frequencies of particular alleles, or genotypes, in set of cases and controls
- Typically, relies on standard contingency table tests…
  - Chi-squared Goodness-of-Fit Test
  - Cochran-Armitage Trend Test
  - Likelihood Ratio Test
  - Fisher's Exact Test

Construct Contingency Table

- Rows
  - One row for cases, another for controls
- Columns
  - One for each genotype
  - One for each allele
- Individual cells
  - Count of observations, with double counting for allele tests

Simple Association Study

	Genotype		
	1/1	1/2	2/2
Affecteds	$n_{a,11}$	$n_{a,12}$	$n_{a,22}$
Unaffecteds	$n_{u,11}$	$n_{u,12}$	$n_{a,22}$

Organize genotype counts in a simple table...

Notation

- Let index *i* iterate over rows
  - E.g. *i* = *1* for affecteds, *i* = *2* for unaffecteds
- Let index *j* iterate over columns
  - E.g. *j* = *1* for genotype 1/1, j = 2 for genotype 2/2, etc.
- Let O_{ij} denote the observed counts in each cell
  - Let $O_{\bullet\bullet}$ denote the grand total
  - Let $O_{i\bullet}$ and $O_{\bullet j}$ denote the row and column totals
- Let E_{ij} denote the expected counts in each cell
  - $E_{ij} = O_{i\bullet} O_{\bullet j} / O_{\bullet\bullet}$

Goodness of Fit Tests

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- If counts are large, compare statistic to chi-squared distribution
  - p = 0.05 threshold is 5.99 for 2 df (e.g. genotype test)
  - p = 0.05 threshold is 3.84 for 1 df (e.g. allele test)
- If counts are small, exact or permutation tests are better

Likelihood Ratio Test

$$G^2 = -\sum_{ij} 2O_{ij} \ln \frac{O_{ij}}{E_{ij}}$$

- If counts are large, compare statistic to chi-squared distribution
  - p = 0.05 threshold is 5.99 for 2 df (e.g. genotype test)
  - p = 0.05 threshold is 3.84 for 1 df (e.g. allele test)
- If counts are small, exact or permutation tests are better

Haplotype Association Test
A Simple Straw Man Approach

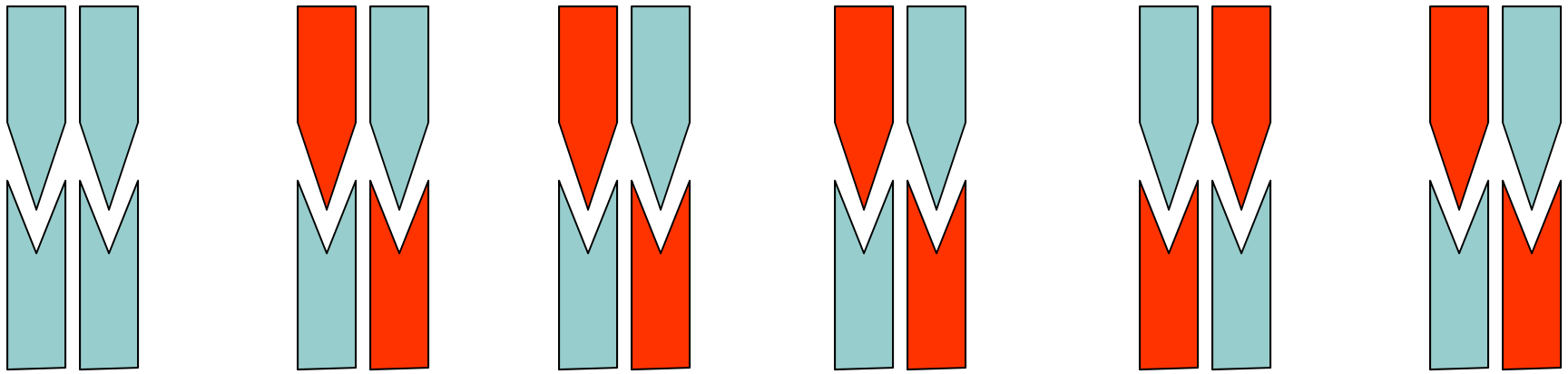
- Calculate haplotype frequencies in each group
- Find most likely haplotype for each individual
- Fill in contingency table to compare haplotypes in the two groups

Haplotype Association Test
A Simple Straw Man Approach

- Calculate haplotype frequencies in each group
- Find most likely haplotype for each individual
- Fill in contingency table to compare haplotypes in the two groups

NOT RECOMMENDED!!!

