The

Machinery of Parametric Linkage Analysis

David Duffy

Queensland Institute of Medical Research
Brisbane, Australia

Introduction

- Mendelism
- Linkage
- Statistical distributions
- Maximum likelihood linkage analysis
- The generalized single major locus model

Mendel and Mendelism

- Mendel studied **binary** traits
- Had parental lines that bred true for traits (**homozygous**)
- F_1 hybrid offspring were homogenous
- F_2 generation exhibited Mendelian ratios
    - 3:1
    - 1:2:1

Backcross

- F_1 with P_1 or P_2
- Simpler ratios
- Simpler interpretation in case of linkage

		Paternal Genotype = **Ff** (F_1) **Slightly frizzled**	
		F (50%)	**f** (50%)
Maternal Genotype = **FF** **Frizzled** (P_1)	F (50%)	**FF** (25%) **Frizzled**	**Ff** (25%) **Slightly Frizzled**
	F (50%)	**FF** (25%) **Frizzled**	**Ff** (25%) **Slightly Frizzled**

The Other Backcross

Maternal Genotype = **ff**	f (50%)	**Ff** (25%)	**ff** (25%)
Normal (P_2)		**Slightly Frizzled**	**Normal**
	f (50%)	**Ff** (25%)	**ff** (25%)
		Slightly Frizzled	**Normal**

Dihybrid testcross

- Backcross involving two traits
- If both are dominant, see a 1:1:1:1 ratio in the (informative) **testcross**

Two traits in the potato plant: **Tall** v. **Dwarf**, and **Cut leaf** v. **Potato cut leaf**.

Counts in the backcross generation (MacArthur 1931): **Tall, Cut** (F_1) x **Dwarf, Potato**

	Tall	**Dwarf**	
Cut	77	72	149
Potato	62	73	135
	139	145	284

Linkage in a dihybrid testcross

- Deviation from a 1:1:1:1 ratio is due to linkage between the trait loci

Two traits in the chicken: **Frizzled** v. **Normal**, and **White** v. **Coloured**.

Counts in the testcross (Hutt 1931): *White, Frizzled* (F_1) x *Coloured, Normal*

	White	**Coloured**	
Frizzled	18	63	81
Normal	63	13	76
	81	76	157

The **recombination fraction c** = (18+13)/157 = 0.197.

Phase: Coupling and repulsion

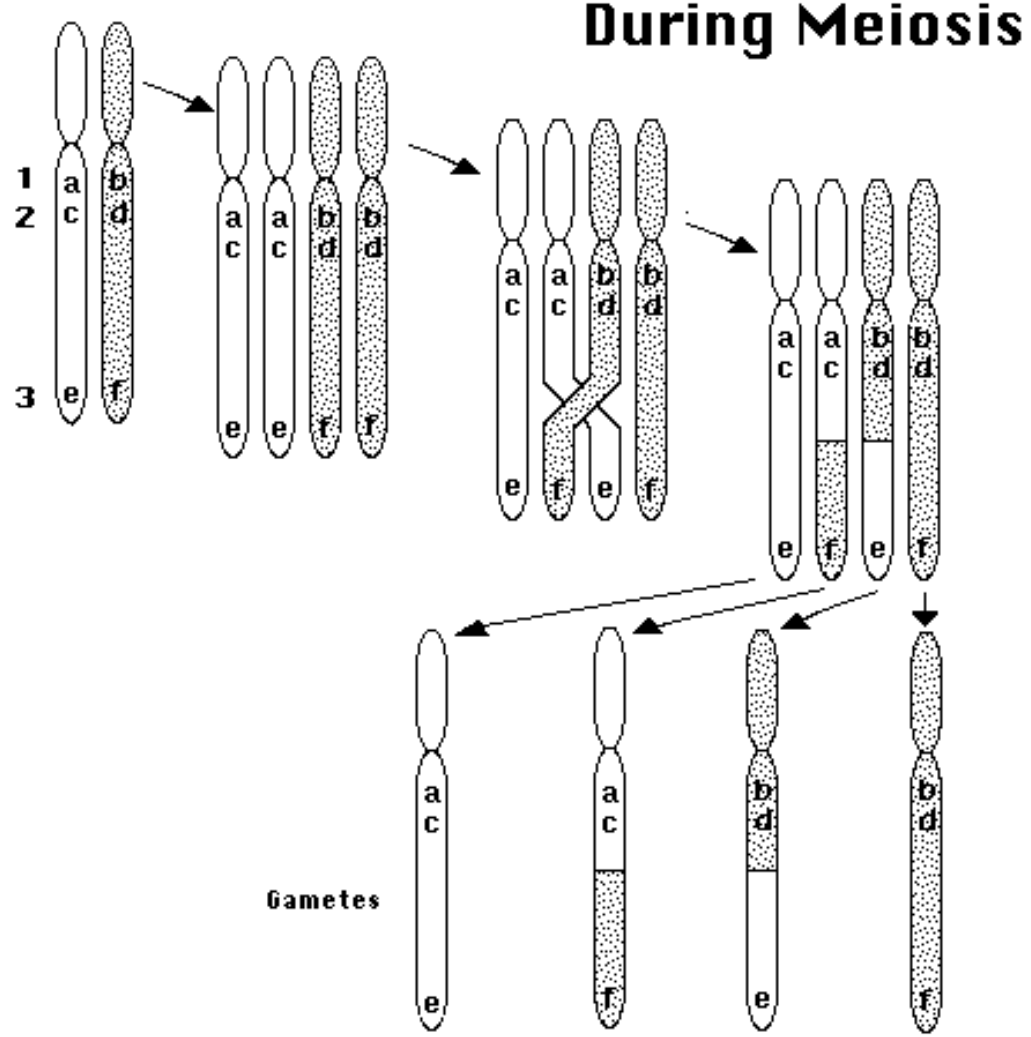
Counts from another mating (Hutt 1933): *White, Frizzled* (F_1) x *Coloured, Normal*

	White	**Coloured**	
Frizzled	15	2	17
Normal	4	12	16
	19	14	33

The recombination fraction $c = (4+2)/33 = 0.182$.

In this family, the dominant traits *White* and *Frizzled* are in **coupling**, but in the previous family, they were in **repulsion**.

