Haplotype Inference for the MHC Haplotype Map Project

Jonathan Marchini

Department of Statistics, University of Oxford, Oxford, UK

January 11, 2006

Summary

This short note describes the details of the how haplotypes were infered for MHC Haplotype Map Project.

CEU and YRI Analysis Panels

For each panel we split each autosome in to non-overlapping segments of 100 SNPs and a new version of PHASE (v2.1) that deals with father-mother-child trios was applied to each segment in turn. We used the default settings of PHASE and used the recommended parent-independent mutation (PIM) model at the HLA loci. We took the most likely haplotype reconstruction returned by the program as the infered haplotypes. To link the haplotypes from each segment back together we used the information at non-ambiguous sites in each trio to infer the transmission status of each parental haplotype. That is, for each parent we worked out which haplotype was transmitted to the child and which haplotype was untrans-

mitted. For each trio we then aligned the transmitted and untransmitted haplotypes across the segments to produce haplotypes across each autosome.

CHB and JPT Analysis Panels

For each panel we split each autosome in to non-overlapping segments of 100 SNPs and ran PHASE (v2.1) on each segment in turn. We used the default settings of PHASE and used the recommended parent-independent mutation (PIM) model at the HLA loci. We took the most likely haplotype reconstruction returned by the program as the infered haplotypes. To link the haplotypes from each segment back together we developed an approach that infers the best alignment of sets of haplotypes across two adjacent segments. The details of this approach are as follows.

Let $H^1 = \{H^1_1, H^1_2, \ldots, H^1_n\}$ and $H^2 = \{H^2_1, H^2_2, \ldots, H^2_n\}$ be two sets of infered haplotypes for n individuals on two adjacent segments of SNPs where $H^i_j = \{H^i_{j1}, H^i_{j2}\}$ are the two infered haplotypes for the jth individual in the ith segment. We use $H = \{H_1, H_2, \ldots, H_n\}$ to denote the set of pairs of haplotypes for the n individuals across the two segments that we wish to infer. We use $Z = \{z_1, \ldots, z_n\}$ to denote a set of n binary variables that denote the alignment of haplotypes across the two segments in each individual. This is, if $z_k = 0$ then the haplotypes across the two segments in the kth individual, H_k, are $H^1_{k1} \oplus H^2_{k1}$ and $H^1_{k2} \oplus H^2_{k2}$ where the $\oplus$ is used to denote two haplotypes joined together end-to-end. If $z_k = 1$ then $H_k = \{H^1_{k1} \oplus H^2_{k2}, H^1_{k2} \oplus H^2_{k1}\}$. Thus, Z can be thought as a vector of switches that indicate the alignment of haplotypes across the two segments and we wish to infer the most likely alignment. To do this we try to

optimise the following pseudo-likelihood function

$$S(H) = \prod_{j=1}^{n} \pi(H_{j1}|H_{-j1}, \rho, \mu)\pi(H_{j2}|H_{-j2}, \rho, \mu)$$

where π is the conditional distribution from Li and Stephens (2003) and the H_{-j1} denotes the set of haplotypes H minus the haplotype H_{j1}. The paramter ρ is an estimate of the population-scaled recombination rate from the PHASE algorithm applied to each segment. The recombination rate in the interval between the two segments was taken as a local average of the rates in the 5 intervals either side of that interval. The parameter μ is controls the mutation rate used in the model (see Li and Stephens (2003) for more details).

To optimize the function we used simulated annealing with multiple start points. The output of the method is the vector of switches Z that result in a set of haplotypes H that gave the highest pseudo-likelihood. We tested this method on simulated datasets analyzed in a comparison of leading phasing algorithms (Marchini et al., 2006). We found that the combining PHASE infered haplotypes across segments produced more accuratly infered haplotypes across the whole region than any of the other methods compared in the study. This method (called PHASElink) was applied to all pairs of adjacent segments of PHASE infered haplotypes in the CHB and JPT panels and a set of haplotypes across each autosome was reconstructed from the infered set of swicthes between each pair of segments.

Software

The PHASElink software is available upon request from Jonathan Marchini at `marchini@stats.ox.ac.uk`.

References

Li N, Stephens M (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165:2213–2233.

Marchini J, Cutler D, Patterson N, Stephens M, Eskin E, Halperin E, Lin S, et al (2006). A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet (in press).