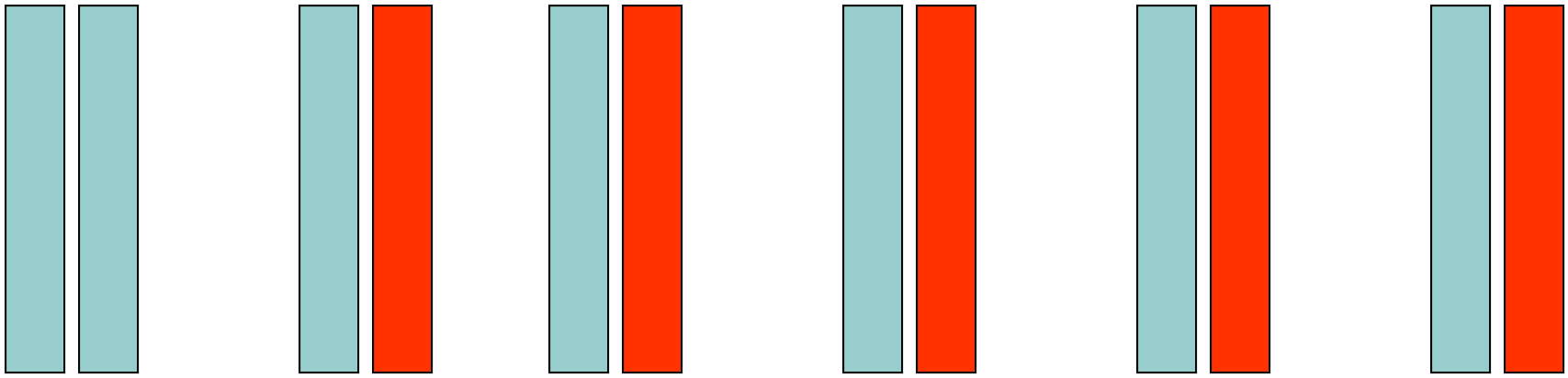
Observed Case Genotypes



1 2 3 4 5 6

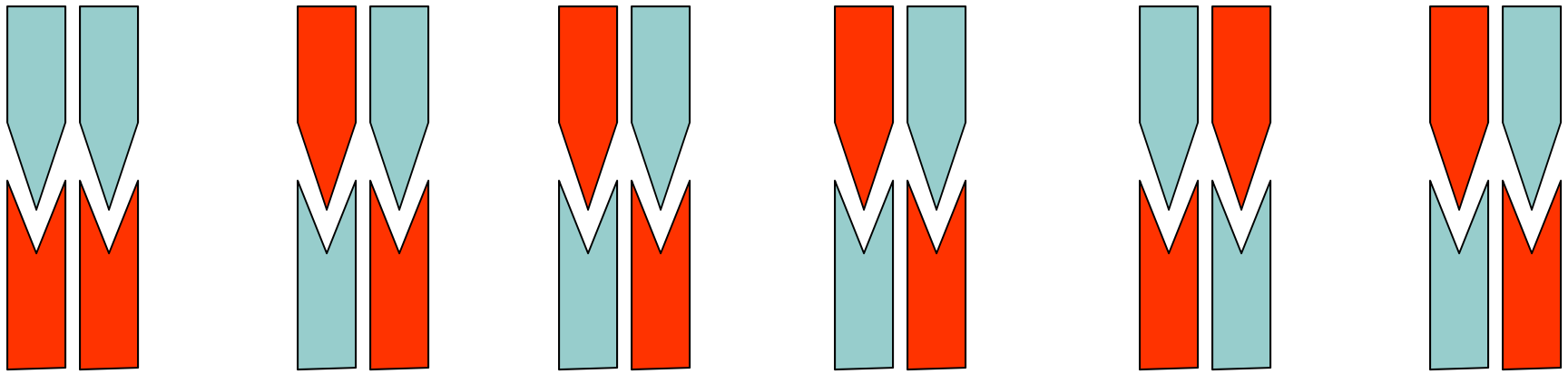
The phase reconstruction in the five ambiguous individuals will be driven by the haplotypes observed in individual 1 …

Inferred Case Haplotypes



This kind of phenomenon will occur with nearly all population based haplotyping methods!

