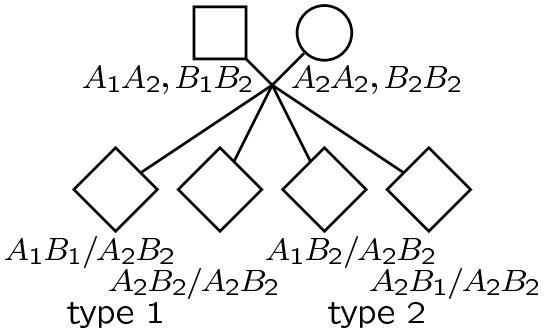
3.3.1 THE PHASE-UNKNOWN BACKCROSS DESIGN:

• In human pedigrees, we often cannot classify all individuals as recombinant and non-recombinant.



• One possibility is a *phase-unknown backcross*. As before, one parent is A_1A_2, B_1B_2 and the other is A_2A_2, B_2B_2, but now we do not know whether the first parent is A_1B_1/A_2B_2 (type 1 haplotypes), or A_1B_2/A_2B_1 (type 2 haplotypes).

• Either "type-1" is recombinant, and "type-2" are not, or
"type-1" is non-recombinant, and "type-2" are recombinant.

• Suppose we have n such families, and in each type just two offspring. Each gets A_2B_2 from the father, so, as before, we know what each got from the mother.

3.3.2 TESTING FOR LINKAGE IN FAMILIES SIZE 2:

• 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 $\rho^2 + (1-\rho)^2$.

• If there is one of each type, then one offspring must be a recombinant and the other not. This event has probability $\rho^* = 2\rho(1-\rho)$.

• Instead of a $T \sim B(n, \rho)$ recombinants, we have a $W \sim B(n, \rho^*)$ families with one offspring of each "type".

• For $0 \le \rho \le \frac{1}{2}$, ρ^* is a 1-1 monotone increasing function of ρ.
Also, when $\rho = \frac{1}{2}$, $\rho^* = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$.
So testing $H_0 : \rho = \frac{1}{2}$ against $H_1 : \rho < \frac{1}{2}$, is equivalent
to testing $H_0^* : \rho^* = \frac{1}{2}$ against $H_1^* : \rho^* < \frac{1}{2}$.

• The test is as before; reject $\rho^* = \frac{1}{2}$ and infer linkage if $W < w_0$, where the critical value w_0 is determined by the desired type-1 error.

• 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.

3.3.3 THE INTERCROSS LINKAGE DEISGN:

• Two phase-known hybrid parents, each of type A_1B_1/A_2B_2 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	$A_1A_1, B_2B_2;\ A_2A_2, B_1B_1$	2	$\rho^2/4$
II	A_1A_2, B_1B_2	1	$\frac{1}{2}(\rho^2 + (1-\rho)^2)$
III	A_1A_1, B_1B_2 etc.	4	$\frac{1}{2}\rho(1-\rho)$
IV	$A_1A_1, B_1B_1;\ A_2A_2, B_2B_2$	2	$(1-\rho)^2/4$

• Group II includes both double-heterozygote two-locus genotypes A_1B_1/A_2B_2 and A_1B_2/A_2B_1. Group III includes the four types heterozygous at one of the two loci: A_1A_1, B_1B_2, A_1A_2, B_1B_1, A_2A_2, B_1B_2 and A_1A_2, B_2B_2.

3.3.4 INTERCROSS EXPERIMENT–TESTING FIT:

Types	H_2 : general	H_1:total prob	H_0 : $\rho = \frac{1}{2}$
I	q_1	$\frac{1}{2}\rho^2$	0.125
II	q_2	$\frac{1}{2}(\rho^2 + (1-\rho)^2)$	0.25
III	q_3	$2\rho(1-\rho)$	0.5
IV	q_4	$\frac{1}{2}(1-\rho)^2$	0.125

• Consider a sample of size n, with n_j in class j, $j = 1, 2, 3, 4$. The log-likelihood for these multinomial data is,

$$\ell_n(\mathbf{q}) = \text{const} + \sum_{j=1}^{4} n_j \log_e q_j(\rho).$$

• The probabilities of each phenotype group are shown, under the general multinomial model H_2, the general linkage model H_1, and in the absence of linkage H_0.

• For example, suppose $\mathbf{n} = (1,\ 72,\ 42,\ 85)$.
Under H_2 : $general\ q_j,\ \sum_{j=1}^{4} q_j = 1, \widehat{q_j} = n_j/n$,
or $\widehat{\mathbf{q}}$ = (0.005, 0.36, 0.21, 0.425). dim$(H_2) = 3$.