Crossing-over and Recombination During Meiosis

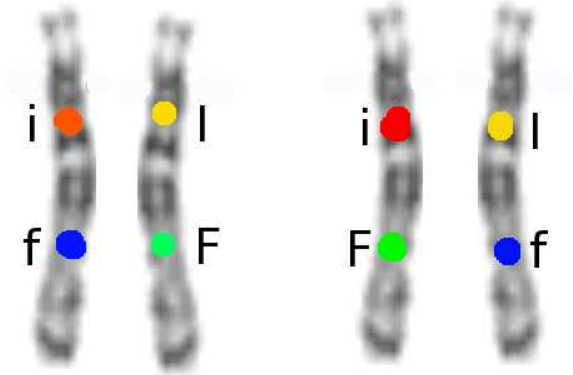


QIMR

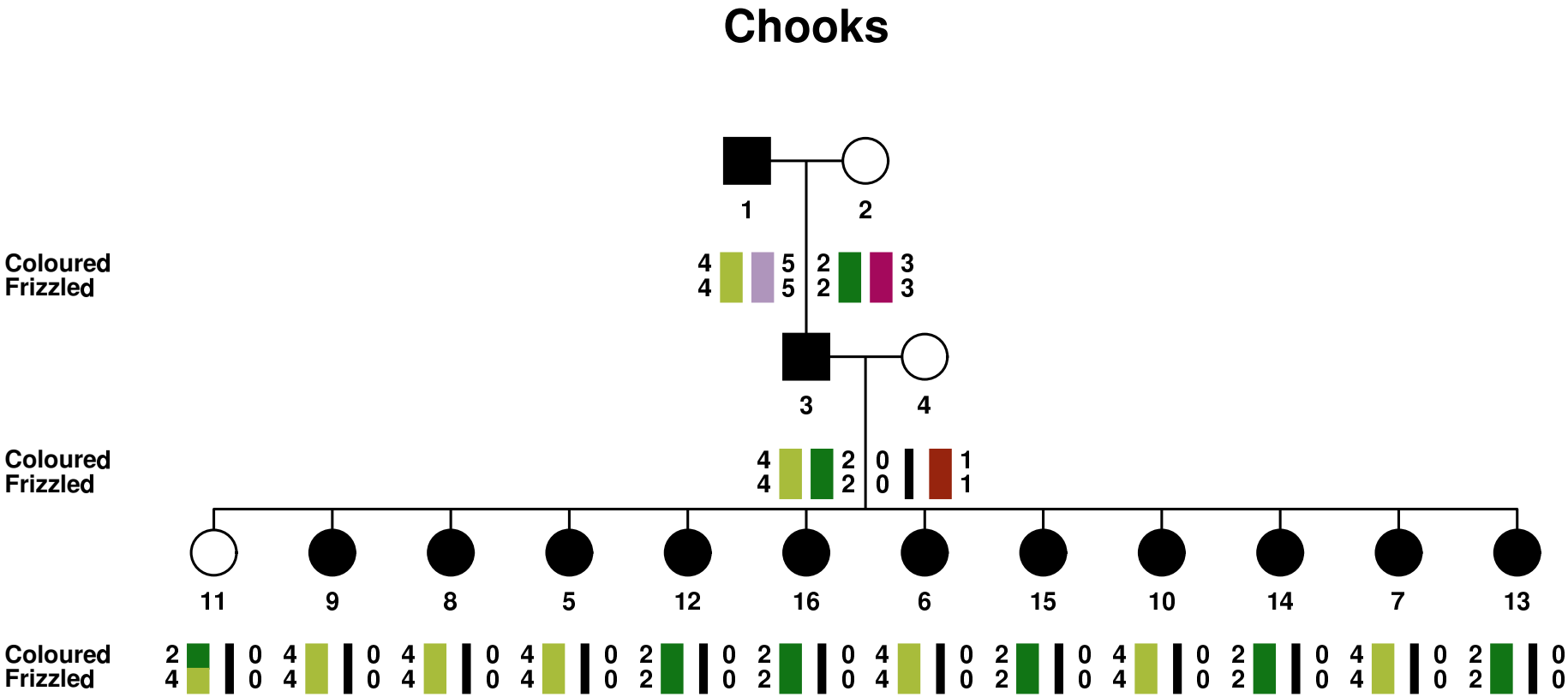
Phase: Coupling and repulsion of frizzled and coloured

In the backcross, only one parent is doubly heterozygous and contributes to the linkage information.

In double heterozygotes, there are two possible arrangements on the chromosomes (the pairs of alleles on each chromosome are **haplotypes**):



	Gametic frequencies			
	IF	**If**	**iF**	**if**
IF/if (coupling)	(1-c)/2	c/2	c/2	(1-c)/2
If/iF (repulsion)	c/2	(1-c)/2	(1-c)/2	c/2



QIMR

Mapping and Multipoint Analysis

- The experimental cross can be extended to involve more loci: *three-point* cross, etc
- The recombination fractions between pairs of loci can be used to order loci in the same *linkage group*

The presence of **double recombinants** and **interference** means that recombination fractions are only roughly additive. A **mapping function** adjusts for one or both of these phenomena, allowing us to estimate consistent **genetic map distances**.

So they address questions like, "if c_{AB}=0.4 and c_{BC}=0.4, what should c_{AC} be?". One map unit (1 Morgan) is the (shortest) map distance that is equivalent to c=0.50.

Mapping and Multipoint Analysis

The **Morgan mapping function** is, $x=c$, where x is the distance in map units. This assumes complete interference, and is adequate over small distances.

The **Haldane mapping function** is:

$$x = 0.5 \log(1-2c)$$

$$c = 0.5 (1-e^{-2x})$$

and adjusts for double recombination only. Trow's formula assumes the Haldane mapping function: $c_{AC} = c_{AB} + c_{BC} - 2c_{AB}c_{BC}$.

The **Kosambi mapping function** also allows for interference, but is **not multipoint consistent**, so it very occasionally causes problems in multipoint linkage analysis.

$$x = 0.25 \log[(1+2c)/(1-2c)]$$

$$c = 0.5 (e^{4x}-1)/(e^{4x}+1)$$

Mapping and Multipoint Analysis

Data from three-point cross of corn (*colourless*, *shrunken*, *waxy*) due to Stadler.

	Progeny Phenotype	Count
1	A B C	17959
2	a b c	17699
3	A b c	509
4	a B C	524
5	A B c	4455
6	a b C	4654
7	A b C	20
8	a B c	12
	Total Tested	45832

Statistical Underpinnings

In these experimental crosses, the numbers of offspring per mating is large, so we can neglect statistical uncertainty about:

- The accuracy of the genotypes
- The phase of the mating
- The counts of recombinants and nonrecombinants

Recombination is a binary (*yes-no*, *R-NR*) phenomenon. For a given parental genotype of known phase, the probability of a recombination event in production of a gamete is a constant (c). Each meiosis is an independent **Bernoulli trial**. The count of **recombination events** arising from a number of meioses therefore comes from the **binomial distribution**.

The Binomial Distribution

If two loci are unlinked, c=0.50. For a testcross giving rise to 3 offspring, we expect eight outcomes to be equally likely. While if the two loci are linked, with c=0.10 say, the outcomes with fewer recombinants will be observed more often.

Outcome	c=1/2	c=1/10
R, R, R	1/8	1/1000
R, R, NR	1/8	9/1000
R, NR, R	1/8	9/1000
R, NR, NR	1/8	81/1000
NR, R, R	1/8	9/1000
NR, R, NR	1/8	81/1000
NR, NR, R	1/8	81/1000
NR, NR, NR	1/8	729/1000

The Binomial Distribution

If the order of the events making up each outcome is irrelevant (as it is this case), we say the events are **exchangeable**, and we can summarize the outcomes as counts:

R	NR	c=1/2	c=1/10
3	0	1/8	1/1000
2	1	3/8	27/1000
1	2	3/8	243/1000
0	3	1/8	729/1000

The expected number of recombination events if c=0.5 is E(R)=cN=1.5.

If c=0.1, then E(R)=cN=0.3.

The Likelihood Ratio

If we wish to make a decision about whether two loci are linked, we usually evaluate a **likelihood ratio comparing two hypotheses about our observed data**.

If in our testcross sibship we observed 0 out of 3 recombinants, then the likelihood ratio comparing the two hypotheses c=0.1 and c=0.5 is **the ratio of the probability of observing the data under the two hypotheses**.

Since these probabilities are not "actual" probabilities, but contingent on the underlying hypothesis, Fisher suggested we call them **likelihoods**.

$$L(R = 0, NR = 3 \mid c = 0.5) = 0.125$$

$$L(R = 0, NR = 3 \mid c = 0.1) = 0.001$$

$$LR = 125$$

We interpret this as saying that the hypothesis that c=0.1 is 125 times more likely than the hypothesis that the loci are unlinked.

The Lod Score

Newton Morton suggested in 1955 that a likelihood ratio testing the hypothesis of linkage should be "significant" if it was 1000:1 in favour of a hypothesis where $c < 0.5$. This was based on a sequential testing argument and the length of the *human* genetic map. It is thus a **genome-wide critical significance level**, adjusting for the number of possible tests that could be done.

If the likelihood ratio was 100:1 in favour of the $c = 0.5$ null hypothesis, then he suggested this be accepted as significant evidence for **exclusion of linkage** for that value of c (eg c=0.1). Intermediate ratios were regarded as inconclusive.

Following Barnard (1947), he presented the likelihood ratio as the **decimal log odds** or **lod score**. The lod scores from different families testing the same linkage hypothesis can be added together to obtain a total lod score for that hypothesis. Similarly, for large datasets, the likelihoods for particular hypotheses are usually very small, so **model log likelihoods** are a convenient summary for computations.

Linkage in outbred human families

Human families are relatively small, so phase is harder to evaluate.