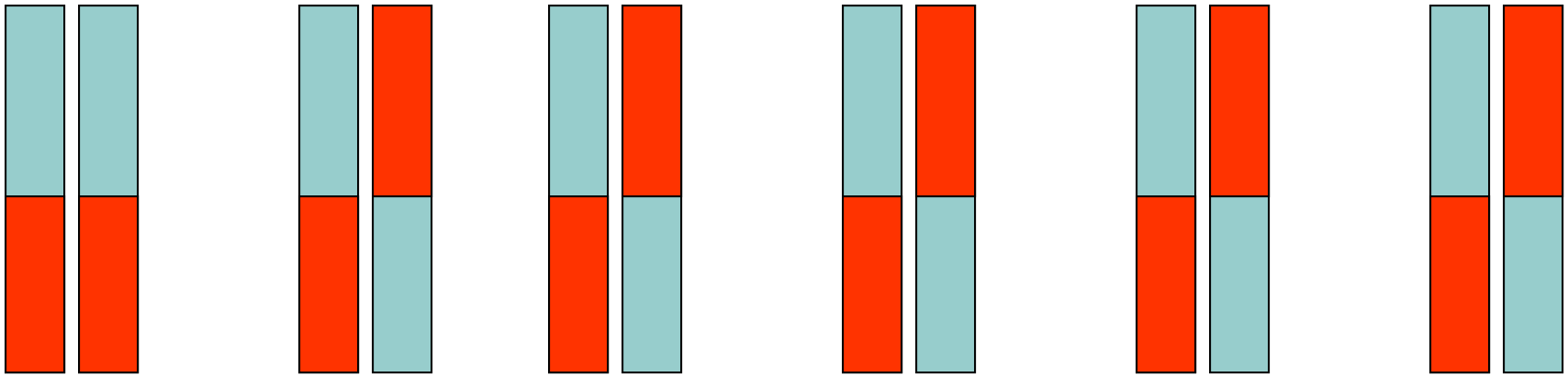
Observed Control Genotypes



1 2 3 4 5 6

Note these are identical, except for the single homozygous individual …

Inferred Control Haplotypes



1 2 3 4 5 6

Ooops… The difference in a single genotype in the original data has been greatly amplified by estimating haplotypes…

Common Sense Rules for Haplotype Association Tests

- Never impute haplotypes in two samples separately
- Use maximum likelihood
  - Does not require imputing individual haplotypes
  - Likelihood statistic can allow for uncertainty
- If haplotypes imputed, treat cases and controls jointly
  - Schaid et al (2002) *Am J Hum Genet* **70**:425-34
  - Zaytkin et al (2002) *Hum Hered.* **53**:79-91

Likelihood Function for Haplotype Data

- Estimated haplotype frequencies, imply a likelihood for the observed genotypes

$$L = \prod_{i} \sum_{H \sim G_i} P(H)$$

Likelihood Function for Haplotype Data

- Estimated haplotype frequencies, imply a likelihood for the observed genotypes

$$L = \prod_{i} \sum_{H \sim G_i} P(H)$$

individuals

possible haplotype pairs, conditional on genotype

haplotype pair frequency

Likelihood Ratio Test For Difference in Haplotype Frequencies

- Calculate 3 likelihoods:
  - Maximum likelihood for combined sample, L_A
  - Maximum likelihood for control sample, L_B
  - Maximum likelihood for case sample, L_C

$$2\ln\left(\frac{L_B L_C}{L_A}\right) \sim \chi^2_{df}$$

df corresponds to number of non-zero haplotype frequencies in large samples

Significance in Small Samples

- In realistic sample sizes, it is hard to estimate the number of *df* accurately
- Instead, use a permutation approach to calculate empirical significance levels

Permutation Approach ...

- Can you propose one?

A More General Approach

- Zaykin, Westfall, Young, et al (2002) *Hum Hered* 53:79-91

- Provides estimates of haplotype effects
- Can be used with quantitative traits
- Can incorporate covariates

Regression Model

- Predictors
  - Haplotype counts
- Regression Parameters
  - Phenotypic effect of each haplotype
- Outcome
  - The phenotype of interest

Exemplar Design Matrix

μ	h_1	h_2	h_3

$$E\begin{Bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{Bmatrix} = \begin{Bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1/2 & 1/2 \\ 1 & 1/2 & 0 & 1/2 \end{Bmatrix} \begin{Bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{Bmatrix}$$

Hypothetical set-up when observed haplotypes are:

h_1/h_1 for individual 1
h_2/h_3 for individual 2
h_1/h_3 for individual 3

Permutations Are Very Efficient

$$\hat{\beta} = \mathbf{P}'\mathbf{Y}$$

$$\mathbf{P} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}$$

Note that $\mathbf{P}$ does not vary with permutation so that we need only recalculate $\mathbf{P}'\mathbf{Y}$.

