Haplotype probabilities in advanced intercross populations

Karl W. Broman[1]

Department of Biostatistics and Medical Informatics,

University of Wisconsin–Madison, Madison, Wisconsin 53706

28 October 2011

Running head: AIL haplotype probabilities

Key words: advanced intercross lines, heterogeneous stock, diversity outcross, map expansion, Collaborative Cross

[1]**Corresponding author:**

Karl W Broman

Department of Biostatistics and Medical Informatics

University of Wisconsin–Madison

1300 University Ave, Rm 4710 MSC

Madison, WI 53706

Phone: 608–262–4633

Fax: 608–265–7916

Email: kbroman@biostat.wisc.edu

ABSTRACT

Advanced intercross populations have the advantage of greater precision of genetic mapping, due to the accumulation of recombination events across the multiple generations. Related designs include heterogeneous stock and the diversity outcross population. We derive the two-locus haplotype probabilities on the autosome and X chromosome with these designs.

Advanced intercross populations, in which multiple inbred strains are mated at random for many generations, have the advantage of greater precision of genetic mapping, due to the accumulation of recombination events across the multiple generations. The most commonly used form, which begins with two inbred strains, was formally introduced by DARVASI and SOLLER (1995) and called advanced intercross lines (AIL). A closely related design is that of heterogeneous stock (HS; see MOTT *et al.* 2000), in which eight inbred strains are randomly mated for many generations. SVENSON *et al.* (2012) developed the diversity outcross population (DO), which was formed with progenitors that were partially inbred individuals drawn from intermediate generations in the development of the Collaborative Cross (so-called pre-CC mice; see AYLOR *et al.* 2011).

The mapping of quantitative trait loci (QTL) in such populations, whether by interval mapping (LANDER and BOTSTEIN 1989) or Haley-Knott regression (HALEY and KNOTT 1992), generally requires conditional genotype probabilities at putative QTL, given the available marker genotype data. Such probabilities are often calculated using a hidden Markov model (HMM; see BROMAN and SEN 2009, App. D). An HMM for this purpose formally requires the calculation of two-locus diplotype probabilities, though if the populations are formed with a large number of mating pairs, the two haplotypes within an individual are independent, and so it is sufficient to calculate two-locus haplotype probabilities.

DARVASI and SOLLER (1995) derived the two-locus haplotype probabilities for the autosome in AIL. I am not aware of any work considering the X chromosome. In this paper, I derive the two-locus haplotype probabilities for the autosome and X chromosome in AIL, HS and the DO. The calculations for the DO rely on recent results on haplotype probabilities in

pre-CC mice (BROMAN 2012). Throughout, I assume an effectively infinite set of mating pairs at each generation, no sex difference in recombination, and no selection or mutation.

Let us first revisit the two-locus autosomal haplotype probabilities in AIL, as they serve as a simple example of the technique used in these calculations (see also BULMER 1980, Ch. 3). Let p_s denote the frequency of the AA haplotype at generation F_s. Then $p_1 = \frac{1}{2}$ and we have the recurrence relation

$$p_{s+1} = (1 - r)p_s + r \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} \tag{1}$$

where r is the recombination fraction (in one meiosis) between the two loci. Equation (1) is derived by noting that an AA haplotype drawn from generation F_{s+1} is either an intact AA haplotype at generation F_s, transmitted without recombination, or it is a recombinant haplotype bringing two independent A alleles together. Note that the frequency of the A allele is $\frac{1}{2}$ at every generation.

The solution of this recurrence relation (see GRAHAM *et al.* 1994) is, for $s \geq 2$,

$$p_s = \tfrac{1}{4}\left[1 + (1 - 2r)(1 - r)^{s-2}\right]. \tag{2}$$

The frequency of recombinant haplotypes at generation F_s is $1 - 2p_s$.