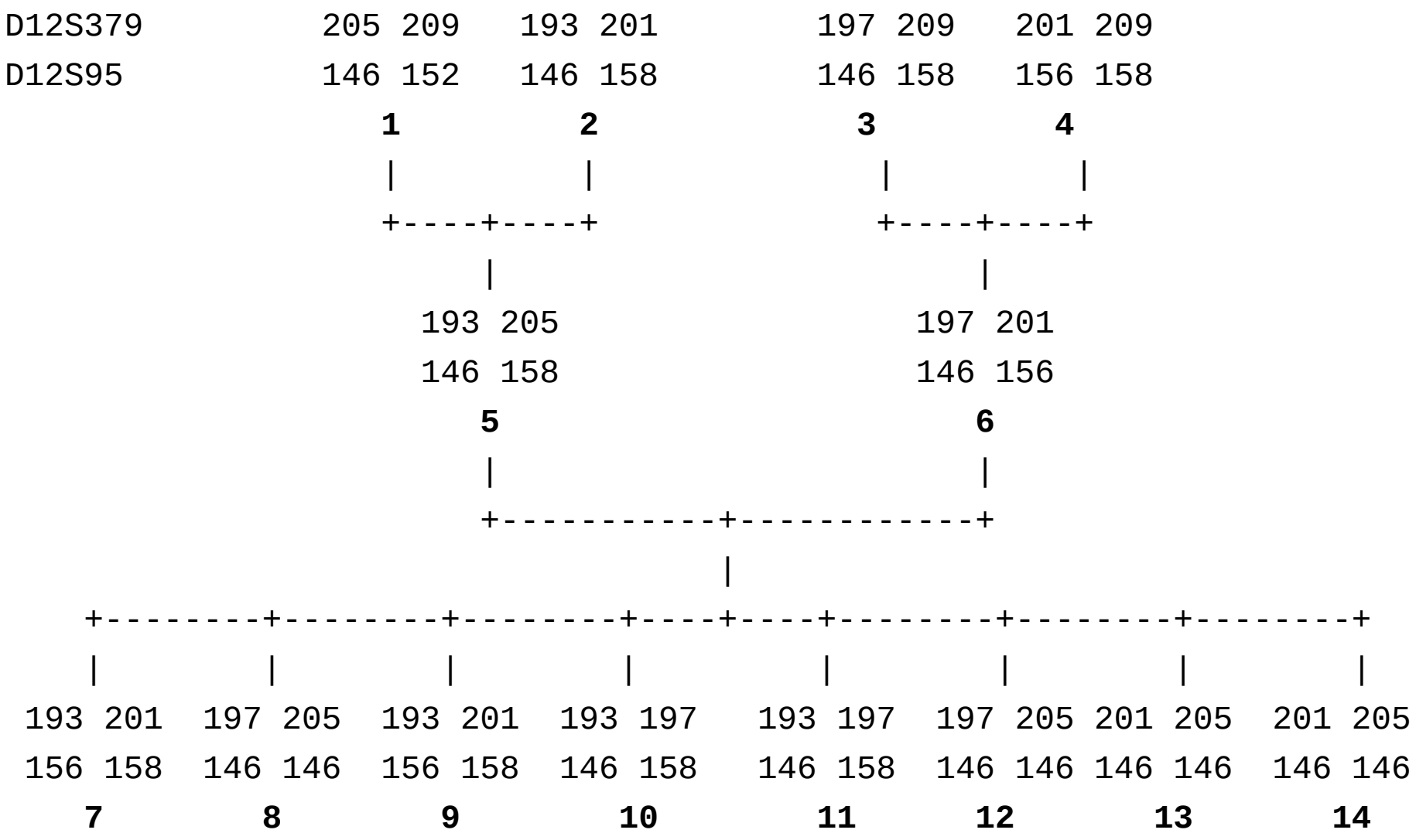
Matings are relatively random, so only a proportion of families in the population are **informative** for linkage analysis at any given marker.

Codominant marker loci and the direct method

One way to work out the phase of a mating is to genotype three generations of a family.

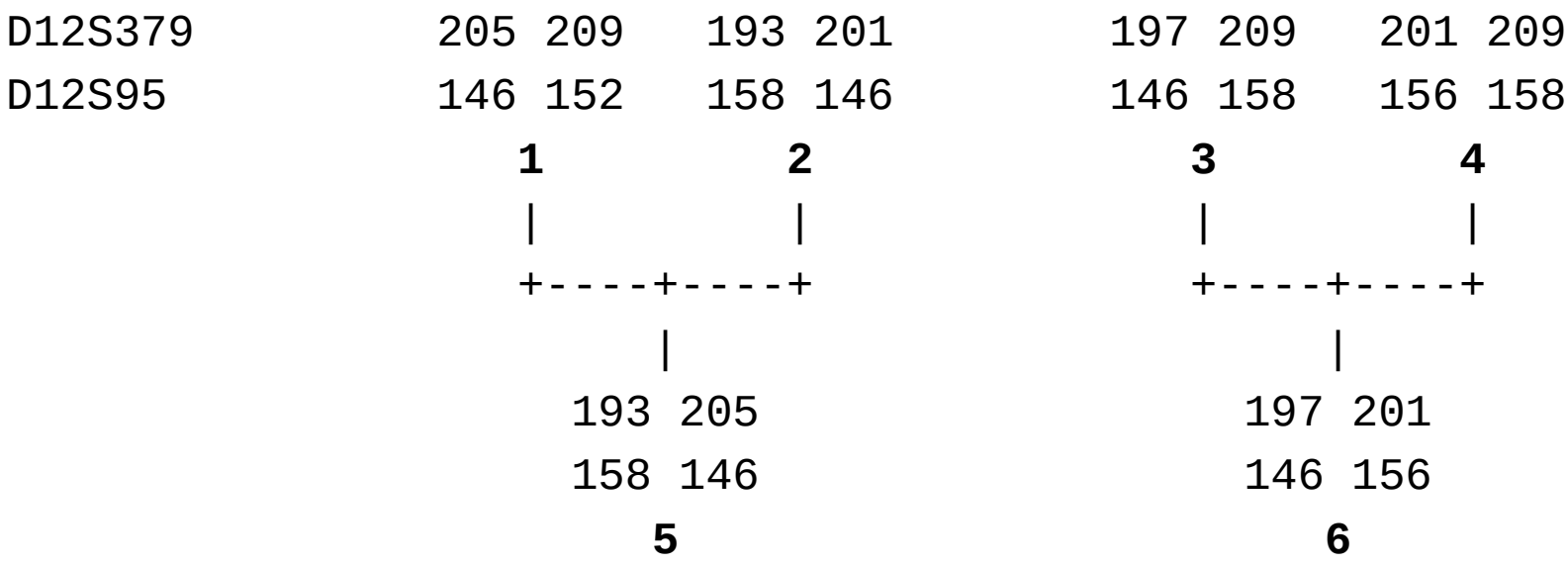
Where there enough doubly heterozygous parents, one can count up the recombination events, as in a planned cross.

Genotypes at D12S379 and D12S95 in an Amish family



QIMR

Direct estimation of recombination fraction 2



The grandparental data allows us to work out that the four gametes that gave rise to the parents **5** and **6** were:

