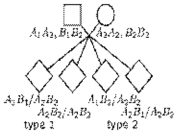
3.3 Linkage Designs and Information
3.3.1 Phase unknown backcross

- In human pedigrees, we often cannot classify individuals as recombinant and non-recombinant.
- One possibility is a *phase-unknown backcross*.
- As before, one parent is A1A2,B1B2 and the other is A2A2,B2B2, but now the first parent may be A1B1/A2B2 (type 1 haplotypes), or A1B2/A2B1 (type 2 haplotypes).
- Suppose we have *n* such families, and in each type just two offspring. Each gets A2B2 from the mother, so, as before, we know what each got from the father.
- If both offspring get the same ``type" of haplotype (type 1 or type 2), then either both are recombinant, or neither is, so this event has probability ρ^2 + (1- ρ)^2.
- Or there is one of each: then one offspring must be a recombinant and the other not. This event has probability ρ* = 2 ρ (1 - ρ).



Phase unknown backcross: analysis

- Instead of a T ~ B(*n*, ρ) recombinants, we have a W ~ B(*n*, ρ*) families.
- For 0 ≤ ρ ≤ 1/2, ρ* is a 1-1 monotone increasing function of ρ, and when ρ = 1/2, ρ = 2x (1/2) x (1/2) = 1/2.
- So testing H0: ρ = 1/2 against H1: ρ < 1/2, is equivalent to testing H0*: ρ* = 1/2 against H1*: ρ* < 1/2.
- Thus the test is as before; reject ρ* = 1/2 and infer linkage if W < w0, where the critical value w0 is determined by the desired size (type 1 error) of the test.
- The critical values are exactly as for the phase-known case, with ρ* replacing ρ, and *n* now denoting the number of two-child families.
- Of course, the tests properties are different. When ρ=0.3, for example, ρ*=2x0.3x0.7= 0.42, which is closer to 1/2. It will be correspondingly harder to detect linkage.
- We return to this in section 3.4.

3.3.2 INTERCROSS EXPERIMENT

- Another classic design for experimental organisms is the *intercross*.
- Two phase-known parents, each of type A1B1/A2B2 are mated. There are nine types of offspring, but these fall into four groups.
- Each type within a group has the same probability, as a function of ρ, and hence the total count of offspring in each group contains all the available information for linkage.
- These total counts are the sufficient statistics for ρ.

Type	genotypes	number	each prob.
I	A1A1,B2B2; A2A2,B1B1	2	ρ^2/4
II	A1A2,B1B2	1	(ρ^2 + (1-ρ)^2)/2
III	A1A1,B1B2 etc.	4	ρ (1- ρ)/2
IV	A1A1,B1B1; A2A2,B2B2	2	(1 – ρ)^2 /4

3.3.3 INTERCROSS EXPERIMENT
Analysis: the type probabilities

- Group II includes both double-heterozygote two-locus genotypes A1B1/A2B2 and A1B2/A2B1.
- Group III includes the four types heterozygous at one of the two loci: A1A1,B1B2; A1A2,B1B1; A2A2,B1B2 and A1A2,B2B2.
- The following table gives the type probabilities under alternative hypotheses:

Types	H2:general	H1: total prob	H0:ρ=1/2
I	q1	ρ^2 /2	0.125
II	q2	(ρ^2 + (1- ρ)^2)/2	0.25
III	q3	2 ρ (1 – ρ)	0.5
IV	q4	(1- ρ^)*2 /2	0.125

INTERCROSS EXPERIMENT
Analysis: testing fit.

- Consider a sample of size *n*, with nj in class j, j=1,2,3,4.
- The log-likelihood for these multinomial data is,
  . λ(**q**) = const + sum_{j=1}^4 nj log qj(ρ).
- The probabilities of each phenotype group are shown, under the general multinomial model H2, the general intercross linkage model H1, and in the absence of linkage H0.
- For example, suppose ***n*** = (1, 72, 42, 85).
- Under H2: general q_j, q1+q2+q3+q4 = 1. MLE qj* = nj/n, or **q*** = (0.005, 0.36, 0.21, 0.425). dim(H2) = 3.
- Under H1: general ρ, for these data we find, by evaluating the log-likelihood, that ρ* = 0.12 giving **q***(ρ) = (0.007, 0.394, 0.211, 0.387). dim(H1) = 1.
- The null hypothesis is of no linkage; H0: ρ = 1/2.
- **q***(1/2)= (0.125, 0.25, 0.5, 0.125) and dim(H0) = 0.
- Estimated cell probabilities under H1 and H2 are in good agreement, but quite different from those under H0.

Intercross experiment: testing hypotheses

- Computing the maximized log-likelihoods for Hi, i=0,1,2, we find that they are -307.76, -217.87, and -217.14 respectively.
- For testing null H0 against H1, the (base e) lod score is 89.9. Twice this value (179.8) has approximately a χ^2_1 if H0 is true. So H0 is rejected.
- For testing null H1 against alternative H2, the lod score is 0.73, and twice this value (1.46) is χ^2_2 if H1 is true. So H1 is not rejected.

