G6.22 alessandra.stella@tecnoparco.org

The use of phenotypic information for refinement of haplotype reconstruction using the Expectation-Maximization algorithm

A. Stella[1,], and P. J. Boettcher[2];*
[1] *Parco Tecnologico Padano, Lodi, Italy;*
[2] *Istituto di Biologia e Biotecnologia Agraria, Consiglio Nazionale delle Ricerche, Milano, Italy*

INTRODUCTION

Standard genotyping methods generally do not provide the information about linkage phase necessary to define haplotypes of several loci. Therefore, *in silico* methods have been developed for this purpose (Niu, 2004). One of the most popular approaches is the Expectation-Maximization (EM) approach of Excoffier and Slatkin (1995). Explained simply, this method determines the population frequencies for all haplotypes (estimation). Each individual is then assumed to have the most common haplotype configuration (maximisation) among those that are plausible (based on its genotype). Reconstruction is based entirely on population genotype data. In some cases, haplotype reconstruction is performed with the final goal of estimating genetic associations between the haplotypes and a given phenotype. The evaluation of haplotype effects is usually a 2-process: 1) reconstruction of haplotypes and 2) estimation effects. However, when a true association exists, the phenotype should theoretically provide information that could be used to reconstruct haplotypes more accurately, which should also lead to more precise estimation of haplotype effects. The objective of this work was to develop a method based on the EM algorithm that simultaneously reconstructs haplotypes and estimates their effects. Stochastic simulation was used to test the utility of this method.

MATERIAL AND METHODS

Steps in Haplotype Reconstruction

1. Obtain haplotype probabilities for each individual using standard EM method.
   - P_{H1}= p(haplotype | population frequencies)
2. Estimate haplotype effects (using an iterative solving procedure)
3. Adjust haplotype probabilities based on phenotype and estimated effects
   - P_{H2} = (haplotype | population frequencies, phenotype, haplotype effects)
   - $P_{H2} = P_{H1} \times P(y|$ haplotype effects)

-the latter term is calculated based on the residual and the normal p.d.f

4. Repeat steps 1 to 3 until population haplotype frequencies are stable.

MATERIAL AND METHODS

Simulation

1. 500 individuals, 5 locus (biallelic) haplotypes
2. High, medium, or low haplotype effects (25, 10, or 2% of population variance)
3. Infinite (all haploytypes possible) or selected (10 possible haplotypes) populations
4. 500 replicates

Comparisons

1. Haplotype Reconstruction
   - Accuracy of haplotype definition
2. Effect Estimation
   - Correlation with true values
   - F-test

RESULTS

- Considering phenotype during reconstruction increases accuracy, with the benefit increasing as the size of the effect and number of haplotypes increases (Figure 1).
- Simultaneously reconstructing haplotypes and estimating effects increases significance tests, decreasing Type II error (Figure 2).
- However, effects tend to be over-estimated (increasing Type I error) and accuracy tends to be higher only when true effects are large (>10% of total variance)
- Accuracy of reconstruction is decreased when no effects of the haplotype exist (data not shown)

Figure 1. Accuracy of haplotype reconstruction for different populations and haplotype effects.

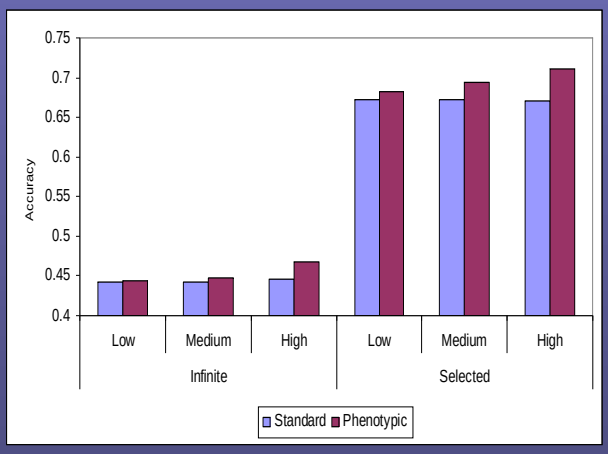
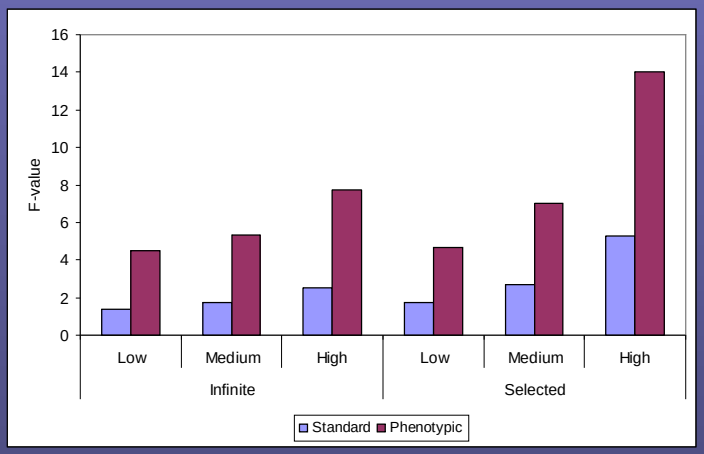


Figure 2. F-values for haplotype effects for different populations and haplotype effects.



CONCLUSIONS

- In certain situations, considering the phenotype increases the accuracy of haplotype reconstruction with the simple EM algorithm
- This general concept could be developed for other haplotype reconstruction algorithms.
- One possible use of this algorithm is in developing countries, where lack of recording and historical data precludes the availability of information from the parental generation

ABSTRACT NUMBER: 477 -Session G6.22, 56th EAAP Annual Meeting, Uppsala, Sweden