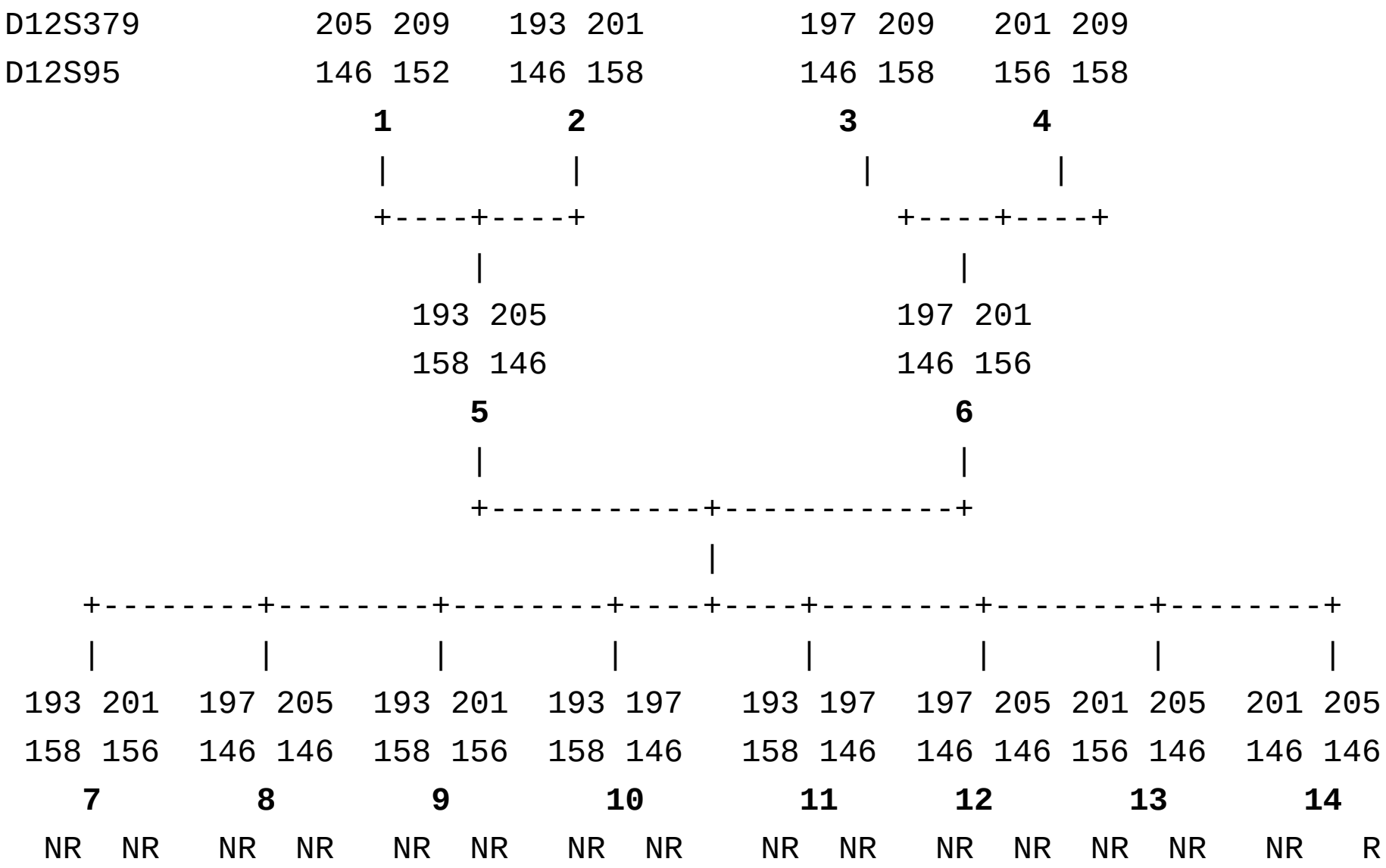
{205,146} from individual **1**,

{193,158} from **2**,

{197,146} from **3**,

{201,156} from **4**.

Direct estimation of recombination fraction 3



QIMR

Direct estimation of recombination fraction 4

This allows us to score the children as to whether these haplotypes have been broken up by a recombination event or not.

Our estimate of the recombination distance between these loci from this family is

$c = 1/16 = 0.0625$.

Because there are so few observations, the 95% confidence interval is wide, from 0.002 to 0.302. Actually, D12S379 and D12S95 are approximately 6 cM apart.

The Lod Score for this Example Pedigree

In our sibship of eight children, one recombinant and fifteen nonrecombinants were observed, so the likelihood for the family is:

$$L(R=1,\ NR=15;\ \mathbf{c}) \ = \ \mathbf{c}^{\mathbf{1}}\ (1-\mathbf{c})^{15}\ .$$

For our example pedigree, the likelihood ratio and the lod score are:

$$LR = \frac{(\frac{1}{16})^1\ (\frac{15}{16})^{15}}{(\frac{1}{2})^1\ (\frac{1}{2})^{15}} = 1555.712\ ; \quad lod = \log_{10}\frac{(\frac{1}{16})^1\ (\frac{15}{16})^{15}}{(\frac{1}{2})^1\ (\frac{1}{2})^{15}} = 3.19$$

An Alternative Interpretation of the Lod

An alternative interpretation of the likelihood ratio, is that (aymptotically),

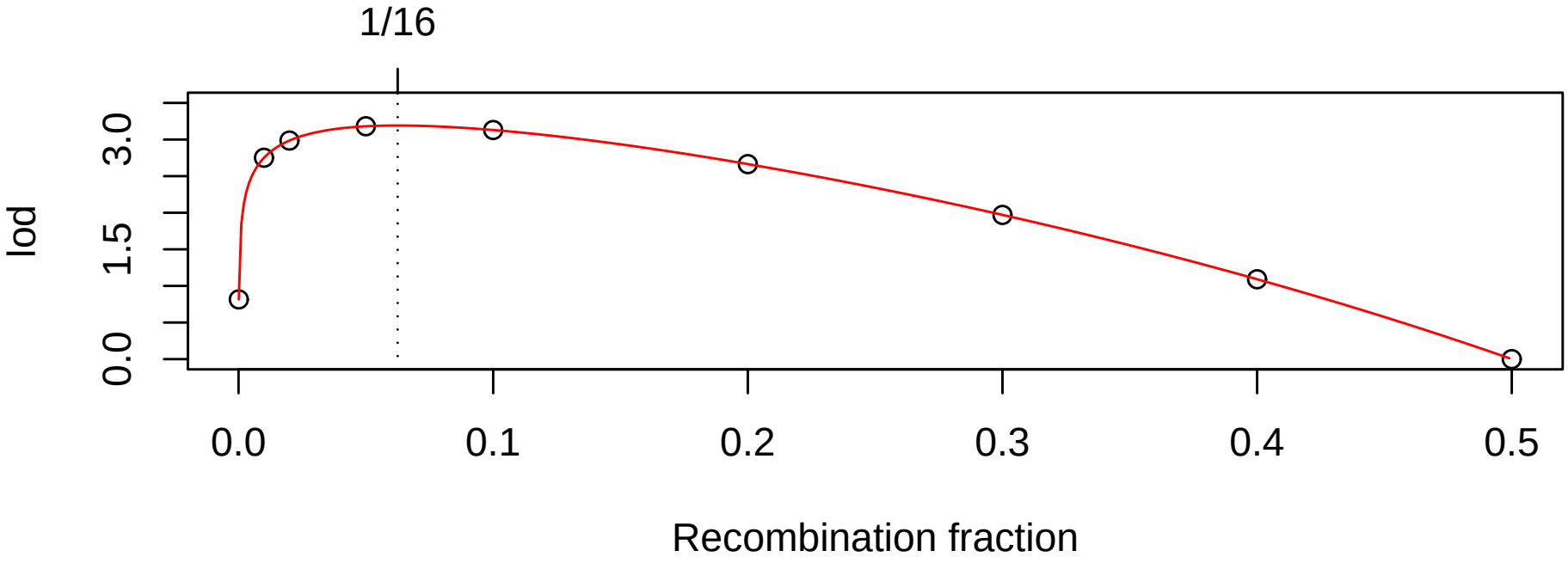
$2\log_e(LR) \sim \chi_1^2$, the chi-square distribution.

So, we can calculate a P-value for a lod score:

lod	P-value
0	0.5
1	0.016
2	0.0012
3	0.00010
4	0.000009

Maximizing the lod score

Computer programs for linkage analysis calculate the lod score for a grid of different values of **c**. The value of **c** which maximizes the lod score as the **maximum likelihood estimate**:



QIMR

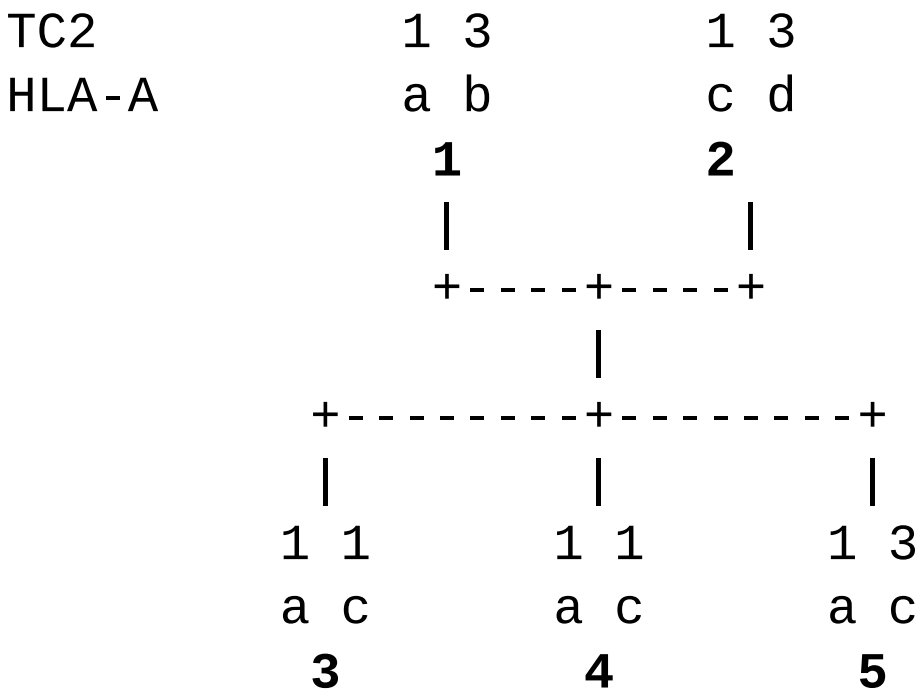
Evaluating the lod score for ambiguous families 1

In most situations, the grandparents are unavailable, or grandparents or parents may be homozygous at a marker.