Dealing With Unphased Data

- Calculate weights for each configuration
  - Function of observed genotype
  - Function of estimated frequencies
- Fill in design matrix with partial counts

$$\Pr(h_2, h_3 \mid G_i) = \frac{\Pr(G_i \mid h_2, h_3) p_{h_2} p_{h_3}}{\sum_{u,v} \Pr(G_i \mid h_u, h_v) p_{h_u} p_{h_v}}$$

Simulated Example, Single Marker Analysis

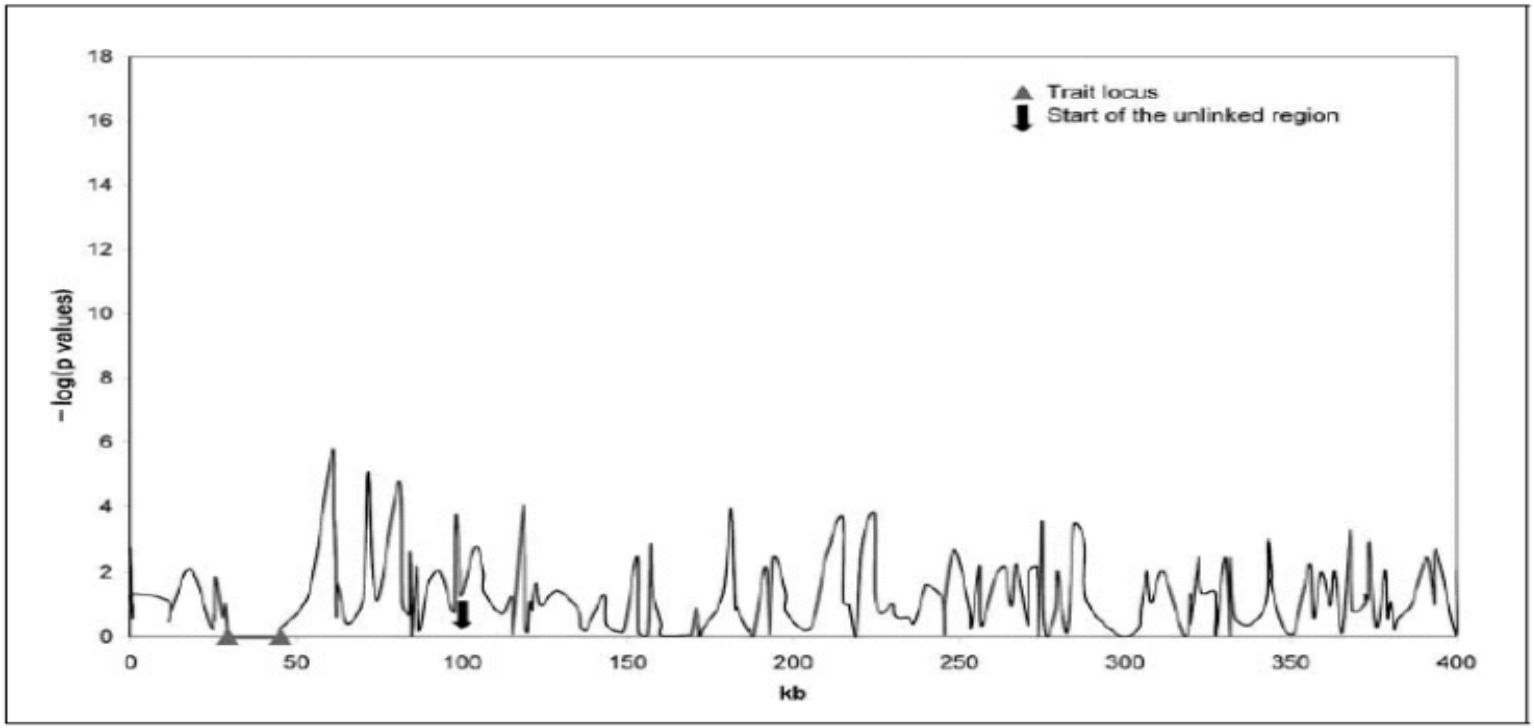


Fig. 1. Sample –log(p values) against the marker map plots for window size of 1 using p values from the asymptotic *F* test.

