GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Transmission Disequilibrium Test for fine mapping based on haplotypes

Xiangdong Ding and Henner Simianer

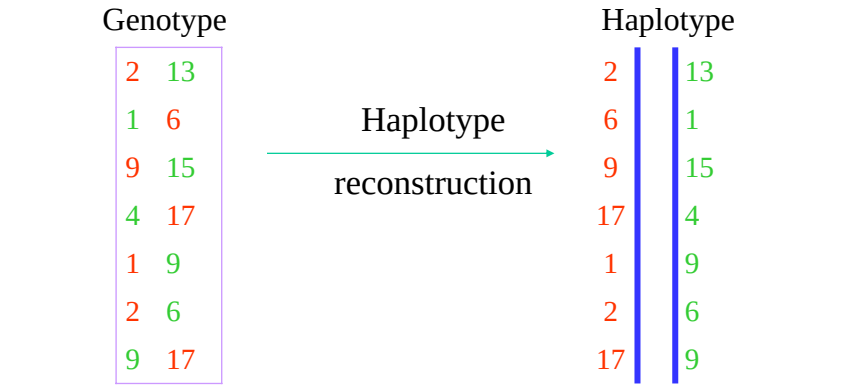
Institute of Animal Breeding and Genetics,
University of Goettingen,Germany

Outline

- Haplotype inference
  - Introduction of important methods
    - Parsimony (Clark,1990)
    - EM (Excoffier and Slatkin, 1995)
    - Bayesian (Stephens and Donnelly,2001)
  - Haplotype inference using family information
- Transmission Disequilibrium Test (TDT)
- Haplotype-based TDT

Genotype and Haplotype

A collection of alleles derived from the same chromosome

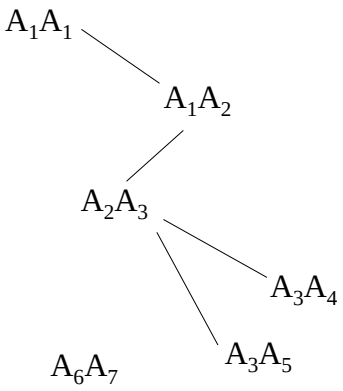


Algorithms for haplotype reconstruction

- Statistical methods
  - Parsimony (Clark,1990)
  - EM
    - Excoffier and Slatkin (1995); Hawley and Kidd, (1995); Qin *et al*. (2002)
  - Bayesian
    - Stephens and Donnelly (2001); Niu *et al*. (2002)
- Rule-based methods
  - Minimum recombination principle
    - Qian and Beckmann (2002); Li and Jiang (2003); Baruch, *et al*. (2006)

Parsimony (Clark,1990)

1. Start from a homozygote
2. Determine any other ambiguous sequence using the definitive haplotype from 1
3. Continue this procedure until all haplotypes are resolved or until no more new haplotypes can be found



Clark's Parsimony

- Disadvantages:
  - No starting point for algorithm;
  - Individuals may remain phase indeterminate;
  - Biased estimates of haplotype frequencies.

EM algorithm: Excoffier and Slatkin (1995)

- Numerical method of finding maximum likelihood estimates for parameters given incomplete data.

1. Initial parameter values: haplotype frequencies
2. *Expectation step*: compute expected values of missing data based on initial data
3. *Maximization step*: compute MLE for parameters from the complete data
4. Repeat with updated set of parameters until changes in the parameter estimates are negligible.

