

Stat 246: Statistical Genetics

Week 4, Lecture 2

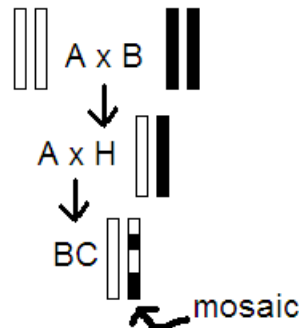
Instructor: Terry Speed
Scribe: Greg Hather

February 13, 2006

1 Breeding Experiments

An inbred strain is any strain of animal or plant obtained by a breeding strategy that leads to homozygosity across the whole genome. An inbred strain could be created by brother sister matings. For example, we could catch a male mouse and a female mouse from the wild, let them mate, and then select one one of the male offspring and one of the female offspring to be mated together. After 10 or more generations, we would have a litter of mice that were nearly genetically identical.

Now, suppose that we had two different inbred strains of mice, A and B. If we were to mate an A mouse with a B mouse, their offspring would all be heterozygous at loci at which A and B differ. The offspring in such an experiment are sometimes denoted by F1 (first filial generation), but in these notes we shall denote them by H (heterozygotes). All of the H offspring should be (nearly) genetically identical. Now, if we took an H mouse and mated it back to an A mouse, the resulting offspring are described as BC (backcrossed). The BC mice would have one set of chromosomes that were identical to the A mice. However, the BC mice would have another set of chromosomes that were mosaics of strain A and strain B chromosomes. Also, each BC mouse would be genetically different from its siblings.



Now suppose that the B mouse had a certain trait, T, that was of interest to us¹. Suppose that through previous experiments, we had determined that T was caused by a single dominant allele of

¹One might obtain a mouse with a trait of interest via mutagenesis. For example, we could start with an ordinary male B mouse and irradiate the mouse's testicles. One could then mate the male B mouse to a female B mouse and screen the offspring for traits of interest. After identifying a mouse with a trait of interest, one could use backcrossing

unknown chromosomal location. Suppose mice exhibit trait T if and only if they have that specific allele². Let's denote the allele's location on the chromosome by *, and let's denote the allele by b_* . Suppose that the B mouse was homozygous for b_* and that we were interested in determining *. As described in the following section, we can estimate * by studying the genotypes and phenotypes of the BC mice.

2 Genetic Mapping

Genetic mapping is the determination of the relative positions of genes on a DNA molecule and of the genetic or physical distances between them.

2.1 Single Locus Case

Consider a marker locus l at which A and B mice differ. Let H_l be the event that a given mouse is heterozygous at locus l . Let A_l be the event that a given mouse is homozygous at locus l . Let H_* and A_* be defined in a similar manner. Form the following table from n BC individuals. Let r be the recombination fraction between l and *. Then

$$Pr(A_l \& A_*) = Pr(A_l)Pr(A_*|A_l) = \frac{1}{2}(1-r). \quad (1)$$

In this way, we can compute the expected fraction of offspring in each quadrant of the table. Then, we could estimate r using

$$\hat{r} = \operatorname{argmax}_{0 \leq r \leq \frac{1}{2}} L(r), \quad (2)$$

where

$$L(r) \propto \left(\frac{1}{2}r\right)^{n_2+n_3} \left(\frac{1}{2}(1-r)\right)^{n_1+n_4}. \quad (3)$$

We could convert this estimate for r into a distance if desired. Sometimes, we wish to test $H_0 : r = \frac{1}{2}$ vs $K : r < \frac{1}{2}$. In genetics, researchers often use the LOD score, which is defined by

$$LOD = \log_{10} \left(\frac{L(\hat{r})}{L(\frac{1}{2})} \right) \quad (4)$$

and reject H_0 if $LOD \geq 3$.