For the X chromosome in AIL, I will first consider a balanced case, begun with equal proportions of F_1 individuals from reciprocal crosses, $A \times B$ and $B \times A$, so that the F_1 males are equally likely to be hemizygous A or B. Let m_s and f_s denote the frequency of the AA

haplotype in males and females, respectively, at generation F_s. Then $m_1 = f_1 = \frac{1}{2}$ and we have

$$\begin{aligned} m_{s+1} &= (1-r) f_s + \tfrac{r}{4} \\ f_{s+1} &= \left(\tfrac{1}{2}\right) m_s + \left(\tfrac{1-r}{2}\right) f_s + \tfrac{r}{8} \end{aligned} \tag{3}$$

This recurrence relation is derived in a similar way to that for the autosome, noting that the male haplotype was drawn from his mother, with a chance for recombination, and a random female haplotype is equally likely to have been drawn from her father, without recombination, or from her mother, with the potential for recombination. I again make use of the fact that the frequency of the A allele is $\frac{1}{2}$ in both males and females at every generation. The solution to this relation is, for $s \geq 2$,

$$\begin{aligned} m_s &= \tfrac{1}{8}\left[2 + (1-2r)(w^{s-2} + y^{s-2}) + \left(\tfrac{3-5r+2r^2}{z}\right)(w^{s-2} - y^{s-2})\right] \\ f_s &= \tfrac{1}{8}\left[2 + (1-2r)(w^{s-2} + y^{s-2}) + \left(\tfrac{3-6r+r^2}{z}\right)(w^{s-2} - y^{s-2})\right] \end{aligned} \tag{4}$$

where $z = \sqrt{(1-r)(9-r)}$, $w = (1-r+z)/4$ and $y = (1-r-z)/4$. Note that the frequencies of recombinant haplotypes in males and females are $1 - 2m_s$ and $1 - 2f_s$, respectively, and that the overall frequency is $1 - (2m_s + 4f_s)/3$.

Now I turn to the unbalanced case for the X chromosome, in which all F_1 individuals are derived from the cross female $A \times$ male B, so that all F_1 males are hemizygous A. This appears to be widely used in practice (e.g., NORGARD *et al.* 2008; KELLY *et al.* 2010). The calculations are more difficult, because the allele frequencies are different in males and females and across generations.

I first calculate the single-locus allele frequencies. Let q_s be the frequency of the A allele in females at generation F_s. Note that the frequency in males at F_s is q_{s-1}. The initial values are $q_0 = 1$ and $q_1 = \frac{1}{2}$, and we have the recurrence relation $q_{s+1} = \frac{1}{2}q_s + \frac{1}{2}q_{s-1}$, which comes from the fact that a random allele drawn from the female at generation F_{s+1} is equally likely to be an allele from the female or male at generation F_s, and the allele in the male at F_s is a random allele from the female at F_{s-1}. The solution of the recurrence relation is $q_s = \frac{2}{3} + (\frac{1}{3})(-\frac{1}{2})^s$, for $s \geq 0$.

I now turn to the two-locus haplotype probabilities. Let m'_s and f'_s denote the frequencies of the AA haplotype on the X chromosome in males and females at generation F_s in an unbalanced AIL, and note that $m'_1 = 1$ and $f'_1 = \frac{1}{2}$. The haplotype probabilities satisfy a recurrence relation similar to that in equation (3):

$$\begin{aligned} m'_{s+1} &= (1-r)f'_s + rq_{s-1}q_{s-2} \\ f'_{s+1} &= \left(\tfrac{1}{2}\right)m'_s + \left(\tfrac{1-r}{2}\right)f'_s + \left(\tfrac{r}{2}\right)q_{s-1}q_{s-2} \end{aligned} \tag{5}$$

Note the distinction between equations (3) and (5): if a recombinant haplotype is transmitted from the F_s female, the chance that it brings two A alleles together depends on the frequency of the A allele in males and females in the F_{s-1} generation. In the balanced case, these are each $\frac{1}{2}$; in the unbalanced case, they are different from each other and vary across generations.