Simulated Example, Three Marker Windows

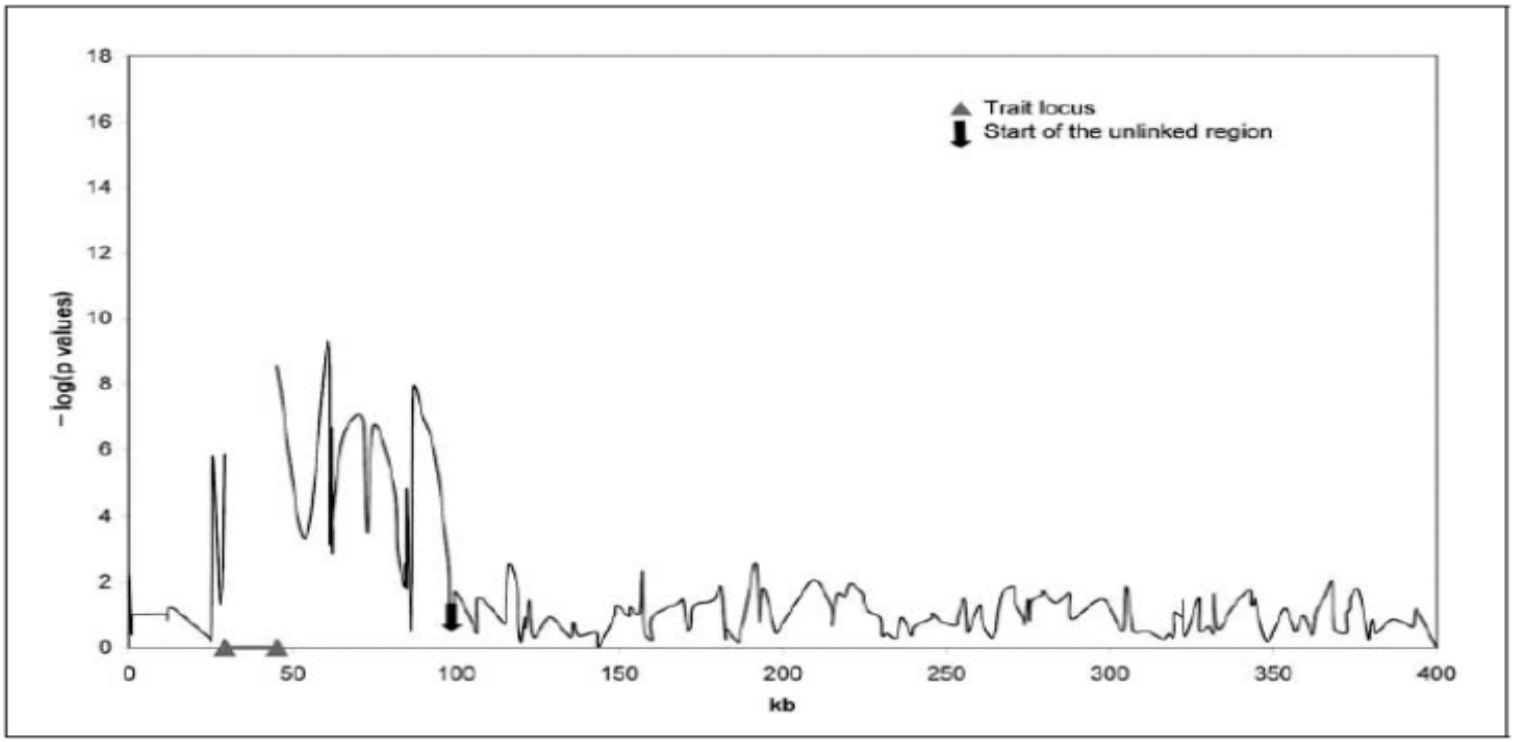


Fig. 2. Sample –log(p values) against the marker map plots for window size of 3 using p values from the asymptotic *F* test.