We can still calculate a pedigree likelihood:

- List all the possible haplotype arrangements
- Calculate a likelihood for each arrangement
- Calculating the average of these likelihoods

Evaluating the lod score for ambiguous families 2



The likelihood for this family is:

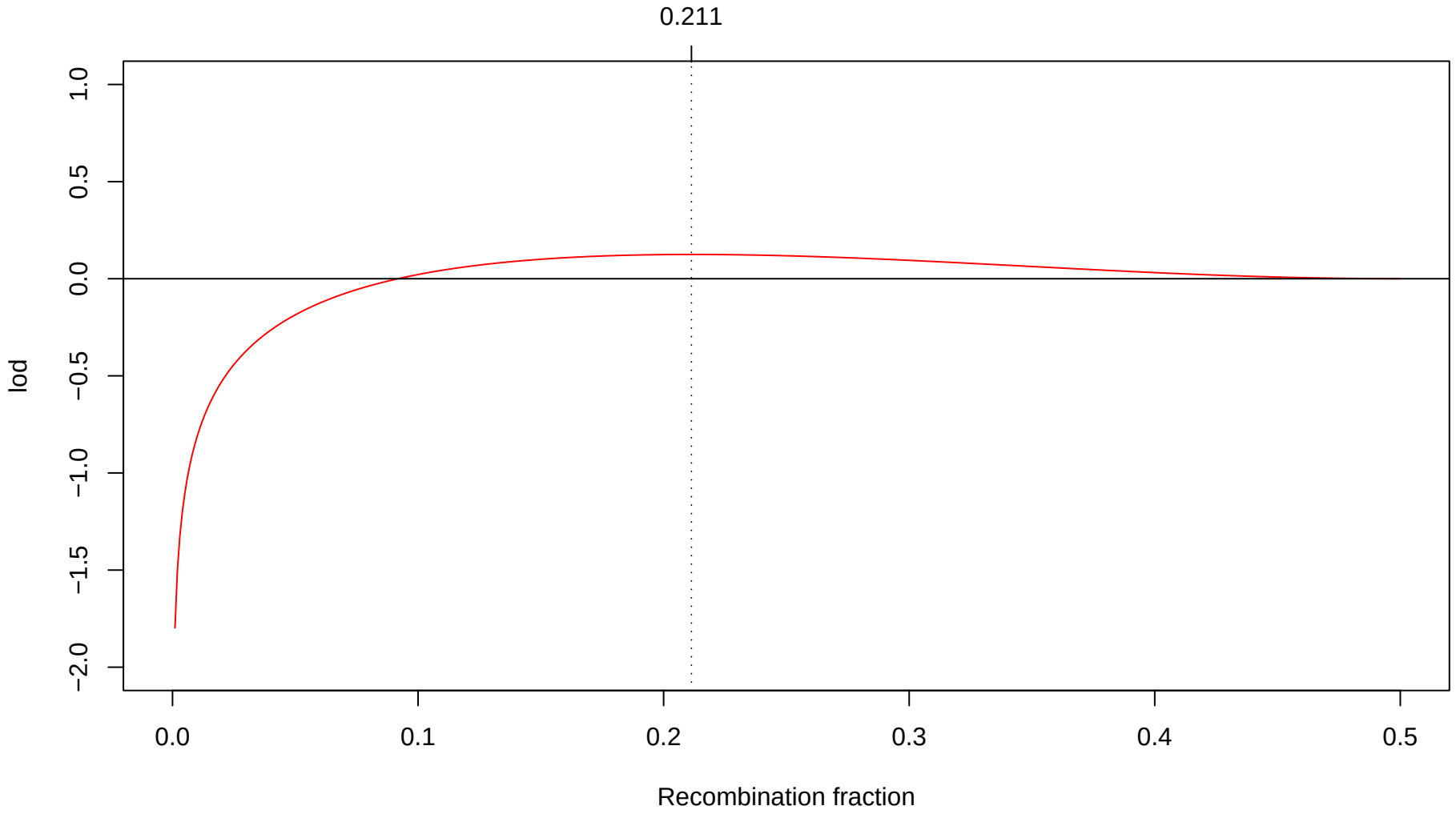
$$L(c) = \frac{1}{4}c(1-c)[1-3c(1-c)].$$

Evaluating the lod score for ambiguous families 3

This formula arises from the fact that both parents are phase unknown, as is the individual 5. Each of the eight possible arrangements is equally likely:

Person **1**	Person **2**	Person **5**	Recombinants	Likelihood
1a/3b	1c/3d	1a/3c	NR,NR,NR,NR,NR,R	$\mathbf{c}(1-\mathbf{c})^5$
1a/3b	1c/3d	1c/3a	NR,NR,R,NR,NR,NR	$\mathbf{c}(1-\mathbf{c})^5$
1a/3b	1d/3c	1a/3c	NR,NR,NR,R,R,NR	$\mathbf{c}^2(1-\mathbf{c})^4$
1a/3b	1d/3c	1c/3a	NR,NR,R,R,R,R	$\mathbf{c}^4(1-\mathbf{c})^2$
1b/3a	1c/3d	1a/3c	R,R,R,NR,NR,R	$\mathbf{c}^4(1-\mathbf{c})^2$
1b/3a	1c/3d	1c/3a	R,R,NR,NR,NR,NR	$\mathbf{c}^2(1-\mathbf{c})^4$
1b/3a	1d/3c	1a/3c	R,R,R,R,R,NR	$\mathbf{c}^5(1-\mathbf{c})$
1b/3a	1d/3c	1c/3a	R,R,NR,R,R,R	$\mathbf{c}^5(1-\mathbf{c})$

The lod for the family is the average of these eight possibilities. It reaches its maximum value Z_{max} at **c**=0.21.



QIMR

Parametric linkage analysis of a trait locus 1

For highly penetrant trait loci, we can infer the underlying genotype based on the observed phenotype.

We need to know the likely mode of inheritance, and how common the risk allele is in the general population.

For example, for a rare familial disease that appears to be dominantly inherited, we can score each affected person as **Dd**, each unaffected person as **dd**, and take each **D** allele as coming from a single pedigree founder.

For a condition that appears to be recessively inherited, we score affected persons as **DD**, and their parents as **dD**.

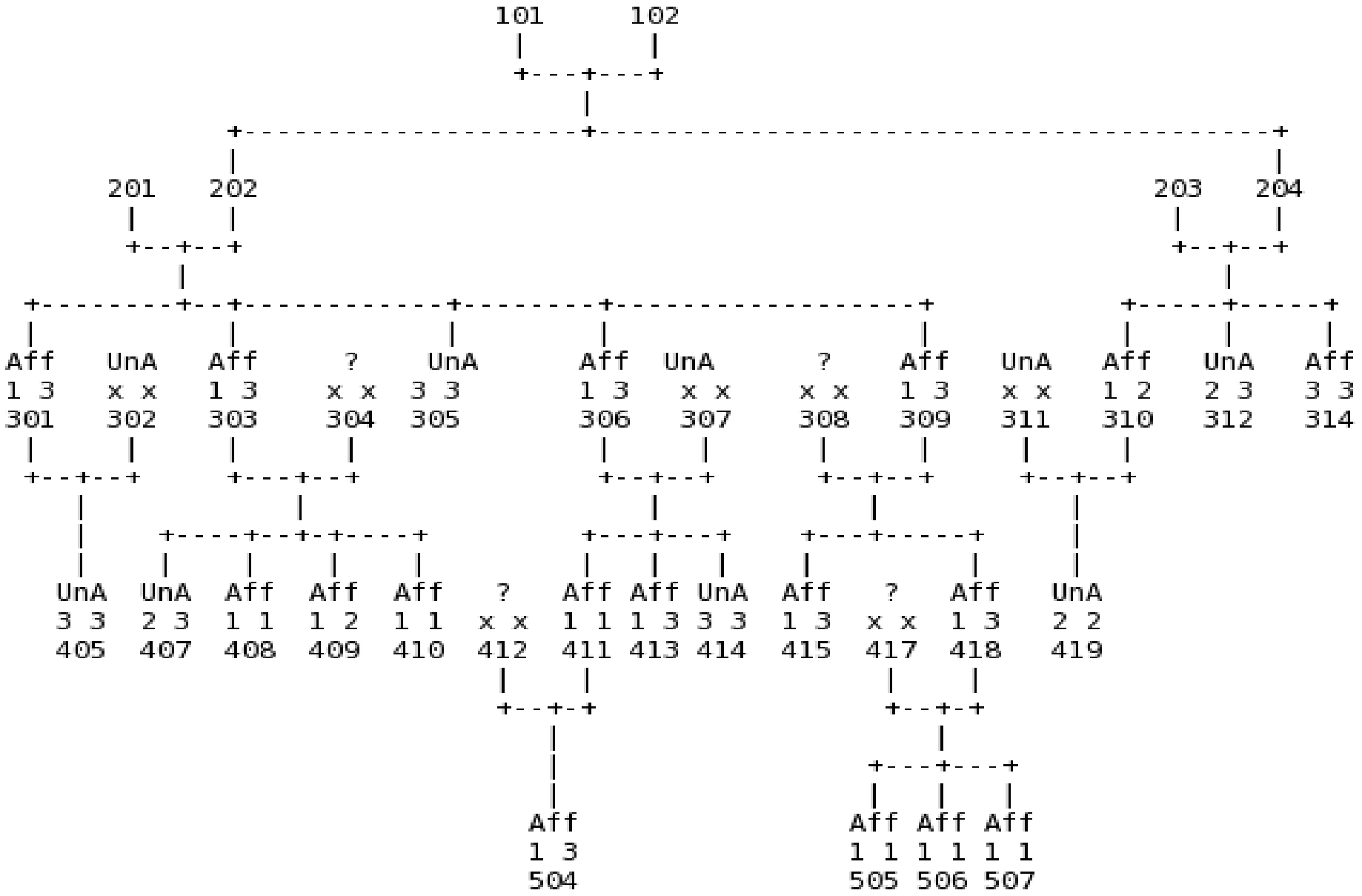
Parametric linkage analysis of a trait locus 2