2.2 Multiple Locus Case

Assume a dense set of mapped markers at loci $1, 2, \dots, \mathcal{L}$ at which A and B mice differ. Suppose the nearest marker to the left of * is L and the nearest marker to the right of * is R . L and R can be identified as the loci with the top two LOD scores. To estimate *, we can use data on markers at L and R only, but this will usually involve ignoring individuals with missing data. The genotype at L may be missing, the genotype at R may be missing, the phenotype may be missing, or some combination thereof.

or brother sister mating to confirm that the trait was associated with a single locus and to obtain mice that were homozygous for the associated allele.

²In general, $Pr(T|\text{have allele}) = p_1$ and $Pr(\bar{T}|\text{don't have allele}) = p_2$. p_1 and p_2 are called the penetrances.

	\bar{T}	T
A_L	n_1	n_2
H_L	n_3	n_4

	\bar{T}	T
A_L	$.5(1-r)$	$.5r$
H_L	$.5r$	$.5(1-r)$

	A_R	H_R	$-$
A_L			
H_L			
$-$			
	\bar{T}		

	A_R	H_R	$-$
A_L			
H_L			
$-$			
	T		

	A_R	H_R	$-$
A_L			
H_L			
$-$			
	$-$		

Exercise: Estimate $*$ using the complete data only. Assume the Poisson model.

We want to calculate

$$\hat{*} = \operatorname{argmax}_* L(\text{locus is } * | \text{data}), \quad (5)$$

where

$$L(\text{locus is } * | \text{data}) = \prod_{\text{BC mice}} \operatorname{Pr}(\text{data for BC mice} | \text{locus is } *). \quad (6)$$

Note that $*$, as a parameter, is allowed to vary continuously over the entire chromosome. Our problem boils down to calculating $\operatorname{Pr}(\text{one mouse's data} | \text{trait locus at } *)$. Let G_i be a variable which equals 1 when the mouse is heterozygous at locus i and 0 otherwise. Until further notice, we shall assume the Poisson model. Thus, if every locus is scored, and if there is no missing data, we have

$$\begin{aligned} & \operatorname{Pr}(G_1, G_2, G_3, \dots, G_L, G_*, G_R, \dots, G_{\mathcal{L}}, T | \text{trait locus is } *) \\ &= \operatorname{Pr}(G_1) \operatorname{Pr}(G_2 | G_1) \operatorname{Pr}(G_3 | G_2) \dots \operatorname{Pr}(G_* | G_L) \operatorname{Pr}(G_R | G_*) \dots \operatorname{Pr}(G_{\mathcal{L}} | G_{\mathcal{L}-1}) \\ &= \frac{1}{2} \prod_{i \in \{1, \dots, L-1\}} (1 - r_{i, i+1})^{I(G_i = G_{i+1})} r_{i, i+1}^{I(G_i \neq G_{i+1})} \\ & \quad (1 - r_{L, *})^{I(G_L = G_*)} r_{L, *}^{I(G_L \neq G_*)} (1 - r_{*, R})^{I(G_* = G_R)} r_{*, R}^{I(G_* \neq G_R)} \\ & \quad \prod_{i \in \{R, \dots, \mathcal{L}\}} (1 - r_{i, i+1})^{I(G_i = G_{i+1})} r_{i, i+1}^{I(G_i \neq G_{i+1})} \end{aligned}$$

Apparently we have a Markov Chain with transition matrix

$$\begin{array}{c} A_i \\ A_{i+1} \\ H_{i+1} \end{array} \begin{array}{cc} & H_i \\ \left[\begin{array}{cc} 1 - r_{i,i+1} & r_{i,i+1} \\ r_{i,i+1} & 1 - r_{i,i+1} \end{array} \right] \end{array}$$

2.3 HMMs in Genetics

The challenge is dealing with incomplete data, and this is where HMMs (Hidden Markov Models) can help.

3 A Reference on Mouse Genetics

Lee M. Silver's book *Mouse Genetics: Concepts and Applications* is good reference. It available online at <http://www.informatics.jax.org/silver/>