Simulated Example, Five Marker Windows

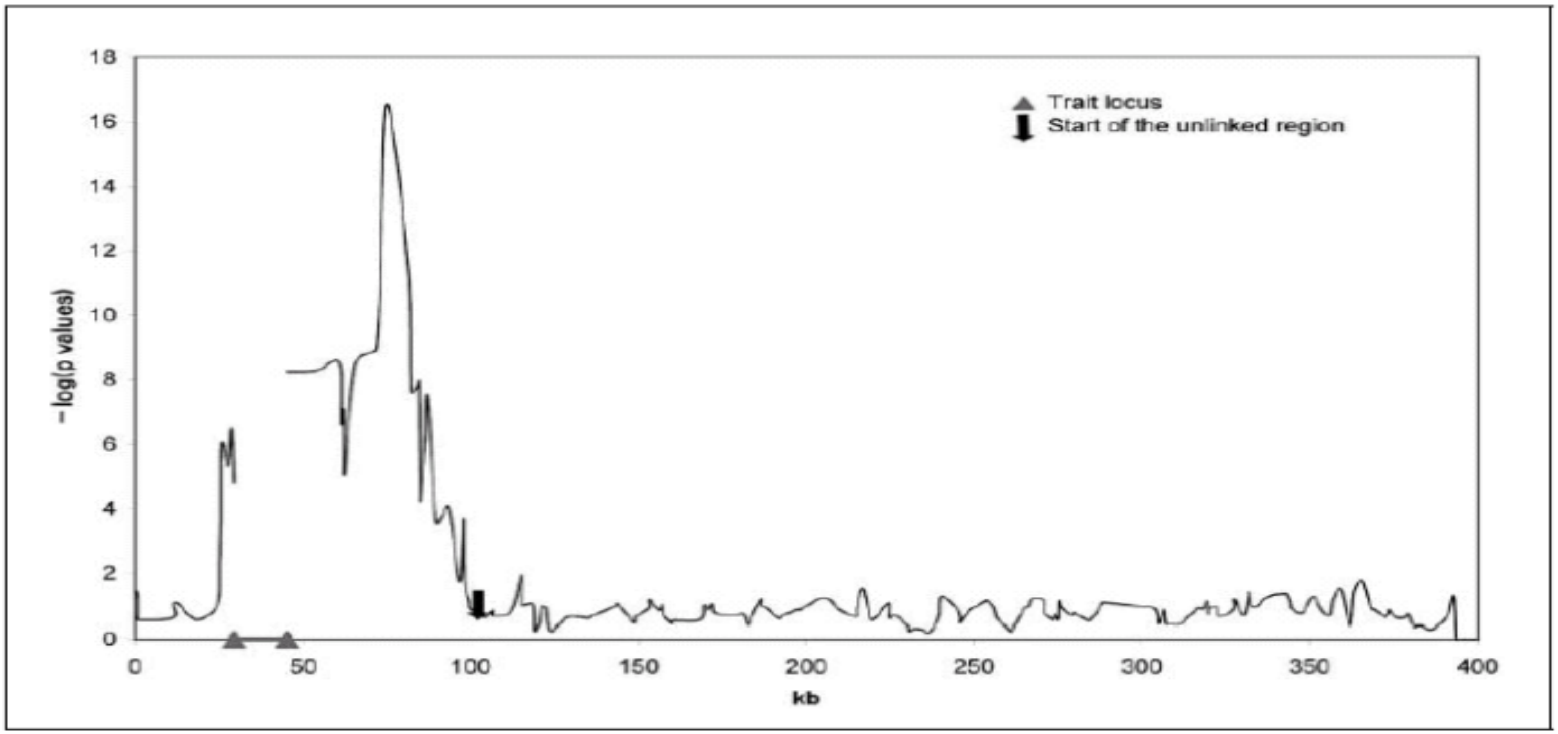


Fig. 3. Sample –log(p values) against the marker map plots for window sizes of 5 using p values from the asymptotic *F* test.

Loss of Power
Due to Unobserved Haplotypes

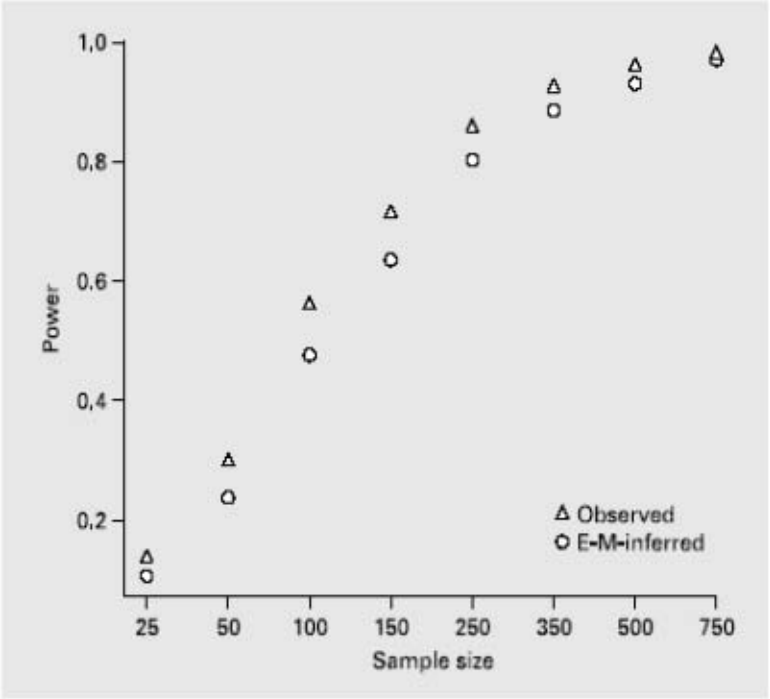


Fig. 5. Power values against the sample size for observed and E-M-inferred three marker haplotypes (HTR tests).