I have been unable to obtain closed-form solutions for m'_s and f'_s. However, the values can be quickly calculated numerically, using equation (5). Note that $\lim_{s\to\infty} f'_s = \lim_{s\to\infty} m'_s = \frac{4}{9}$.

Haplotype probabilities in the DO are calculated similarly. The progenitors for the DO

were pre-CC mice. I assume a large number of progenitors, that they were drawn from independent lines, and that the order of the crosses that generated the different lines were random, giving complete balance across the eight alleles.

In a potential abuse of notation, I will redefine the q's, p's, m's and f's used above. Let q_k denote the frequency of the AA haplotype at generation $\mathrm{G}_2 : \mathrm{F}_k$ in the pre-CC; this is $\frac{1-r}{2}$ times the haplotype probability in Table 4 of BROMAN (2012). Let p_s be the probability of the AA haplotype at generation s of the diversity outcross.

The pre-CC progenitors of the DO were drawn from independent lines at a variety of different generations along the course to inbreeding. Let α_k denote the proportion of the pre-CC progenitors that were at generation $\mathrm{G}_2 : \mathrm{F}_k$, and note that a pre-CC progenitor at generation $\mathrm{G}_2 : \mathrm{F}_k$ will transmit the AA haplotype with frequency q_{k+1} (that is, the frequency of the AA haplotype at generation $\mathrm{G}_2 : \mathrm{F}_k$). Thus, the frequency of the AA haplotype at the first generation of the DO is $p_1 = \sum_k \alpha_k q_{k+1}$.

The recurrence relation for the p_s is like that in equation (1): $p_{s+1} = (1-r)p_s + r/64$. The solution is

$$p_s = \tfrac{1}{64} + (1-r)^{s-1}\left(p_1 - \tfrac{1}{64}\right) \tag{6}$$

Note that the recombinant haplotypes are all equally likely, due to the random order of the initial crosses, and so each has probability $(1 - 8p_s)/56$.

HS corresponds to the DO with $\alpha_1 = 1$ (that is, $k \equiv 1$), in which case $p_1 = q_2 = 7 - 24r + 24r^2 - 8r^3$.

I now turn to the X chromosome. Let m_s and f_s denote the frequency of the AA haplotype on the X chromosome in males and females in the DO at generation s. Assuming random

orders of crosses to generate the pre-CC progenitors,

$$f_1 = \sum_k \alpha_k \left(\tfrac{1}{8}\right) \left[(2-r)h_{k+1}^{AA} + (1-r)h_{k+1}^{CC}\right] \tag{7}$$

where h_{k+1}^{AA} and h_{k+1}^{CC} are the frequencies of the AA and CC haplotypes, respectively, on the X chromosome in females at generation $G_1 : F_{k+1}$ in the construction of four-way RIL by sibling mating (see BROMAN 2012, Table 4). m_1 is calculated in the same way. The recurrence relations are much like equation (3):

$$\begin{aligned} m_{s+1} &= (1-r)f_s + \tfrac{r}{64} \\ f_{s+1} &= \left(\tfrac{1}{2}\right) m_s + \left(\tfrac{1-r}{2}\right) f_s + \tfrac{r}{128} \end{aligned} \tag{8}$$

The solutions are the following:

$$\begin{aligned} m_s &= \tfrac{1}{128}\left\{2 + \left[\tfrac{(64m_1-256f_1+3)(1-r)}{z}\right](y^{s-1} - w^{s-1}) - (1-64m_1)(w^{s-1}+y^{s-1})\right\} \\ f_s &= \tfrac{1}{128}\left\{2 + \left[\tfrac{-64f_1(1-r)-128m_1+3-r}{z}\right](y^{s-1} - w^{s-1}) - (1-64f_1)(w^{s-1}+y^{s-1})\right\} \end{aligned} \tag{9}$$

where w, y and z are as in equation (4).

Again, HS corresponds to DO with $\alpha_1 = 1$, in which case $f_1 = (4 - 5r + r^2)/32$ and $m_1 = (2 - 3r + r^2)/16$.