EM algorithm: Excoffier and Slatkin (1995)

Probability of the i^{th} diplotype made up of haplotype k and l

# of diplotypes to j^{th} phenotype

Different phenotypes

$$L(p_1, p_2, \ldots, p_h) = \prod_{j=1}^{m} \left(\sum_{i=1}^{c_j} P(h_{ik} h_{il}) \right)^{n_j}$$

Population frequencies of all haplotypes

EM algorithm

Expectation step: caculate the probability of each possbile diplotype for j^{th} phenotype

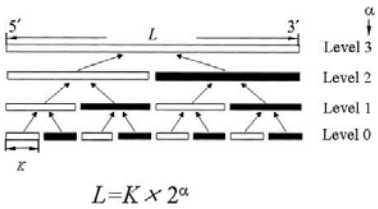
$$P_j(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{P(h_k h_l)}{P_j^{(g)}}$$

Maximization step: update the haplotype frequencies

$$\hat{p}_t^{(g+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P_j(h_k h_l)^{(g)}$$

EM algorithm efficiency

- Heavy computational burden with large number of loci
  - Partition-ligation algorithm (Niu et al., 2002)
  - PL-EM (Qin et al., 2002)

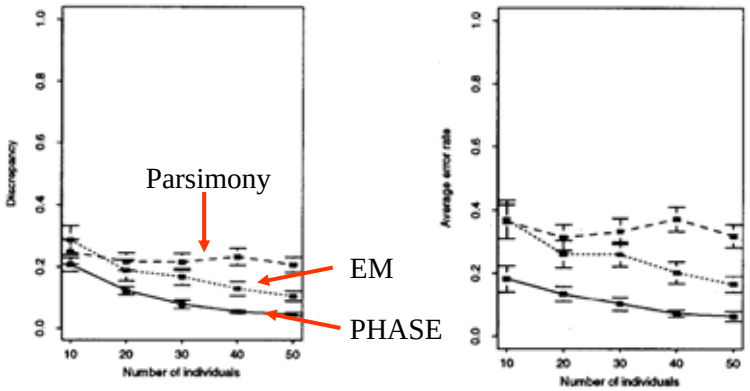


- Accuracy and departures from HWE
  - Assumption of HWE in most EM-based methods
  - Robust to departure from HWE (Fallin and Schork, 2000)

Bayesian haplotype reconstruction

- PHASE (Stephens and Donnelly, 2001)
  - Based on coalescent model
  - Use Gibbs sampling
  - So far, very accurate, but also complicated.

Comparison of Parsimony,EM and PHASE



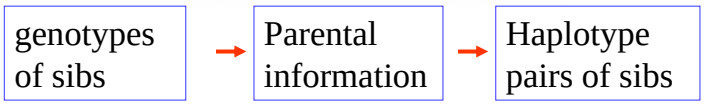
- PHASE performs better than parsimony and EM (Stephen, 2001)
- PHASE and EM-based methods exhibited similar performances (Zhang et al. 2001; Xu et al. 2002)

Haplotype inference using family data (1)

- Haplotype inference based on close relatives
  - Reduces haplotype ambiguity and improves the efficiency
- Rohde and Fuerst (2001) – EM algorithm
  - Families with both parents and their children
  - The genotyped offspring reduce the number of potential haplotype pairs for both parents.
- Ding and Simianer (2006) - EM algorithm
  - Families with only one parent available
  - Parent-child pair with one shared haplotype.

Haplotype inference using family data (2)

- Ding and Simianer – EM algorithm
  - Families with only sibs



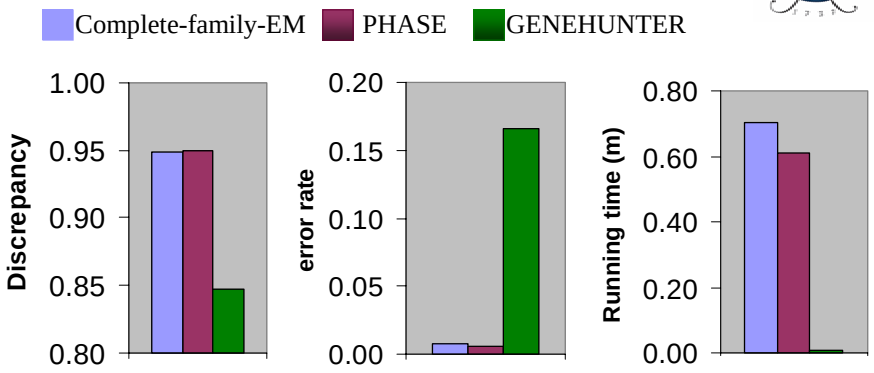
- Mixed family data
  - Complete families
  - Incomplete families
    - One parent
    - Only sibs