Testing the model, and testing for linkage; ctd:

• Under H_1 : $general$ ρ, for these data we find, by evaluating the log-likelihood, that $\widehat{\rho} = 0.12$ giving $\mathbf{q}(\widehat{\rho})$ = (0.007, 0.394, 0.211, 0.387). dim$(H_1) = 1$.

• The null hypothesis is of no linkage; H_0 : $\rho = \frac{1}{2}$.
dim$(H_0) = 0$, with probs $\mathbf{q}(\frac{1}{2}) = (0.125, 0.25, 0.5, 0.125)$.

• Estimated cell probabilities under H_1 and H_2 are in good agreement, but quite different from those under H_0.

• Computing the maximized log-likelihoods for H_i, $i = 0, 1, 2$, we find that they are -307.76, -217.87, and -217.14 respectively.

• For testing null H_0 against H_1, the (base e) lod score is 89.9. Twice this value (179.8) has approximately a χ^2_1 if H_0 is true. So H_0 is rejected.

• For testing null H_1 against alternative H_2, the lod score is 0.73, and twice this value (1.46) is χ^2_2 if H_1 is true. So H_1 is not rejected.

Added notes:

• Thanks to Aaron:
In the test of H_1 vs the general H_2 we might be testing the assumed 50-50 Mendelian segregation ratios at each locus, for example.

• 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.

• Likelihood ratio tests may be used to test equality of male and female recombination frequencies.

• Note for the intercross experiment, each offspring gives us a male and a female meiosis.

• However, the formulae get messy. For example:
Group III: $2\rho(1-\rho)$ becomes $\rho_m(1-\rho_f) + \rho_f(1-\rho_m)$ since we do not know which gamete is recombinant.

3.3.5 POWER and SAMPLE SIZE:

• If ρ is the true value, the probability $H_0 : \rho = 1/2$ is rejected is the power function of the test.
We reject if $T < k^* = (n/2) + (\sqrt{n}/2)\Phi^{-1}(\alpha)$, so

$$\Pr(T < k^*; \rho) \;=\; \Pr\left(\frac{T - n\rho}{\sqrt{n\rho(1-\rho)}} < \frac{k^* - n\rho}{\sqrt{n\rho(1-\rho)}}\right)$$

$$\approx\; \Phi\left(\frac{k^* - n\rho}{\sqrt{n\rho(1-\rho)}}\right) \;=\; \Phi\left(\frac{\Phi^{-1}(\alpha) + \sqrt{n}(1-2\rho)}{2\sqrt{\rho(1-\rho)}}\right)$$

again using the Normal approximation to the Binomial distribution.

• Power decreases over $0 \leq \rho \leq \frac{1}{2}$. Clearly, for a given sample size, linkage is more easily detected when ρ is small. Conversely, for given ρ, one may determine the sample size n required for given power.

• For phase-unknown backcross with 2-kid families:
The power and sample-size computations are exactly as for the phase-known case, with ρ^* replacing ρ, and n now denoting the number of two-child families.

3.3.6 Kullback-Leibler information:

• For general, we can find the Kullback-Leibler information, which is just expected log-likelihood difference.

• For a sample of size n and true value of the parameter $\mathbf{q}$:

$$K_n(\mathbf{q_0}; \mathbf{q}) \;=\; \mathbf{E}_{\mathbf{q}}(\ell_n(\mathbf{q}) \;-\; \ell_n(\mathbf{q_0}))$$

where $\mathbf{q_0}$ is some hypothesized value.

• For multinomial data, K_n takes a simple form. suppose there are c categories and cell probabilities $\mathbf{q} = (q_j),\ j = 1, ..., c$.
Then $\ell_n(\mathbf{q}) \;=\; \sum_{j=1}^{c} n_j \log_e q_j$ and

$$K_n(\mathbf{q_0}; \mathbf{q}) \;=\; n\sum_{j=1}^{c} q_j \log_e q_j \;-\; n\sum_{j=1}^{c} q_j \log_e q_{0j}$$

or, for a single observation,

$$K_1(\mathbf{q_0}; \mathbf{q}) \;=\; \sum_{j=1}^{c} q_j \log_e \left(\frac{q_j}{q_{0j}}\right).$$

• In the case of linkage analysis data, $q_j = q_j(\rho)$ and the null hypothesis is $H_0 : \rho = \frac{1}{2} :\ q_{0j} = q_j(\frac{1}{2})$.