In Figure 1, the probabilities of recombinant two-locus haplotypes are displayed for the different populations. For the DO, I used the distribution of k as in Figure 1 of SVENSON *et al.* (2012) and $s = 5$. For HS and AIL, I used $s = 10$ and 12, respectively, to match the total number of generations with recombination (the average k in SVENSON *et al.* (2012) was 6).

Recombinant haplotypes are more frequent on the autosome, and are more frequent in HS than in the DO; inbreeding in the pre-CC progenitors of the DO is accompanied by a loss of recombinants.

It is particularly interesting to consider the map expansion in these populations, which is the frequency of recombination breakpoints on a random chromosome. Let R denote the probability of a recombinant haplotype; then the map expansion is $\frac{dR}{dr}\big|_{r=0}$ (see TEUSCHER and BROMAN 2007). The map expansion on an autosome in AIL is $s/2$. For the DO, on an autosome, the map expansion satisifies $M_s = \frac{7}{8}(s-1) + M_1$, where M_1 is the weighted average (with weights α_k) of the map expansion in the pre-CC at generation $G_2 : F_{k+1}$ (see BROMAN 2012, Table 4). For the particular progenitors detailed in SVENSON *et al.* (2012, Figure 1), this is approximately $(7s + 37)/8$. For HS, we have $M_1 = 3$ and $M_s = \frac{7s+17}{8}$.

For the X chromosome in balanced AIL, HS and DO, the map expansion is $\frac{2}{3}$ that of the autosome. For the case of the X chromosome in unbalanced AIL, in which all F_1 males are hemizygous A, I cannot derive a closed-form solution, but taking the derivatives of the recurrence relations in equation (5), I can derive a simple recurrence relation for the map expansion. (Note that the overall map expansion on the X chromosome can be obtained as the average of the sex-specific map expansions, with $\frac{2}{3}$ weight given to the female, since two-thirds of the X chromosomes are in females.) Let M'_s denote the map expansion at F_s, and again let q_s be the frequency of the A allele in females at F_s. Then we have

$$M'_{s+1} = M'_s + \tfrac{4}{3}(q_s - q_{s-1}q_{s-2}) \tag{10}$$

with the initial conditions $M'_1 = 0$ and $M'_2 = \frac{2}{3}$. While I have not been able to derive a

closed-form solution for M'_s, it is easily calculated numerically.

The haplotype probabilities calculated above provide the key quantities for developing HMMs for advanced intercross populations. However, it should be noted that there are other approaches to handling such data. For example, BESNIER *et al.* (2011) used a variance components model to analyze outbred chicken AIL data, with identity-by-descent (IBD) probabilities calculated using a modified version of the method of PONG-WONG *et al.* (2001), for general pedigree data.

The above result for HS differs from that in MOTT *et al.* (2000) and incorporated into the HAPPY software. They had assumed that the map expansion in HS was $\frac{7}{8}(s+2)$, while I show it to be $\frac{7}{8}(s-1)+3$. In the first three of generations with recombination, individuals are fully heterozygous, and so all recombination events can be seen; in the subsequent $s-1$ generations, there is a $1/8$ chance of homozygosity and so only $7/8$ of recombination events can be seen.

MOTT *et al.* (2000) further assumed that the transition probabilities along an HS chromosome are a function of genetic distance, but that requires knowledge of the map function. It is more direct to express the transition probabilities in terms of the recombination fraction at meiosis.

The green curve in Figure 1 displays the probability of a recombinant haplotype assumed in MOTT *et al.* (2000) for HS with $s = 10$, using the map function corresponding to the gamma model with the level of crossover interference estimated for the mouse in BROMAN *et al.* (2002). The probability is slightly smaller than that from our calculations; at $r = 0.01$, the equation in MOTT *et al.* (2000) gives 0.099, whereas I obtain 0.103.