- As with the phase-known backcross, this all extends to the estimation and testing of two recombination frequencies ρ_m in males, and ρ_f in females.
- Although for the intercross experiment, each offspring gives us a male and a female meiosis, we generally will not know which one is recombinant.
- The probabilities of the Table of 3.3.3 now depend on ρ_m and ρ_f. For example the first is (ρ_m ρ_f)/2.
- A likelihood ratio test may be derived in a similar way to Example 3 of the phase known backcross (3.2.4), to test equality of male and female recombination frequencies.
- However the MLEs of ρ_m and ρ_f are now harder to find.

3.4 POWER and INFORMATION
3.4.1 POWER and SAMPLE SIZE

- If ρ is the true value, the probability a null hypothesis H_0 is rejected is the power function of the test.
- For example, using the Normal approximation for a phase-known backcross (or any example where we count recombinants), the power is
  . P(T < t0 ; ρ) = P ((T - n ρ)/√(n ρ (1- ρ)) < (t0 - n ρ)/√(n ρ (1- ρ)))
  . ≈ Φ((t0 - n ρ)/√(n ρ (1- ρ)) .
- But now (from 3.2.3), t0 = (n/2) + (√n/2)Φ^{-1}(α) , so
  . P(T < t0 ; ρ) ≈ Φ((Φ^{-1}(α) + √n (1-2 ρ))/ (2√(ρ (1- ρ)).
- Note when ρ =1/2 this is equal to α, the test type-1 error.
- It decreases over 0 ≤ ρ ≤ 1/2. Clearly, for a given sample size, linkage is more easily detected when ρ is small. i.e. the power is larger.
- Conversely, for given ρ , one may determine the sample size n required for given power.
- For the phase-unknown backcross. the power and sample-size computations are exactly as for the phase-known case, with ρ*=2 ρ (1- ρ) replacing ρ, and n now denoting the number of two-child families.

3.4.2 Kullback-Leibler information

- The Kullback-Leibler (KL) information is a log-likelihood based measure appropriate for testing hypotheses (as opposed to Fisher Information which concerns estimation.
- For multinomial data in general, we can find the form of the KL information. Suppose there are c categories and suppose **q** is the true value of the cell probabilities qi, i=1,... ,c, and **q**0 is some hypothesized value.
- Then λ(**q**) = sum_{i=1}^c n_j log q_j. where here we use base-e logs.
- So for a sample size n, K_n (**q**0; **q**) = Exp_**q**(λ(**q**) – λ(**q**0))
  = n sum_{j=1}^c qi (log qi – log (q0i)) = n sum_{i=1}^c qi log (qi/q0).
- For a single observation, K = K_1 (**q**0; **q**) = sum_{j=1}^c qi log (qi / q0i).
- In the case of linkage analysis data, qi = qi(ρ) and the null hypothesis is H0: ρ = 1/2: q0i= qi(1/2).
- Evaluating the KL information for testing ρ=1/2, for the binomial (c=2) phase-known and (2-offspring) phase-unknown backcross experiments, and for the intercross experiment (c=4) we obtain the values for the information per offspring sampled shown on the next page.

KL Information in linkage designs

True ρ	0.0	0.1	0.2	0.3	0.4	0.5
Backcross: phase known	0.69	0.368	0.193	0.082	0.021	0
Backcross: phase unknown	0.35	0.111	0.033	0.006	0.0004	0
Intercross	1.04	0.479	0.226	0.089	0.021	0

- This measures information, per offspring sampled, for detecting linkage when ρ is the true value. As expected, the more ρ differs from 1/2 the more information there is.
- Also each phase-known offspring contributes at least twice as much as in the phase-unknown case. When ρ is close to 1/2, the phase-unknown two-offspring design provides very little information.
- As expected, each intercross offspring contains more information than a backcross offspring. But there is not twice as much information, as there would be if the meioses were fully observable.
- As ρ → 1/2, there is almost no additional information in doing an intercross design rather than a backcross.

3.4.3 Elods and sample size

- The Kullback-Leibler information for testing ρ = 1/2 is the expected base-e lod score at the true value of the recombination frequency ρ.
- This, but base-10, is a measure very widely used in linkage analysis and known as the Elod.
- Note we expect the base-e lod score to be approximately nK when n is large. For our intercross data with n=200, we had ρ*= 0.12; in fact, the data were simulated at ρ = 0.1. Then 200x 0.479 is about 95, in good agreement with the (base-e) lod score value of 90 which we obtained (last page of 3.3.3).
- This also tells us that if we had realized that ρ might be around 0.1, it was very wasteful to breed 200 mice. When ρ = 0.1, about 20 mice are expected to give a lod score (base e) of more than 9; this is plenty to detect that ρ ≠ 1/2.
- Note again that we have used natural logarithms in these examples, contrary to standard practice in genetics.