Morton (1956) analysed *familial elliptocytosis* pedigrees collected by Lawler and Sandler (1954) for linkage to Rhesus blood group (a codominant marker).

This paper is also one of the first examples of testing for homogeneity of linkage in different pedigrees.

We will concentrate on one of the linked pedigrees.

Pedigree 5 from Lawler and Sandler (1954) used by Morton (1956).



QIMR

Parametric linkage analysis of a trait locus 3

One needs to know that familial elliptocytosis is extremely rare, so that the population allele frequency of the disease allele is very low.

Examination of this pedigree and others shows the inheritance is consistent with fully penetrant autosomal dominant inheritance.

Also, the allele frequencies for the marker locus (Rhesus blood group) are well known, so for individuals who are untyped, we can weight the possibilities appropriately (0.4076, 0.1411, 0.3886, 0.0627).

Parametric linkage analysis of a trait locus 4

Morton (1956) gives the lod score expression for this family as:

$$\mathbf{Z} = \log_{10} 2^{20}/39168 \quad \{810\mathbf{c}(1-\mathbf{c})^{19} + 324\mathbf{c}(1-\mathbf{c})^{18} + 180\mathbf{c}(1-\mathbf{c})^{17} + 72\mathbf{c}(1-\mathbf{c})^{16} + 90\mathbf{c}^{3}(1-\mathbf{c})^{17} + 72\mathbf{c}^{3}(1-\mathbf{c})^{16} + 40\mathbf{c}^{3}(1-\mathbf{c})^{15} + 24\mathbf{c}^{3}(1-\mathbf{c})^{14} + 90\mathbf{c}^{4}(1-\mathbf{c})^{15} + 20\mathbf{c}^{4}(1-\mathbf{c})^{13} + 90\mathbf{c}^{5}(1-\mathbf{c})^{15} + 432\mathbf{c}^{5}(1-\mathbf{c})^{14} + 20\mathbf{c}^{5}(1-\mathbf{c})^{13} + 104\mathbf{c}^{5}(1-\mathbf{c})^{12} + 1800\mathbf{c}^{6}(1-\mathbf{c})^{14} + 558\mathbf{c}^{6}(1-\mathbf{c})^{13} + 440\mathbf{c}^{6}(1-\mathbf{c})^{12} + 176\mathbf{c}^{6}(1-\mathbf{c})^{11} + 90\mathbf{c}^{7}(1-\mathbf{c})^{13} + 324\mathbf{c}^{7}(1-\mathbf{c})^{12} + 120\mathbf{c}^{7}(1-\mathbf{c})^{10} + 360\mathbf{c}^{8}(1-\mathbf{c})^{12} + 378\mathbf{c}^{8}(1-\mathbf{c})^{11} + 80\mathbf{c}^{8}(1-\mathbf{c})^{10} + 76\mathbf{c}^{8}(1-\mathbf{c})^{9} + 4\mathbf{c}^{8}(1-\mathbf{c})^{4} + 180\mathbf{c}^{9}(1-\mathbf{c})^{11} + 522\mathbf{c}^{9}(1-\mathbf{c})^{10} + 80\mathbf{c}^{9}(1-\mathbf{c})^{9} + 100\mathbf{c}^{9}(1-\mathbf{c})^{8} + 10\mathbf{c}^{9}(1-\mathbf{c})^{3} + 180\mathbf{c}^{10}(1-\mathbf{c})^{10} + 846\mathbf{c}^{10}(1-\mathbf{c})^{9} + 40\mathbf{c}^{10}(1-\mathbf{c})^{8} + 216\mathbf{c}^{10}(1-\mathbf{c})^{7} + 18\mathbf{c}^{10}(1-\mathbf{c})^{4} + 4\mathbf{c}^{10}(1-\mathbf{c})^{2} + 1170\mathbf{c}^{11}(1-\mathbf{c})^{9} + 378\mathbf{c}^{11}(1-\mathbf{c})^{8} + 260\mathbf{c}^{11}(1-\mathbf{c})^{7} + 72\mathbf{c}^{11}(1-\mathbf{c})^{6} + 45\mathbf{c}^{11}(1-\mathbf{c})^{3} + 180\mathbf{c}^{12}(1-\mathbf{c})^{8} + 396\mathbf{c}^{12}(1-\mathbf{c})^{7} + 40\mathbf{c}^{12}(1-\mathbf{c})^{5} + 18\mathbf{c}^{12}(1-\mathbf{c})^{2} + 270\mathbf{c}^{13}(1-\mathbf{c})^{7} + 234\mathbf{c}^{13}(1-\mathbf{c})^{6} + 40\mathbf{c}^{13}(1-\mathbf{c})^{5} + 52\mathbf{c}^{13}(1-\mathbf{c})^{4} + 180\mathbf{c}^{14}(1-\mathbf{c})^{6} + 108\mathbf{c}^{14}(1-\mathbf{c})^{5} + 80\mathbf{c}^{14}(1-\mathbf{c})^{4} + 16\mathbf{c}^{14}(1-\mathbf{c})^{3} + 90\mathbf{c}^{15}(1-\mathbf{c})^{5} + 162\mathbf{c}^{15}(1-\mathbf{c})^{4} + 20\mathbf{c}^{15}(1-\mathbf{c})^{3} + 180\mathbf{c}^{16}(1-\mathbf{c})^{4} + 72\mathbf{c}^{16}(1-\mathbf{c})^{3} + 90\mathbf{c}^{17}(1-\mathbf{c})^{3}\}$$

The reported peak lod score in the paper was 3.31 at a recombination distance of approximately 5% (this may be an error as the equation above has a maximum value of only 2.84; MLINK gives a lod score of 3.40 at **c**=0.05).