3.3.7 Information for testing for linkage; $H_0 : \rho = 1/2$:

• Evaluating K_1 for the above *phase-known intercross* experiment, and for the previous binomial *phase-known* and *phase unknown* backcross experiments, we see the information per offspring individual:

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

• This measures information for rejecting $\rho = \frac{1}{2}$ when ρ is the true value; the more ρ differs from $\frac{1}{2}$ the more information there is.

• Each *phase-known* offspring contributes at least twice as much as in the phase-unknown case (2 kids per family). When ρ is close to $1/2$, the *phase-unknown* two-offspring design has low power.

• Each *intercross* offspring contains more information than a *backcross* offspring, but *not* twice as much information. As $\rho \to \frac{1}{2}$, there is almost no additional information in the intercross design.

3.3.8 Elods and sample size:

• The Kullback-Leibler information for testing $\rho = \frac{1}{2}$ is the expected base-e lod score at the true value ρ_T of the recombination frequency. This base-10, is a measure very widely used in linkage analysis and known as the *Elod*.

• Note that when n is large, we expect the base-e lod score to be broadly equal to its expectation nK_1.

• For our previous data with $n = 200$, we had $\widehat{\rho} = 0.12$; in fact, the data were simulated at $\rho = 0.1$. Then 200×0.479 is about 95, in good agreement with the lod score value of 90 which we obtained.

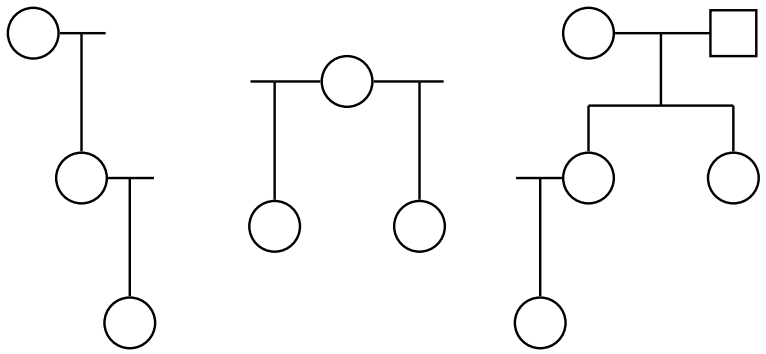
• This also tells us that if we had realized that ρ_T might be around 0.1, it was very wasteful to breed 200 mice. When $\rho_T = 0.1$, about 20 mice are expected to give a lod score (base e) of more than 9; this is plenty to detect that $\rho \neq \frac{1}{2}$.

• Note again that we have used natural logarithms in these examples, contrary to standard practice in genetics.

3.4.1 Data at linked loci can distinguish relationships:

• The simplest example is for the three unilateral ($\kappa_2 = 0$) pairwise relationships of grandmother-granddaughter (G), half-sisters (H), and aunt-niece (N).

• Each of these relationships has $\boldsymbol{\kappa} = (\kappa_0, \kappa_1, \kappa_2) = (\frac{1}{2}, \frac{1}{2}, 0)$, and hence they are indistinguishable on the basis of data at independently segregating (unlinked) loci.



3.4.2 Gene identity at two linked loci:

• For unilateral relationships, gene identity at two linked loci is summarized by

$\kappa_{1,1}(\rho)$ = P(share 1 gene *ibd* at each of 2 loci at recombination ρ).

loc 2	locus 1		
	0	1	
0			$1-\kappa_1$
1		$\kappa_{1,1}$	κ_1
	$1-\kappa_1$	κ_1	1

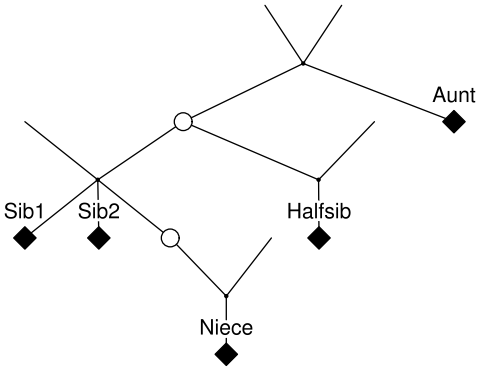
• For the three relationships G, H, N, we have

$$G : \kappa_{1,1}(\rho) = \frac{1}{2}(1-\rho)$$
$$H : \kappa_{1,1}(\rho) = \frac{1}{2}(\rho^2 + (1-\rho)^2) = \frac{1}{2}R \text{ say}$$
$$N : \kappa_{1,1}(\rho) = \frac{1}{2}((1-\rho)R + \rho/2).$$

Thus the relationships are identifiable of the basis of data at two linked loci ($0 < \rho < \frac{1}{2}$), but not on the basis of data at unlinked loci. All the three relationships have $\kappa_{1,1}(0) = \frac{1}{2}$ and $\kappa_{1,1}(\frac{1}{2}) = \frac{1}{4}$.