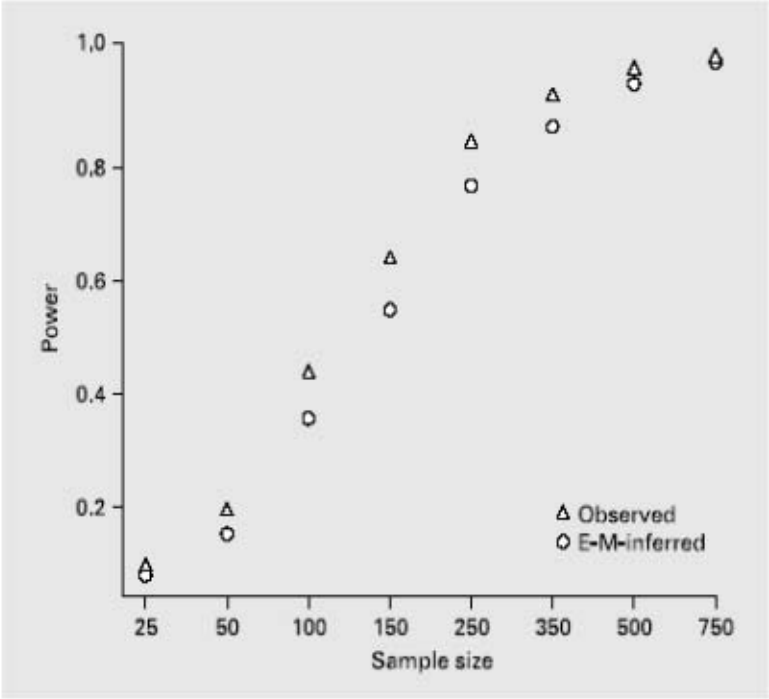


Fig. 6. Power values against the sample size for observed and E-M-inferred five marker haplotypes (HTR tests).

Comparison of Regression and Maximum Likelihood Approaches

	Haplotype size					
	1	2	3	4	5	6
HTR-A (H_0)	0.056	0.034	0.033	0.029	0.022	0.027
HTR-P (H_0)	0.048	0.051	0.050	0.051	0.048	0.049
HTR-A (H_A)	0.321	0.352	0.373	0.412	0.408	0.427
HTR-P (H_A)	0.315	0.365	0.396	0.449	0.448	0.491
LRT-P (H_A)	0.310	0.357	0.388	0.420	0.444	0.436

A = Asymptotic test; P = permutational test.

Zaykin et al. Approach

- Regression based
  - Estimated haplotype counts as predictors
- Can also be applied to discrete traits
  - For example, using logistic regression
- To accommodate multiple correlated tests, significance should be evaluated empirically

Further Refinements

- When there are many haplotypes, fitting one effect per haplotype is inefficient
- Instead, it might be desirable to group haplotypes
  - This may also be helpful when for capturing the effect of unmeasured alleles
- We will summarize the suggestions of
  - Morris et al (2004), *Am J Hum Genet* **75:**35-43