I have assumed an effectively infinite number of mating pairs at each generation. In

practice, with a finite number of mating pairs, there will be some inbreeding and so an increased frequency of homozygosity and a decreased frequency of recombination. In addition, the individuals at the final generation will include siblings, and the relationships among individuals might be used to improve the genotype reconstruction. In practice, for computational efficiency, both the inbreeding and the relationships among individuals would probably be ignored in the genotype reconstruction, and with dense genotype data, there will be little loss of information.

Acknowledgments

Jim Crow generously provided comments for improvement of the manuscript. This work was supported in part by National Institutes of Health grant GM074244.

Literature Cited

AYLOR, D. L., W. VALDAR, W. FOULDS-MATHES, R. J. BUUS, R. A. VERDUGO *et al.*, 2011 Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. **21**: 1213–1222.

BESNIER, F., P. WAHLBERG, L. RÖNNEGÅRD, W. EK, L. ANDERSSON *et al.*, 2011 Fine mapping and replication of QTL in outbred chicken advanced intercross lines. Genet. Sel. Evol. **43**: 3.

BROMAN, K. W., 2012 Genotype probabilities at intermediate generations in the construction of recombinant inbred lines. Genetics **in press [CC-CROSS CITATION CC3]**.

BROMAN, K. W., L. B. ROWE, G. A. CHURCHILL and K. PAIGEN, 2002 Crossover interference in the mouse. Genetics **160**: 1123–1131.

BROMAN, K. W. and S. SEN, 2009 *A guide to QTL mapping with R/qtl*. Springer.

BULMER, M. G., 1980 *The mathematical theory of quantitative genetics*. Claredon Press.

DARVASI, A. and M. SOLLER, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141**: 1199–1207.

GRAHAM, R. L., D. E. KNUTH and O. PATASHNIK, 1994 *Concrete Mathematics*. Addison-Wesley, 2nd edition.

HALEY, C. S. and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69**: 315–324.

KELLY, S. A., D. L. NEHRENBERG, J. L. PEIRCE, K. HUA, B. M. STEFFY *et al.*, 2010 Genetic architecture of voluntary exercise in an advanced intercross line of mice. Physiol. Genomics **42**: 190–200.

LANDER, E. S. and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121**: 185–199.

MOTT, R., C. J. TALBOT, M. G. TURRI, A. C. COLLINS and J. FLINT, 2000 A method for fine mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. USA **97**: 12649–12654.

NORGARD, E. A., C. C. ROSEMAN, G. L. FAWCETT, M. PAVLIC, C. D. MORGAN *et al.*, 2008 identification of quantitative trait loci affecting murine long bone length in a two-generation intercross of LG/J and SM/J mice. J. Bone Miner. Res. **23**: 887–895.

PONG-WONG, R., A. W. GEORGE, J. A. WOOLLIAMS and C. S. HALEY, 2001 A simple and rapid method for calculating identity-by-descent matrices using multiple markers. Genet. Sel. Evol. **33**: 453–471.

SVENSON, K. L., D. M. GATTI, W. VALDAR, C. E. WELSH, R. CHENG *et al.*, 2012 The mouse diversity outcross population. Genetics **in press [CC-CROSS CITATION CC12]**.

TEUSCHER, F. and K. W. BROMAN, 2007 Haplotype probabilities for multiple-strain recombinant inbred lines. Genetics **175**: 1267–1274.

FIGURE LEGENDS

Figure 1. Frequency of a two-locus haplotype being recombinant, as a function of the recombination fraction at meiosis, for the diversity outcross population at $s = 5$ (solid curves), heterogeneous stock at $s = 10$ (dashed curves) and balanced AIL at $s = 12$ (dotted curves), for the autosome (black), male X (blue) and female X (red). The green dashed curve is the recombinant frequency for HS at $s = 10$ assumed in MOTT *et al.* (2000).

