Anton Amirov
Homework 8

1. In the notes, we showed that the probability that two half-sisters share a gene ibd at two loci at recombination frequency r is R/2 where $R = r^2 + (1-r)^2$, but that the probability for an aunt-niece pair is ((1-r)R/2 + r/4).
(a) Show that under the Haldane map function $R=(1+e^{-4d})/2$.

With the Haldane map function

$$r = \frac{1 - e^{-2d}}{2}$$

and

$$R = \left(\frac{1 - e^{-2d}}{2}\right)^2 + \left(\frac{1 + e^{-2d}}{2}\right)^2 = \frac{1 + e^{-4d} - 2e^{-2d}}{4} + \frac{1 + e^{-4d} + 2e^{-2d}}{4} = \frac{1 + e^{-4d}}{2}$$

(b) Find the probability that an aunt niece pair share a gene ibd at both of two loci at recombination frequency r in terms of genetic distance d.

$$\frac{(1-r)R}{2} + \frac{r}{4} = \frac{(1-r)(1+e^{-4d}) + r}{4} = \frac{1 + e^{-4d} - re^{-4d}}{4} = \frac{2 + 2e^{-4d} - e^{-4d} + e^{-6d}}{8}$$
$$= \frac{1}{4} + \frac{1 + e^{-2d}}{8} e^{-4d}$$

(c) Find the probabilities of sharing 2, 1, and 0 ibd at both of two linked loci for a pair of non-inbred full sibs, in terms of R.

Sharing two genes IBD for full sibs is the square of probability of sharing 1 gene ibd between half-sibs, since paternal and maternal meiosis are independent

$$P_2(r) = (K11(r))^2 = \frac{R^2}{4}$$

Probability of sharing 1 gene IBD at both loci will be

$$P_1(r) = 2\big(1 - K11(r)\big)K11(r) = \frac{R(2-R)}{2}$$

And sharing none

$$P_0(r) = \big(1 - K11(r)\big)^2 = \frac{(2-R)^2}{4}$$

2. Three loci are known to be ordered, A, B, C, along a chromosome. Let r_1 denote the recombination frequency between A and B, r_2 between B and C, and r between A and C. Call the gametes that are recombinant between A and B ``AB recombinants''.
(a) Show that, in the absence of chromatid interference, r is at least the max of r_1 and r_2. (You may quote Mather's formula without proof.) Show that, in the absence of genetic interference, $r = r_1 (1-r_2) + r_2 (1-r_1)$.

According to Mather's formula r1 is ½ of probability to have crossovers between A and B, r2 is ½ of crossover probability between B and C, and r is ½ of crossover probability between A and C. It's easy to see that crossover probability between A and C is at least as high as either A and B or B and C (as any meiosis that has crossover between, say, A and B also has meiosis between A and C). Therefore,

$$r \geq \max(r_1, r_2)$$

From

$$P(no\ crossover\ between\ A\ and\ C) = P(no\ crossover\ between\ A\ and\ B)P(no\ crossover\ between\ B\ and\ C)$$

we can derive

$$P(crossover\ between\ A\ and\ C) = 1 - (1 - P(crossover\ between\ A\ and\ B)(1 - P(crossover\ between\ B\ and\ C))$$

or, using Mather's formula,

$$2r = 1 - (1 - 2r_1)(1 - 2r_2)$$

and

$$r = \frac{1 - (1 - 2r_1)(1 - 2r_2)}{2} = \frac{2r_1 + 2r_2 - 4r_1 r_2}{4} = r_1 + r_2 - 2r_1 r_2 = r_1(1 - r_2) + r_2(1 - r_1)$$

(b) Three independent studies are made of linkage between a pair of loci. The first results in x_1 recombinants, and $n_1 - x_1$ non-recombinants between A and B. The second results in x_2 recombinants and n_2-x_2 non-recombinants between B and C. The third results in x_3 recombinants and n_3-x_3 non-recombinants between A and C. Suppose it happens that x_i/n_i (i=1,2,3) are all less than 1/2. Making no assumptions about interference, what are the MLEs of r_1, r_2, and r?

Without assumptions on interference we cannot relate r to r_1 and r_2, so all estimates will be done independently as simple frequencies, giving

$$r_1 = \frac{x_1}{n_1} \quad r_2 = \frac{x_2}{n_2} \quad r = \frac{x_3}{n_3}$$

(c) Show that, in the absence of genetic interference, the expected number of AC recombinants that are recombinant in AB but not in BC is $x_3\ r_1\ (1\text{-}r_2)/\ (r_1\ (1\text{-}r_2) + r_2\ (1\text{-}r_1))$, and that the expected number of AC non-recombinants that are recombinant both in AB and in BC is $(n_3\ \text{-}x_3)\ r_1\ r_2\ /\ (r_1\ r_2 + (1\text{-}r_1\ )(1\text{-}r_2))$.