Multipoint linkage analysis

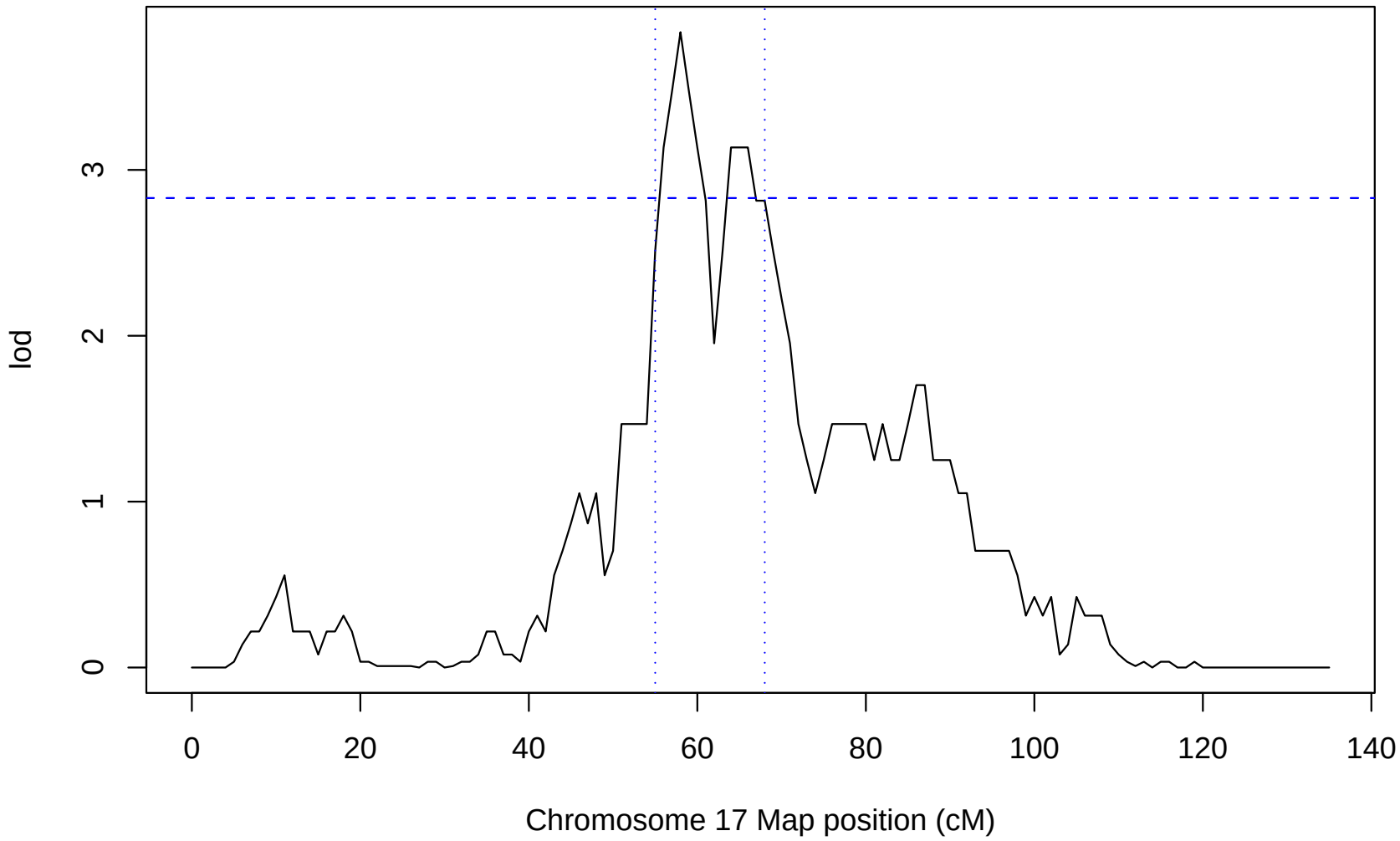
Multipoint linkage analysis simultaneously estimates the recombination fractions between multiple loci. Almost all modern linkage studies will involve multiple markers that can be combined to increase the power to detect linkage.

The usual type of analysis involves testing the position of a single test locus (which may be a marker or a trait locus) with respect to multiple marker loci whose positions are known. The resulting lod score is often called a **location score**.

Although a likelihood involving multiple **c**'s is being evaluated, these are then a function of the test locus position via the **mapping function**. For multipoint analysis, the Haldane function is often used, as strictly speaking, the Kosambi mapping function can be **multipoint inconsistent**.

It is known that multipoint linkage analyses are more sensitive to genotyping errors, so one will usually also carry out a **twopoint** analysis testing every marker in turn versus the test locus.

A multipoint lod score plot



QIMR

The Elston-Stewart algorithm for general pedigrees 1

The lod score formulae for larger pedigrees are difficult to generate and evaluate. This is especially the case where some pedigree members are untyped, or the relationship between phenotype and genotype is not the direct relationship of codominant loci.

Certain computer programs (actually computer algebra systems) can write out these high order polynomials, and then evaluate them.

The standard programs such as the LINKAGE programs (MLINK or ILINK), CRI-MAP, MENDEL, MERLIN, SUPERLINK and GENEHUNTER, do not produce a single closed form expression. They instead numerically evaluate the likelihood in a recursive fashion.

The Elston-Stewart algorithm for general pedigrees 2

For even large pedigrees that meet certain criteria (absence of loops, no more than one founder × founder mating), it is possible to write the likelihood in a form,

$$\mathbf{L(c) = \Sigma\ Pr(x_i|g_i)Pr(g_i|parents,c)\ldots}$$
$$\mathbf{\Sigma\ Pr(x_n|g_n)Pr(g_n|parents,c)}$$

where,

$\mathbf{x_i}$ is the phenotype of the **i**th individual,

$\mathbf{g_i}$ is the (poly-)genotype of the **i**th individual,

Pr($\mathbf{g_i}$|**parents**) is the probability of observing that genotype given the parental genotypes (the population genotype frequencies in the case of founders), and the recombination distance between the loci contributing to the genotype.

The Elston-Stewart algorithm for general pedigrees 3

$$L(c) = \Sigma \Pr(x_i|g_i)\Pr(g_i|\text{parents},c)\ldots \Sigma \Pr(x_n|g_n)\Pr(g_n|\text{parents},c)$$

The summation for each individual is over all possible genotypes consistent with their phenotype (eg two possibilities for the phase-unknown case, two codominant loci),

The individuals are ordered by their position in the pedigree, from founders downwards (to descendants).

The nested sums are evaluated from right to left, so the likelihood of the descendants below a particular individual become summarised in the likelihood of that individual.

General pedigree traversal analysis

For pedigrees where loops or multiple founder matings exist, the more complicated **pedigree traversal** algorithms used in the LINKAGE programs must be used. Given the complexity of evaluating the lod score for large pedigrees, values are usually produced for a grid of fixed values, such as **c**=(0.0,0.01,0.05,0.1,0.2…).

The **Lander-Green algorithm** is an alternative method of ordering the caclulations that is faster in the case of multipoint (more than 2 loci) linkage analysis for smaller pedigrees, but which is not usable in large pedigrees.

The program SUPERLINK tests a variety of different calculation orderings, picking the best approach for a given pedigree using the HUGIN algorithm.

Confidence intervals for the recombination fraction and multipoint location

The lod score or location score curve can also be used to give a confidence region for **c** or the trait location. The easiest method is the "1 unit" confidence interval or "support interval".

This is constructed by taking the closest values of **c** on either side of Z_{max} which have a lod score of Z_{max}-1.

The steepness of the lod curve around the MLE does reflect the precision of the estimate, and asymptotically, this steepness measured as the second derivative of the likelihood function gives the sampling variance of the estimate (as the inverse of the Fisher information).

Introducing the Generalized Single Major Locus Model

So far, we have dealt with **codominant** or **fully penetrant** loci, where there is a simple 1:1 relationship between the underlying genotype and the scored phenotype.

Modern marker loci are invariably *codominant*, but the trait loci that we wish to map are often more complex. For example, a genotype may give rise to a particular phenotype only in a proportion of individuals, and so must be described in a statistical manner.

The probability that a particular phenotype P will be observed in a individual of genotype G, $\Pr(P \mid G)$, is the **penetrance**.

The Generalized Single Major Locus Model

Consider a binary trait under the control of a two allele locus (alleles A and B). We can then write a description of the trait in the population:

Genotype	Frequency in Population (HWE)	Conditional probability, that an individual of that genotype is	
		Affected	Unaffected
A/A	P_A^2	f_2	$1-f_2$
A/B	$2P_A(1-P_A)$	f_1	$1-f_1$
B/B	$(1-P_A)^2$	f_0	$1-f_0$

Knowing the allele frequencies and penetrances, we can calculate the overall proportion of the population expressing the trait (affected),

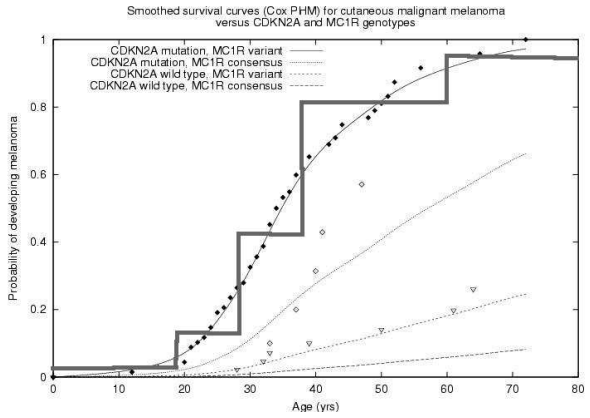
$$\text{Population Risk} = P_A^2 f_2 + 2P_A(1-P_A)f_1 + (1-P_A)^2 f_0$$

SML model with covariates

In the presence of covariates such as age or sex, the model is usually extended in a stratified fashion by defining **liability classes**, and defining penetrances for each liability class:

Sex	PenAA	PenAB	PenBB
Male	f_{0m}	f_{1m}	f_{2m}
Female	f_{0f}	f_{1f}	f_{2f}

For age, we stratify into bands and specify a step function to approximate the age-at-onset curve for each genotype.



QIMR

Estimating genotype carrier probabilities to allow linkage analysis

To carry out a **parametric linkage analysis**, we will use these SML **model parameters** to estimate the probability that an individual carries each of the possible genotypes.