3.4.3 A real-data example of three related individuals:

• Genotypes at many markers were available on three putative sibs.
• Two were very likely sibs, but the third seemed to be an aunt, niece, or half-sister of the pair.
• To distinguish these alternatives we must consider loci jointly (linked), and the three individuals jointly (not pairwise).



See next page for the *ibd* state probabilities.

	Individuals	
	Pairwise	Joint
Loci unlinked	$H \equiv N \equiv A$	$H \equiv N$
Loci linked	$N \equiv A$	H, N, A identifiable

• First thought to be H (usual conjecture). A is genetically unlikely.
Due to missing data, H and N were hard to resolve.
But missing sib, mom of N, was found mislabeled in another family.

3.4.4 The ibd states for the above example:

Gene *ibd* state for two sibs with	Prior (pedigree) probability (a) an aunt	(b) a niece or half-sib
Sibs sharing 2 *ibd*		
1 2 1 2 1 3	1/8	1/8
1 2 1 2 3 4	1/8	1/8
Sibs sharing 1 *ibd*		
1 2 1 3 1 4	1/8	1/8
1 2 1 3 2 3	1/16	0
1 2 1 3 2 4	1/16	1/8
1 2 1 3 3 4	1/16	1/8
1 2 1 3 4 5	3/16	1/8
Sibs sharing 0 *ibd*		
1 2 3 4 1 3	1/16	0
1 2 3 4 1 5	1/16	1/8
1 2 3 4 3 5	1/16	1/8
1 2 3 4 5 6	1/16	0

3.4.5 Two-locus kinship:

• Kinship is the probability that gametes segregating from individuals are *ibd*. Two-locus kinship is the probability gametes are *ibd* at both of two linked loci.

• For a single locus, recall that kinship ψ is given by
$\psi = (\kappa_1 + 2\kappa_2)/4 = \kappa_1/4$ if $\kappa_2 = 0$ (unilateral relatives) .

• For two loci, although $\kappa_{1,1}(\rho)$ is sufficient to specify pairwise genotype and phenotype distributions, it may not determine the two-locus kinship of the individuals, because the shared genes at the two loci may be on the same haplotype in the individual, or on different ones.

• In fact, in H they are necessarily on the same (maternal) haplotype in the two half-sibs, while in G they may be on either haplotype of the grandmother. In fact, G and H have the same two-locus kinship, $(1/8)(1-\rho)^2 R$.

• For N, for the first term they are on the same haplotype in the aunt, while the last term corresponds to the case where the genes at the two loci are on two different haplotypes in the aunt.

3.5.1 HOMOZYGOSITY MAPPING: Likelihoods for linkage:

• Consider the case of just 2 loci: a trait locus with a recessive allele causing the disease, and a marker locus with some number of alleles.

• Suppose the frequency of the recessive disease allele is q, and at the marker locus alleles A_i have frequencies p_i.

• Suppose that the affected individual has inbreeding coefficient f, and probability $f_2(\rho)$ of carrying genes *ibd* at both of two loci between which the recombination frequency is ρ.

loc 2	locus 1 *ibd*		
ibd	no	yes	
no			$1-f$
yes		$f_2(\rho)$	f
	$1-f$	f	1

• Then the probability the individual is autozygous at a specific one of the two loci but not the other is $f - f_2(\rho)$, and the probability he is autozygous at neither is $(1 - 2f + f_2(\rho))$.

• Note at $\rho = 0$, $f_2(0) = f$ (the two loci segregate together), then the four cell probabiities are $(1-f)$, 0. 0, and f.
At $\rho = 1/2$, $f_2(1/2) = f^2$ (the two loci segregate indepedently). and we have $(1-f)^2$, $f(1-f)$, $f(1-f)$, and f^2.