Comparison of four different strategies

Method name	Using family information?	Using LD?	Handling incomplete families?
Complete-family-EM (Rhode and Fuerst, 2001)	YES	YES	NO
Incomplete-family-EM (Ding et al., 2006)	YES	YES	YES
GENEHUNTER (Kruglyak et al., 1996)	YES	NO	YES
PHASE (Stephens et al., 2003)	YES	YES	YES

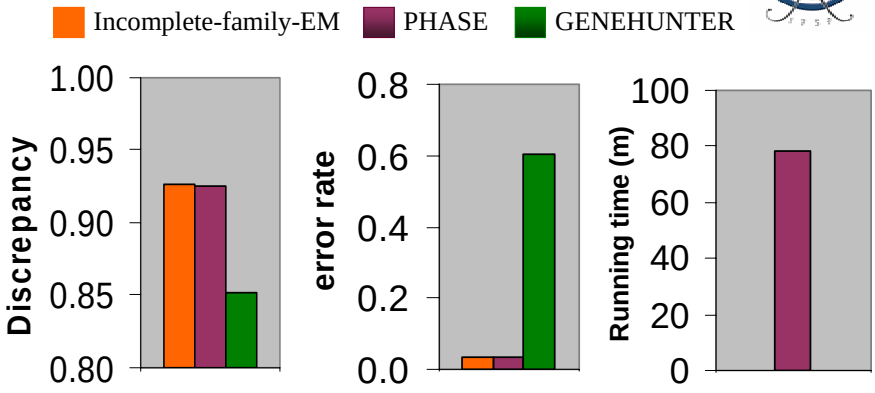
(*Ding and Simianer,2006*)

Result for complete families



- Simulation program based on coalescent model (Schaffner et al.,2005): 30 trios, 20SNPs
- *Discrepancy*: 1- *sum* (|estimated *p* - actual *p*|)
- *Error rate*: the proportion of wrongly haplotyped individuals

Result for incomplete families



Running time of PHASE:

- 3.5 hs for the whole 100 datasets of 30 trios,187 SNPs (Marchini et al.,2006)
- Running time will become prohibitive for large SNPs

Rule-based method

- Minimum recombination principle
  - Qian and Beckmann (2002); Li and Jiang (2003); Baruch, et al. (2006)
- Genetic recombination is rare
- Haplotype with fewer recombinants should be preferred in a haplotype reconstruction

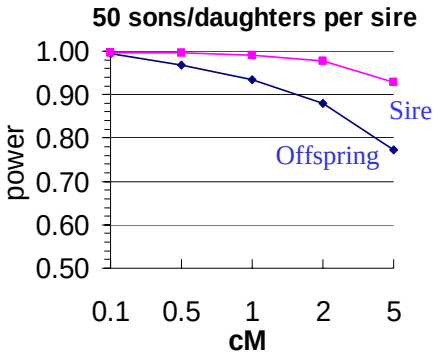
Joint EM and rule-based algorithm for (grand-) daughter design

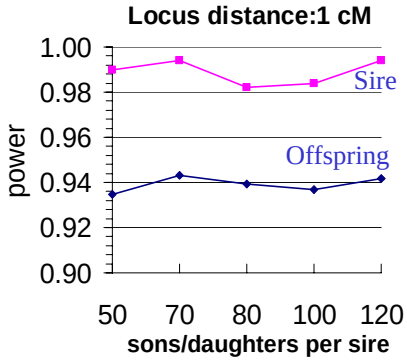
- Assumption of no recombination
  - EM algorithm to construct diplotype
- Taking into account recombination
  - Minimum recombination principle
    - Derive possible diplotypes of sire from all sire-offspring pairs in one sire family
    - Find the diplotype of sire that minimizes the number of recombinations in the sire family

