The C. savignyi Genetic Map and its Integration with the Reference Sequence Facilities Insights into Chordate Genome Evolution

M Hill, KW Broman, E Stupka, W Smith, D Jiang, A Sidow

Supplement on statistical methods for genetic map construction

Here we provide further details on the statistical methods used to construct the genetic map. Much is repeated from the Mapping subsection of the Methods section in the paper, but we expand our explanations in several areas.

Assessment of linkage between pairs of markers

We began by considering pairs of markers. For each pair considered, we estimated the recombination fraction, θ, between them (assuming no sex difference in recombination) and calculated a LOD score comparing the hypothesis that the two markers are linked to the hypothesis that they are not linked ($\theta = 1/2$).

Denoting the ordered parental genotypes at a marker as AB and CD, the F_1 progeny have genotypes either AC, AD, BC or BD. However, precise genotypes for the genetic markers were not observed; rather, each marker exhibited two, three or four distinct banding patterns across the set of progeny. The connection between banding patterns and the underlying genotypes was inferred as part of the process of establishing linkage between marker pairs. In each case, we considered all possible assignments of genotypes to banding patterns.

We first considered all markers with four distinct banding patterns. While at a given marker, there are 24 possible ways to assign banding patterns to genotypes, out inability to distinguish the order of the two parents, or the order of the two haplotypes within a parent, results in 72 unique ways to assign the banding patterns for two markers to genotypes. For a given pair, we consider each of these 72 possible assignments, and estimate the recombination fraction between the markers and calculate the LOD score for linkage, assuming the assignment is the correct one. The inferred assignment is that with the largest LOD score (and so the maximum likelihood). For a given assignment, the recombination fraction can be estimated by simply counting recombination events. The log likelihood is the sum, across individuals, of the log probability of the observed two-locus genotype; these probabilities are displayed in Table 1. The LOD score is the $\log_{10}$ likelihood ratio comparing the hypothesis that the two loci are linked (and using the maximum likelihood estimate of the recombination fraction, $\hat{\theta}$) and the hypothesis that the two loci are not linked ($\theta = 1/2$).

For each of the markers with just two or three distinct banding patterns, we considered linkage to each of the fully informative markers, and not to each other. The establishment of linkage required that we consider all possible partitions of the four marker genotypes to the two or three observed banding patterns. For example, for a marker with two banding patterns, the first banding patterns could be assigned to a single genotype out of the four, to one of the six possible pairs of genotypes, or to one of the four possible groups of three genotypes. Thus there are 14 possible ways to assign the four genotypes to the two banding patterns. For markers with three banding patterns, there are 36 possible ways to assign genotypes to banding patterns.

Table 1: Two-locus genotype probabilities, assuming known phase and no sex difference in recombination.

	Marker 1			
Marker 2	**AC**	**AD**	**BC**	**BD**
AC	$(1-\theta)^2/4$	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$	$\theta^2/4$
AD	$\theta(1-\theta)/4$	$(1-\theta)^2/4$	$\theta^2/4$	$\theta(1-\theta)/4$
BC	$\theta(1-\theta)/4$	$\theta^2/4$	$(1-\theta)^2/4$	$\theta(1-\theta)/4$
BD	$\theta^2/4$	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$	$(1-\theta)^2/4$

θ = recombination fraction

In establishing linkage between a partially informative marker and a fully informative marker, we again consider all possible assignments of marker genotypes to banding patterns. For each such assignment, we again seek to estimate the recombination fraction between the markers and a LOD score for the test of linkage. But, in the case of a partially informative marker, estimation of the recombination fraction can not always be accomplished by simply counting recombination events, as for many individuals it will not be clear, for example, whether there was no recombination or two recombination events.

Consider, for example, the case of Table 2, in which the first marker has three banding patterns, with one assigned to the pair of genotypes AD and BC. Individuals with that pattern at the first marker and with genotype AD at the fully informative marker may have had no recombination event or two recombination events.

Table 2: Two-locus probabilities for a fully informative marker and a partially informative marker, assuming no sex difference in recombination, for a given connection between marker genotypes and banding patterns.