3.5.2 Likelihoods for linkage:

• Probabilities at the trait locus and at the marker, given *ibd* and given non-*ibd*.

	trait affected	marker A_jA_j	marker A_jA_l
ibd	q	p_j	0
not-*ibd*	q^2	p_j^2	$2p_jp_l$

• If the individual has marker phenotype A_jA_l the likelihood ratio is

$$\begin{aligned}\frac{L(\rho)}{L(\rho=\frac{1}{2})} &= \frac{\Pr(\text{data}\ ;\ \rho)}{\Pr(\text{data}\ ;\ \rho=\frac{1}{2})} \\ &= \frac{2p_jp_l(q(f-f_2(\rho))+q^2(1-2f+f_2(\rho)))}{2p_jp_l(q(f-f^2)+q^2(1-f)^2)} \\ &= \frac{(f-f_2(\rho))+q(1-2f+f_2(\rho)))}{(1-f)(f+q(1-f))}.\end{aligned}$$

The coefficient of $f_2(\rho)$ in numerator is $(-1+q)<0$.

LR increases, as $f_2(\rho)$ decreases, and as hence ρ increases: $\widehat{\rho}=\frac{1}{2}$.

The likelihoods ctd.:

• If the individual has marker phenotype A_jA_j the likelihood ratio is

$$\begin{aligned}\frac{L(\rho)}{L(\rho=\frac{1}{2})} &= \frac{\Pr(\text{data}\ ;\ \rho)}{\Pr(\text{data}\ ;\ \rho=\frac{1}{2})} \\ &= \frac{qp_jf_2+q^2p_j(f-f_2)+qp_j^2(f-f_2)+q^2p_j^2(1-2f+f_2)}{qp_jf^2+q^2p_jf(1-f)+qp_j^2f(1-f)+\ q^2p_j^2(1-f)^2} \\ &= \frac{f_2+q(f-f_2)+p_j(f-f_2)+qp_j(1-2f+f_2)}{f^2+qf(1-f)+p_jf(1-f)+qp_j(1-f)^2}.\end{aligned}$$

• The coefficient of $f_2(\rho)$ is $(1-q)(1-p_j)>0$. LR increases as $f_2(\rho)$ increases, i.e. as ρ decreases, so $\widehat{\rho}=0$, $f_2(0)=f$ and

$$\frac{L(\rho=0)}{L(\rho=\frac{1}{2})} = \frac{f+(1-f)qp_j}{(f+(1-f)q)(f+(1-f)p_j)}.$$

This is always ≥ 1, and increases as q or $p_j \to 0$.

• For $q\approx 0$, $L(\rho=0)/L(\rho=\frac{1}{2}) \ \to\ 1/(f+(1-f)p_j)$,

3.5.3 Lod scores and elods:

• Log-likelihoods are additive over unrelated pedigrees i. The base-10 lod score is

$$\mathrm{lod}(\rho) \;=\; \sum_i \log_{10}\left(\frac{L_i(\rho)}{L_i(\rho=\frac{1}{2})}\right)$$

where $L_i(\cdot)$ is the likelihood contributed by pedigree i. The maximized lod score is $\max_{0\le\rho\le\frac{1}{2}}(\mathrm{lod}(\rho))$. Combining over pedigrees, $\widehat{\rho}$ may be neither 0 nor $\frac{1}{2}$. Then $f_2(\rho)$ is also relevant, not only f.

• Again, a useful measure of information for linkage analysis is the *elod*:

$$\textit{elod}(\rho) \;=\; \mathbf{E}_\rho(\mathrm{lod}(\rho)).$$

The *elod* is additive over independent pedigrees.

Example:

• For simplicity consider $\rho = 0$ and $q \approx 0$. Then each affected individual has prob ≈ 1 of having two *ibd* genes at the disease locus, and hence also at marker ($\rho = 0$), and so has prob p_j of being A_jA_j. Hence each contributes

$$-\sum_j p_j \log(f + (1-f)p_j).$$

to the *elod*. For example, for the affected offspring of first-cousin marriages ($f = 1/16$), and a polymorphic marker locus (for example, $p_j = 0.1$ for each of 10 alleles) the value is $\log_{10}(6.4) = 0.81$. (Just 4 such individuals could give a base-10 *elod* of 3.)

• Note the sensitivity of lod scores and elods to the marker allele frequency. For example: As above, a homozygous A_jA_j individual gives lod-score contribution

$$-\log_{10}(f + (1-f)p_j) = \log_{10}(16/(1+15p_j)) \text{ if } f = 1/16.$$

If $p_j = 0.1$ this is 0.81. if $p_j = 0.5$ this is 0.30. Incorrectly assuming too small an allele frequency can give false evidence for linkage.

3.5.4 Allelic association and fine-scale mapping:

• A small number of unrelated affected individuals all homozygous at the same polymorphic marker locus provides strong evidence for linkage. Even more so if they share the same marker allele – WHY? Recall the origin and decay of LD (1.6.2).

blank page: