GENECOUNTING: haplotype analysis with missing genotypes

Jing Hua Zhao1,∗, Sebastien Lissarrague2, Laurent Essioux3 and Pak Chung Sham4

1Department of Epidemiology and Public Health, University College London, 1–19 Torrington Place, London WC1E 6BT, UK, 2Genset SA, Site SNECMA RN7, 91030 Evry, France, 3ValiGen SA, Tour Neptune, 92086 Paris-La-Défense, France and 4Section of Genetic Epidemiology, PO Box 80, Institute of Psychiatry, London SE 8AF, UK

Received on November 22, 2001; revised on March 25, 2001; accepted on May 21, 2001

ABSTRACT

Summary: A general algorithm is described for haplotype analysis of unrelated individuals with missing genotypes. It can handle problems involving multiple polymorphic markers with missing data.

Availability: GENECOUNTING is available from http://www.iop.kcl.ac.uk/IoP/Departments/PsychMed/GEpiBS/software.stm

Contact: j.zhao@public-health.ucl.ac.uk; p.sham@iop.kcl.ac.uk

Maximum likelihood estimation of haplotype frequencies from unphased, multi-locus genotype data can be carried out by a special case of EM algorithm (Dempster et al., 1977) that involves iterative counting of haplotypes. In principle, this algorithm can be extended to take account of missing genotypes, but current implementations of the algorithm only deal with limited amounts of missing data or do not deal with missing data appropriately.

The standard gene-counting algorithm for haplotype frequency estimation formulates the problem of uncertain phase as incomplete data, and consists of an E-step where the expected counts of the phased genotypes are calculated using current haplotype frequency estimates, and an M-step where the expected counts are summed over all individuals to provide revised haplotype frequency estimates. The treatment of missing genotype data requires a generalization of the E-step to consider all possible phased genotypes that are consistent with the non-missing genotypes of each individual. We have implemented this algorithm in a program called GENECOUNTING.

Algorithm G (gene counting with missing genotypes). We classify genotypic configurations into those with and without missing data; the counts of these are denoted as

n and m, respectively, so that the total sample size is

\[ N = \sum_{p=1}^{P} n_p + \sum_{q=1}^{Q} m_q \]

if there are P configurations without missing data and Q configurations with missing data. If the haplotype frequencies are denoted as \( h \), then the probability of each configuration without missing genotype, \( g \), is a function of \( h \), under the assumption of random mating. Furthermore, the probability of each configuration with missing genotypes, \( t \), is a ‘marginal’ probability defined as the sum of all the \( g \)’s which have the same genotypes at the non-missing markers. For clarity below let \( c \), \( c' \) and \( c'' \) be the haplotype counts from complete data, data with ambiguous phase but no missing genotype, and data with with missing genotype, respectively.

\[ G_1 \ [Initialize] \ set \ c, c' \ and \ c'' \ to be zero, set haplotype frequencies \( h \) (e.g. at random or the product of allele frequencies), calculate genotype probabilities from haplotype frequencies and obtain log-likelihood \( l \).
\[ G_2 \ [save log-likelihood] \ l_s \leftarrow l \]
\[ G_3 \ For each configuration with no missing data and at most one heterozygous marker, deduce the two haplotypes and count the \( 2^n_p \) haplotypes into \( c \).
\[ G_4 \ For other configurations, do iterative counting through steps \( G_5 \) to \( G_8 \).
\[ G_5 \ [test for missing genotype] \ If there is missing genotype goto step \( G_7 \).
\[ G_6 \ [count using data without missing genotypes] \]
- count number of heterozygotes \( m \)
- obtain phase probabilities for \( 2^{n-1} \) phases
- count for each phase the two haplotypes each by \( (phase \ probability) \times n_p \) into \( c' \)
\[ G_7 \ [count using data with missing genotypes] \]
- list all possible genotypes for each configuration
- for each possible genotype calculate its probability \( (g) \)

*To whom correspondence should be addressed.
of these configurations in terms of haplotype frequencies. The
probabilities of phases at steps $G_5$ and $G_7$ are recursive,
making it easy to accommodate different numbers of loci.
Full details of the implementation has been described
elsewhere (Zhao and Sham, 2002).

We illustrate the algorithm for the simple case of two
biallelic markers. Table 1 defines the possible configura-
tions and their counts, while Table 2 gives the probabilities
of these configurations in terms of haplotype frequencies.

<table>
<thead>
<tr>
<th>Marker 1</th>
<th>Marker 2</th>
<th>1/1</th>
<th>1/2</th>
<th>2/2</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n_3$</td>
<td>$m_1'$</td>
<td></td>
</tr>
<tr>
<td>1/2</td>
<td>$n_4$</td>
<td>$n_5$</td>
<td>$n_6$</td>
<td>$m_2'$</td>
<td></td>
</tr>
<tr>
<td>2/2</td>
<td>$n_7$</td>
<td>$n_8$</td>
<td>$n_9$</td>
<td>$m_3'$</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>$m_1$</td>
<td>$m_2$</td>
<td>$m_3$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Genotypic probabilities for two biallelic markers

<table>
<thead>
<tr>
<th>Marker 1</th>
<th>Marker 2</th>
<th>1/1</th>
<th>1/2</th>
<th>2/2</th>
<th>$i'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/1</td>
<td>$h_{11}^2$</td>
<td>$2h_{11}h_{12}$</td>
<td>$h_{12}^2$</td>
<td>$i_3'$</td>
<td></td>
</tr>
<tr>
<td>1/2</td>
<td>$2h_{21}h_{11}$</td>
<td>$2h_{21}h_{12}$ + $2h_{22}h_{11}$</td>
<td>$2h_{22}h_{12}$</td>
<td>$i_2'$</td>
<td></td>
</tr>
<tr>
<td>2/2</td>
<td>$h_{21}^2$</td>
<td>$2h_{21}h_{22}$</td>
<td>$h_{22}^2$</td>
<td>$i_3'$</td>
<td></td>
</tr>
<tr>
<td>$t$</td>
<td>$t_1$</td>
<td>$t_2$</td>
<td>$t_3$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- accumulate total probabilities for configuration ($t$)
- perform $G_6$ using $m_q g/t$ as $n_p$ to obtain ($c''$)

$G_8$ [obtain haplotype frequencies and log-likelihood] set
$h ← (c + c' + c'')/(2N)$ and calculate log-likelihood $l$
$G_9$ [test for convergence] if $l - l_s > \epsilon$ save log-
likelihood $l_s ← l$ and goto step $G_4$

Algorithm G is suitable when data are missing at
random. In GENECOUNTING the initialization of $h$
at step $G_1$ (when assuming linkage equilibrium) and
enumeration of phases at steps $G_5$ and $G_7$ are recursive,
making it easy to accommodate different numbers of loci.

We thank Dr Andrew Pakstis for his willingness to analyse
a SNP data using HAPLO. We also thank editor Dr Babara
Cox and reviewers for their help and comments during
preparation of the manuscript. This work is supported by
Wellcome project grant 055379.

REFERENCES

Dempster,A.P., Laird,N. and Rubin,D.B. (1977) Maximum likeli-
hood from incomplete data via the EM algorithm. J. Royal Stat.

inference for multiple linked single-nucleotide polymorphisms.

method for haplotype reconstruction from population data. Am.

Zhao,J.H. and Sham,P.C. (2002) Generic number system and