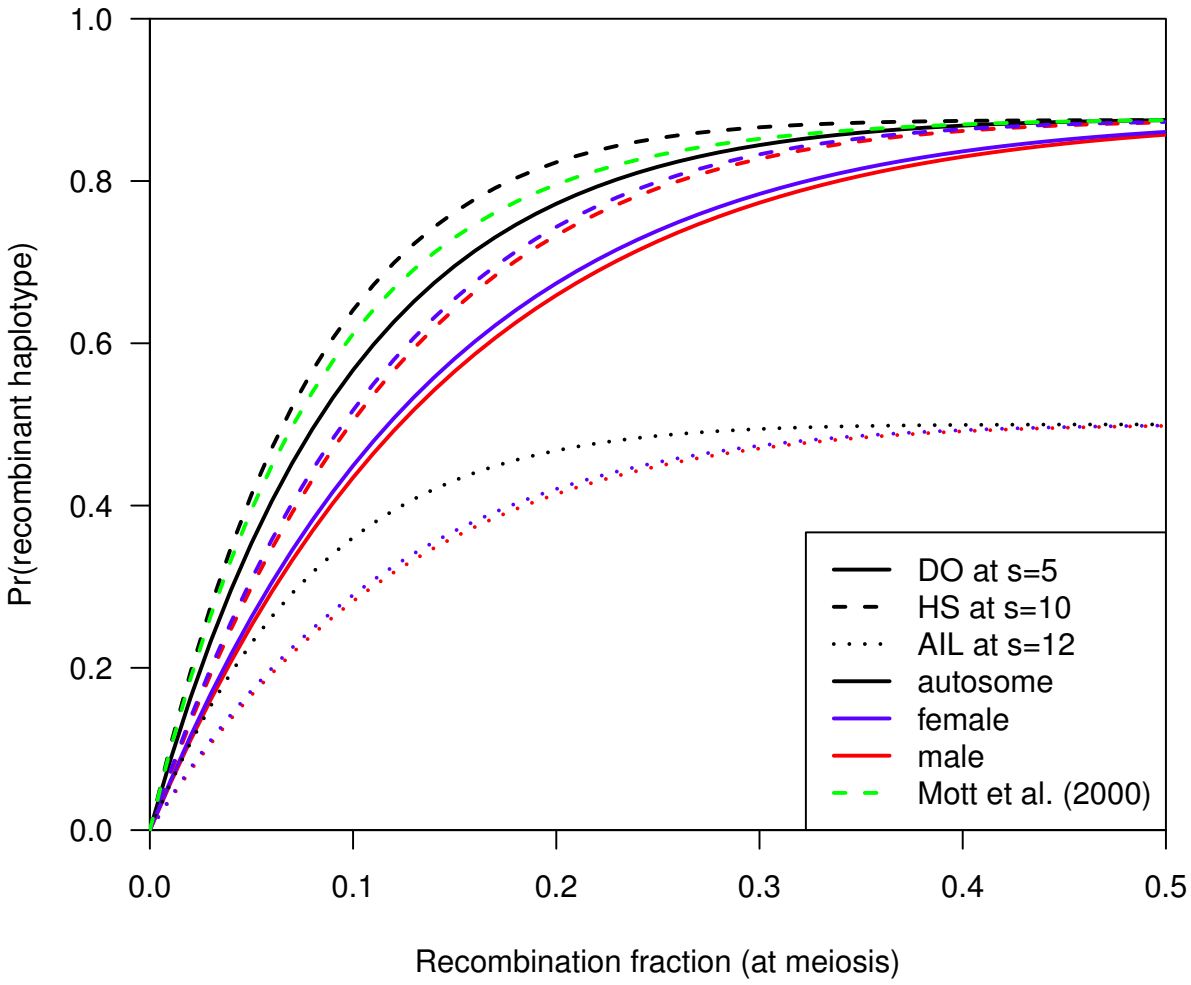

Figure 1: Frequency of a two-locus haplotype being recombinant, as a function of the recombination fraction at meiosis, for the diversity outcross population at $s = 5$ (solid curves), heterogeneous stock at $s = 10$ (dashed curves) and balanced AIL at $s = 12$ (dotted curves), for the autosome (black), male X (blue) and female X (red). The green dashed curve is the recombinant frequency for HS at $s = 10$ assumed in MOTT *et al.* (2000).