Example:

Possible diplotypes	*recom. events*
1. 54731722 31761329	47
2. 51731729 34761322	46
3. 51731722 34761329	**45**
4. 34731729 51761322	47
5. 34761729 51731322	47
6. 51761329 34731722	46
7. 54731329 31761722	48
8. 54731322 31761729	49
9. 51731329 34761722	46

Result





- 10 sires
- 5 markers, 6 alleles with equal allele frequency each

TDT (Transmission Disequilibrium Test)

- Compares the distribution of transmitted and non-transmitted alleles by parents of affected offspring (Spielman et al. 1993)

$$X^2 = (b-c)^2 / (b+c) \sim \chi^2_{1df}$$

	Non-transmitted allele		**total**
transmitted allele	**M_1**	**M_2**	
M_1	**a**	**b**	**a+b**
M_2	**c**	**d**	**c+d**
total	**a+c**	**b+d**	**2n**

If the marker is unlinked to the causative locus then we expect b = c, else, one of the alleles will tend to be transmitted more often

TDT (Transmission Disequilibrium Test)

- Good for fine-mapping, poor for initial detection
- Robust for population stratification/admixture
- Initially for test of linkage,currently used for association
- Extension of TDT
  - Multi allelic markers (Sham and Curtis, 1995)
  - Multiple siblings (Spielman et al.,1998; Boehnke et al., 1998)
  - Missing parental data (Sun, 1999)
  - Extended pedigree (Martin et al., 2000)
  - Quantitative traits (Allison,1997; Rabinowitz,1997; Sun,2000)

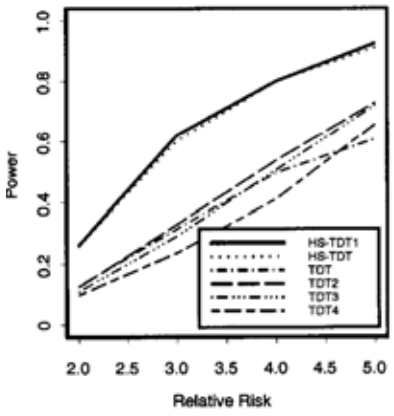
Haplotype-based TDT

- The original TDT and most of its extensions consider one marker at a time. Haplotypes are more informative than single markers.
- Two categories of haplotype-based TDT
  - Haplotype reconstruction first
    - Sethuraman (1997); Wilson (1997); Clayton and Jones (1999); Zhao et al. (2000); Zhang et al.(2003)
  - Implicit haplotype reconstruction
    - Dudbridge (2003)

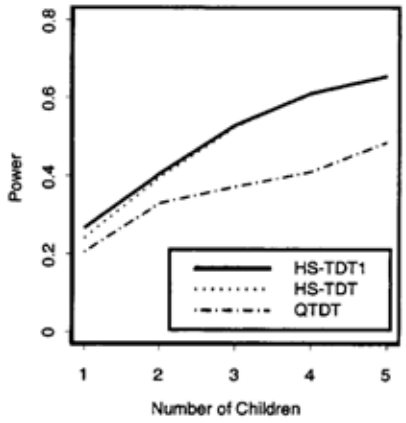
Haplotype-based TDT vs TDT

Zhang et al. (2003)

Qualitative trait



Quantitative trait ($h^2 = 0.06$)



Haplotype-based TDT

- Problem of multiple comparisons
  - Increase in the degree of freedom

More markers → A large number of haplotypes → A large number of degrees of freedom → Limit the power of test

Method to reduce degree of freedom

- Group the haplotypes
  - Estimated evolutionary relationships (Setman et al. 2001)
- Maximum identity length contrast
  - Compare the mean shared length of the transmitted haplotypes and the mean shared length of the non-transmitted haplotypes
    - Bourgain et al. (2000,2001,2002); Zhang et al. (2003)

Thanks FUGATO program of the German Federal Ministry of Education and Research

Thanks for your attention!