Grouping Haplotypes to Learn About Unobserved Alleles

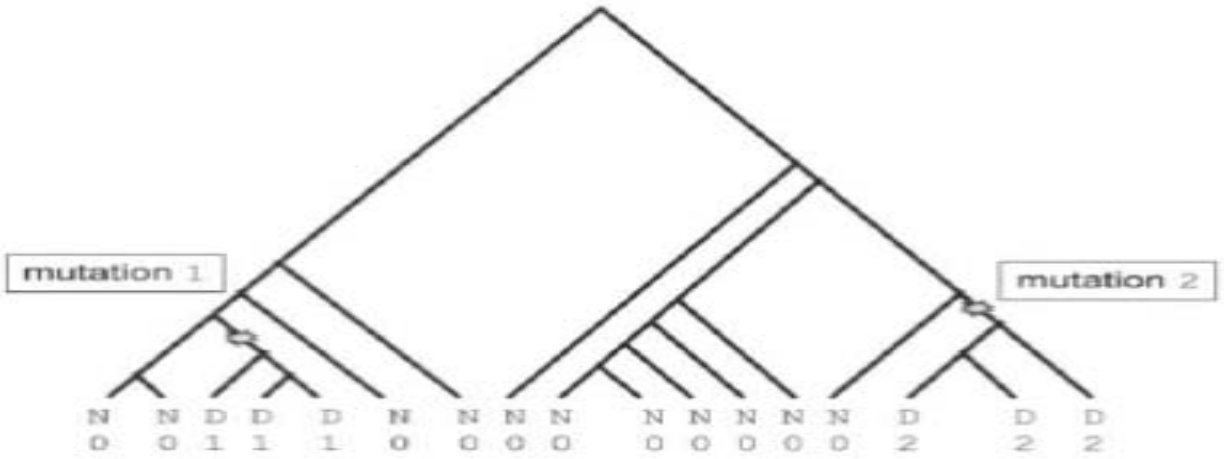


Figure 1 Example of a genealogical tree representing the shared ancestry of chromosomes at the disease gene. Disease chromosomes (D) carrying the same mutation (1 or 2), share more recent common ancestry than normal chromosomes (N) carrying no mutation (0).

Morris et al. (2004) Approach

- Assume that haplotypes are observed
  - In practice, assign most likely haplotype
- Calculate a distance between haplotype pairs and build simple cladogram
  - Using hierarchical group averaging

Haplotype Grouping Reduces Number of Effects in the Model

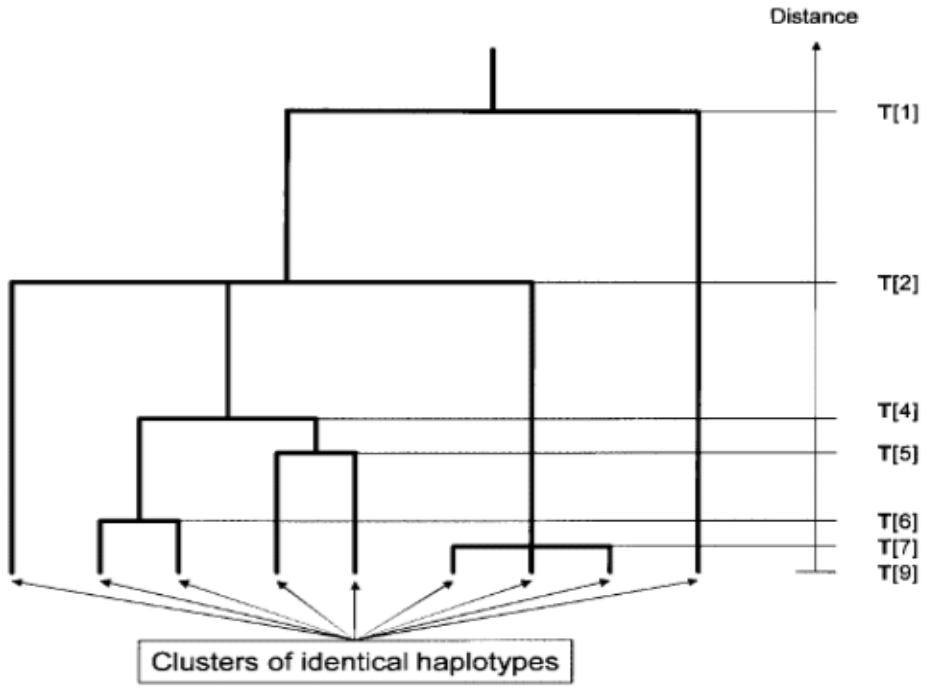


Figure 2 Example of a cladogram representing haplotype diversity within a window of SNPs. The cladogram is constructed using hierarchical group average clustering on pairwise haplotype differences, expressed in terms of the proportion of marker mismatches within the window of SNPs.

Then ...

- Each level of cladogram suggests one possible analysis
- Carry out all possible analyses
  - 9 groups at level T[9]
  - 7 groups at level T[7]
  - etc.
- Select the best fitting model
- Evaluate significance by permutation

Final thoughts…

- Haplotype analyses can improve power
  - Must be carefully planned
- Always evaluate significance empirically
  - Randomize case-control labels

Summary

- Today we discussed issues relating to haplotype based association tests

- A good paper to read is:

  - Zaykin, Westfall, Young, et al (2002) *Hum Hered* 53:79-91