	Marker 1		
Marker 2	**AC**	**AD/BC**	**BD**
AC	$(1-\theta)^2/4$	$\theta(1-\theta)/2$	$\theta^2/4$
AD	$\theta(1-\theta)/4$	$[(1-\theta)^2+\theta^2]/4$	$\theta(1-\theta)/4$
BC	$\theta(1-\theta)/4$	$[(1-\theta)^2+\theta^2]/4$	$\theta(1-\theta)/4$
BD	$\theta^2/4$	$\theta(1-\theta)/2$	$(1-\theta)^2/4$

θ = recombination fraction

To obtain the maximum likelihood estimate of the recombination fraction in such cases, we use the EM algorithm (Dempster et al. 1977). This is an iterative algorithm in which we begin with an initial estimate of the recombination fraction, and then using that estimate and conditioning on the observed data, calculate the expected numbers of individuals in each of the 16 two-locus genotype classes. (For example, suppose there are 5 individuals with genotype AD or BC at the partially informative marker and genotype AD at the fully informative marker. If our current estimate of the recombination fraction is 0.2, then the expected number of individuals who are AD at both markers is $5(1-\hat{\theta})^2/[(1-\hat{\theta})^2+\hat{\theta}^2] \approx 4.7$, and the expected number of indivduals who are BC at the partially informative marker and AD at the fully informative marker is $5\hat{\theta}^2/[(1-\hat{\theta})^2+\hat{\theta}^2] \approx 0.3$.) We

then re-estimate the recombination fraction, using expected counts for the 16 two-locus genotype classes in place of true counts (which are unknown). The process is repeated to convergence. The log likelihood is again the sum, across individuals, of the log probability of the observed two-locus genotype, but using probabilities from Table 2, or an analogous table.

Formation of linkage groups

The pairwise linkage results (among all fully informative markers and between the partially informative markers and the fully informative markers) were used to establish linkage groups. Two markers were placed in the same group if the estimated recombination fraction between them was no more than 0.25 and the LOD score for a test of linkage was at least 4.5. The transitive property (if A is linked to B and B is linked to C, then A is linked to C) was used to close the linkage groups.

Parental haplotype inference within each linkage group

We used the pairwise linkage information to infer the parental haplotypes within each linkage group, though recognizing that the order of the two parents and the order of the two haplotypes within each parent cannot be identified. (Thus, the haplotypes in one linkage group cannot immediately be attached to the haplotypes in another linkage group.) The haplotypes were formed starting with a pair of closely linked markers, and then working through the rest of the markers in the linkage group, considering one additional marker at a time. (For each marker pair, we used the inferred connection between marker genotypes and the banding patterns, from the pairwise linkage information.)

Ordering of markers within linkage groups and multipoint estimation of inter-marker distances

Taking the inferred parental haplotypes to be known, marker order was determined by considering all possible orders of markers or, for the larger linkage groups, all possible orders for a sliding window of markers. The chosen order was that with the maximum likelihood (that is, the marker order for which the observed data were most probable). Multipoint estimates of the recombination fractions between markers were also estimated by maximum likelihood, assuming no crossover interference. Likelihood calculations were performed via the Lander-Green algorithm (Lander and Green 1987). The estimated recombination fractions between adjacent markers were transformed to genetic distances using the Haldane map function. After the initial establishment of marker order, we used the locations of markers within reftigs to refine marker order, when possible.

Large gaps in the estimated linkage maps indicated the possibility that a linkage group should be split in two. In such cases, we calculated a LOD score comparing the hypothesis that the two linkage group should remain merged to the hypothesis that they should be distinct. A linkage group was split into two if this LOD score was not large (< 3). We similarly considered merging pairs of linkage groups; doing so required the consideration of the 8 possible connections between the inferred haplotypes in one linkage group with those in a second linkage group.

Pairwise linkage calculations, the establishment of linkage groups, and the inference of parental haplotypes, were accomplished with perl scripts written specific to the current data. The establishment of marker order and the multipoint estimation of inter-marker distances were accomplished

with R/qtl (Broman et al. 2003), an add-on package for the general statistical software, R (Ihaka and Gentleman 1996).

References

Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889–890

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Stat Soc B 39: 1–38

Ihaka R, Gentleman R (1996) R: A language for data analysis and graphics. J Comp Graph Stat 5: 299–314

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84: 2363–2367