Someone is recombinant in AC if he is either recombinant in AB and not in BC (unconditional probability $r_1(1-r_2)$) or in BC and not in AB (unconditional probability $r_2(1-r_1)$). Conditional probability to be recombinant in AB and not in BC if someone is recombinant in AC by Bayes' theorem is

$$\frac{r_1(1-r_2)}{r_1(1-r_2)+\ r_2(1-r_1)}$$

and given that there are x_3 AC recombinants, expected number of AB/not-BC recombinants is

$$x_3\frac{r_1(1-r_2)}{r_1(1-r_2)+\ r_2(1-r_1)}$$

Someone can be non-recombinant in AC if he is either non-recombinant in both AB and BC (unconditional probability $(1-r_1)(1-r_2)$) or recombinant in both (unconditional probability r_1r_2). Conditional probability for someone who is non-recombinant in AC to be recombinant in AB and BC is

$$\frac{r_1r_2}{r_1r_2+(1-r_1)(1-r_2)}$$

with expectation, give that there are n_3-x_3 non-recombinants between A and C

$$(n_3-x_3)\frac{r_1r_2}{r_1r_2+(1-r_1)(1-r_2)}$$

Describe an EM algorithm which will provide the MLEs of r, r_1 and r_2 in the absence of genetic interference.

1. On E step using approximations of r_1 and r_2 we will find number of AB recombination among known tests. We have n_1+n_3 tests and within them expected number of AB recombinants is (know recombinants from first test + number of AB/non-BC recombination and AB/BC recombinants from above)

$$E_1 = x_1 + x_3\frac{r_1(1-r_2)}{r_1(1-r_2)+\ r_2(1-r_1)} + (n_3-x_3)\frac{r_1r_2}{r_1r_2+(1-r_1)(1-r_2)}$$

For BC recombinants, using the same logic

$$E_2 = x_2 + x_3 \frac{r_2(1-r_1)}{r_1(1-r_2) + \ r_2(1-r_1)} + (n_3 - x_3) \frac{r_1 r_2}{r_1 r_2 + (1-r_1)(1-r_2)}$$

2. On E step we just build estimation of r1 and r2 as frequencies

$$r_1 = \frac{E_1}{n_1 + n_3}$$

$$r_2 = \frac{E_2}{n_2 + n_3}$$

and calculate r as function of r_1 and r_2

$$r = r_1(1-r_2) + r_2(1-r_1)$$

(d) How would you test the absence of genetic interference, given the above data?

If genetic interference is absent, after stabilization of EM algorithm we should have x_3 distributed as $B(n_3,r)$, x_1 as $B(n_1,r_1)$ and x_2 as $B(n_2,r_2)$, so we can the hypothesis that x is distributed accordingly – i.e. that x_i fall within confidence interval of $n_i r_i$

3. A second investigator has a different design for estimating recombination frequencies between the same three loci, A, B, C of the previous question. She types m= (2/3)(n_1+n_2+n_3) gametes at all three loci. (So she has done the same total number of typings as in Qu. 2.) She finds y_{11} that are both AB and BC recombinants, y_{10} that are AB recombinants and BC non-recombinants, y_{01} that are AB non-recombinants and BC recombinants, and y_{00}= m - y_{11} -y_{10} -y_{01} that are non-recombinant in both intervals.
(a) How should this investigator estimate the recombination frequencies r_1 and r_2 ?

She should just use frequencies

$$r_1 = \frac{y_{10} + y_{11}}{m}$$

$$r_2 = \frac{y_{01} + y_{11}}{m}$$

(b) How would you advise this investigator to test the absence of genetic interference?

She can use some statistical test for absence of relationship between events, for example 2x2 contingency table

(c) Which design (Qu.2 or Qu. 3) do you expect to provide more information for detecting genetic interference? Very briefly, how would you measure this information in order to make this comparison?

I would expect second design to provide more evidence about interference. It obviously would provide more evidence if the same amount of gametes were typed, since each gamete contains more information than in the 1^{st} design. However since the number of gametes is reduced to have the same amount of typings, it is difficult to have a definite answer without quantitative analysis, result of which can depend on n_1, n_2, n_3 and r_1, r_2 (for example if $n_1=n_2=n_3$ second design is providing the same information as the first one if we type the same group each time).

One way to conduct such analysis is to have simulations with some values of n and r, as well as with presence and absence of genetic interference between AB and BC (with some value of interference coefficient). In each scenario confidence intervals can be optimized to achieve some desired measure of accuracy (for example amount of errors in detection of presence of interference), and we can decide which approach is better by looking at the resulting optimal value of that measure.