Genotype	A/A	A/B	B/B
Probability	$P_A^2 f_2/R$	$2P_A (1-P_A) f_1/R$	$(1-P_A)^2 f_0/R$

- We must specify the SML model in order to carry out parametric linkage analysis
- The model does not have to be correct
- But power to detect linkage is best when the model is correct
- For complex diseases, fitting two models often covers most possibilities
- All “nonparametric” linkage models have a parametric equivalent

Non-parametric linkage analysis

If one of the loci of interest is not a simple Mendelian trait, then it becomes difficult to determine what the underlying genotypes are.

One approach is to take penetrance and allele frequency information from other sources, and use to those to estimate the probabilities of each genotype in each member of the pedigree.

Another is to perform simple tests looking for effects of ascertainment on segregation of the codominant marker locus in the selected families.

The affected sib pair (ASP) method

This method is used where a trait locus *A* is dichotomous (affected or unaffected), with unknown penetrances and allele frequencies for the underlying trait locus. The other locus *B* is a codominant marker (ideally). In this case, we ascertain families with two affected children. For backcross matings, we obtain

Sibship type, *Bb* x *BB* mating		Frequency of each sibship type	
Child 1	Child 2	Observed	Expected under null hypothesis
BB	BB	O_1	N/4
BB	bB	O_2	N/4
bB	BB	O_3	N/4
bB	bB	O_4	N/4
Total Number of Sibships		N	N

The null hypothesis is that there is no distortion of the segregation proportions due to linkage between the trait locus and the marker locus.

The affected sib pair (ASP) method 2

We can simplify this table to,

Sibship type	Number of families	
Children same type (both B, or both b)	O_1+O_4	$N/2$
Children different types	O_2+O_3	$N/2$
Total	N	N

Deviations in the expected counts from the null expectations occur when $c<0.5$.

The affected sib pair (ASP) method 3

We can work out theoretical expectations for particular values of *c*, penetrances (f_2, f_1, f_0) and allele frequencies (trait P_A and mmarker P_B). Assuming both Hardy-Weinberg and linkage equilibrium,

$$\Pr(\text{Children same type}) = 1/2 + (2c-1)^2(4c(c-1)(V_D-1)-1+2V_A+3V_D)/(16R+8V_A+4V_D)$$
$$\Pr(\text{Different}) = 1/2 - (2c-1)^2(-4c(c-1)(V_D-1)+1+2V_A+V_D)/(16R+8V_A+4V_D)$$

where,

$$R = P_A f_2 + 2P_A(1-P_A)f_1 + (1-P_A)^2 f_0,$$
$$V_A = 2P_A(1-P_A)(P_A(f_1-f_0)+(1-P_A)(f_2-f_1))^2$$
$$V_D = P_{A^2}(1-P_A)^2(f_2-2f_1+f_0)^2.$$

The affected sib pair (ASP) method 3

When *c* is 0.5, the second term disappears, giving the null expectations. If *c* is zero, then

Pr(Children same type)= $1/2 + (2V_A+3V_D-1)/(16R+8V_A+4V_D)$
Pr(Different)=$1/2 - (2V_A+V_D+1)/(16R+8V_A+4V_D)$.

In the case of a multiallelic marker, the test is exactly the same, the numbers for each heterozygous parent genotype still contributing to the sib pair being concordant or discordant at the marker.

Identity by descent and identity by state

In the backcross example above, the heterozygous parent is informative for linkage analysis, in that we can determine whether each child received an allele from the **same parental chromosome** (or **same grandparental gamete**).

This is termed **identity by descent** information. If each child received an allele from the same grandparental gamete, this allele is **identical by descent**.

If a parent is homozygous at the marker, each child receives the same allele, but we do not know whether these came from the same grandparent.

The term **identical by state** describes the situation where two relatives carry the same allele, regardless of whether it was inherited from a common ancestor or not.

Identity by descent and identity by state probabilities

In ambiguous cases, we will often calculate the **identity by descent probabilities**.

For example, if one parent is *BB* and the other *bb*, then the probability that both children carry the *B* allele is 100%. The probability that the *B* allele in one child is identical by descent with the *B* allele in the other child is 50%.

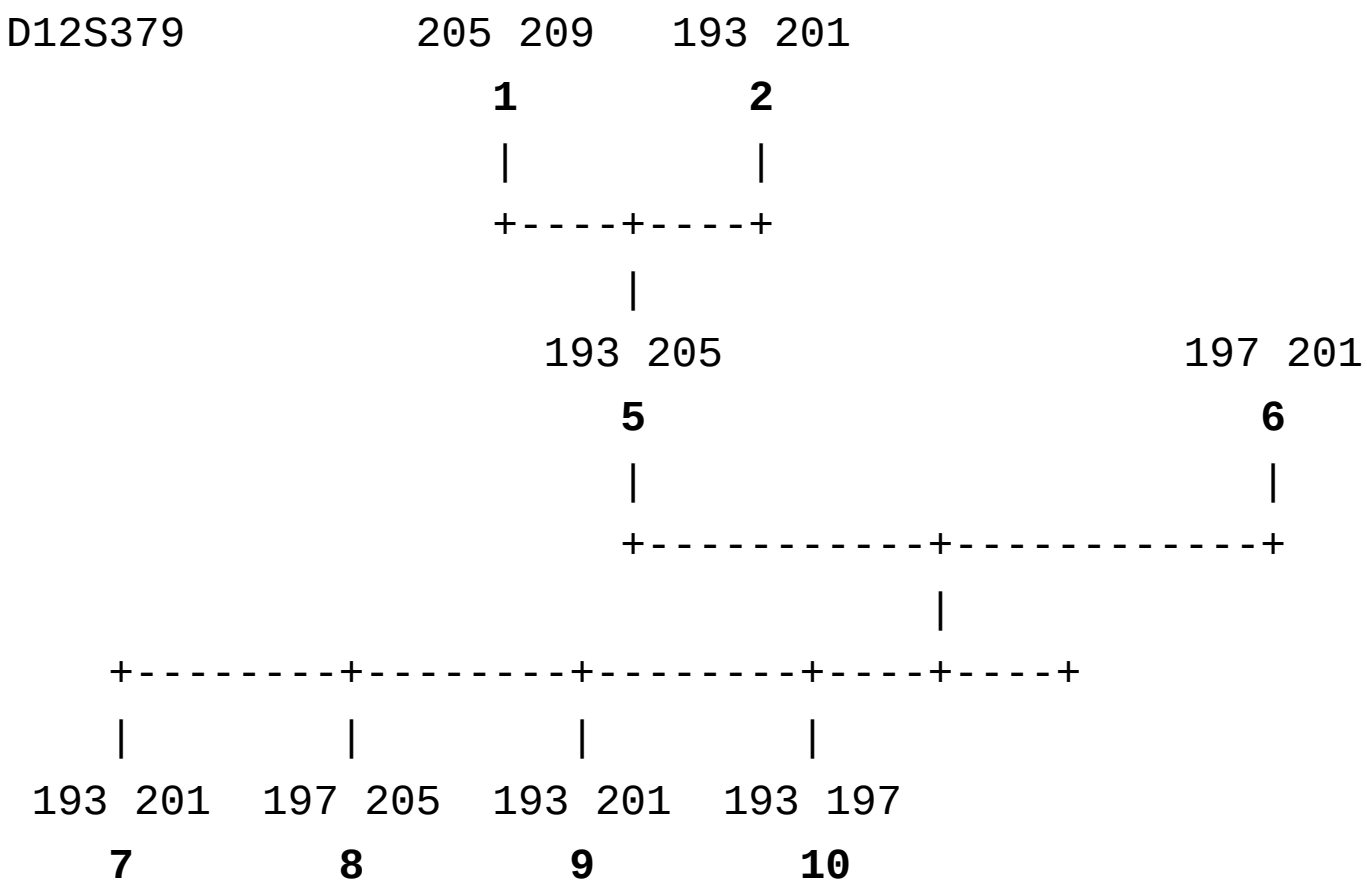
Identity by descent probabilities, or **ibd** are useful because:

- Can be calculated for any pair of relatives
- Can be estimated where one or both relatives is untyped at a marker
- Haplotype transmission in a pedigree is encoded by **ibd**
- The **ibd** probabilities are the empirical **kinship coefficients** for that locus, and any tightly linked trait loci

The **ibd** probabilities are often summarised as the mean probability of sharing an allele *ibd*

at that locus (the empirical coefficient of relationship or “pi-hat” – $\hat{\Pi}$). The set of these **ibd coefficients** for a pedigree is often represented as an **ibd** matrix.

Examples of IBD and IBS 1

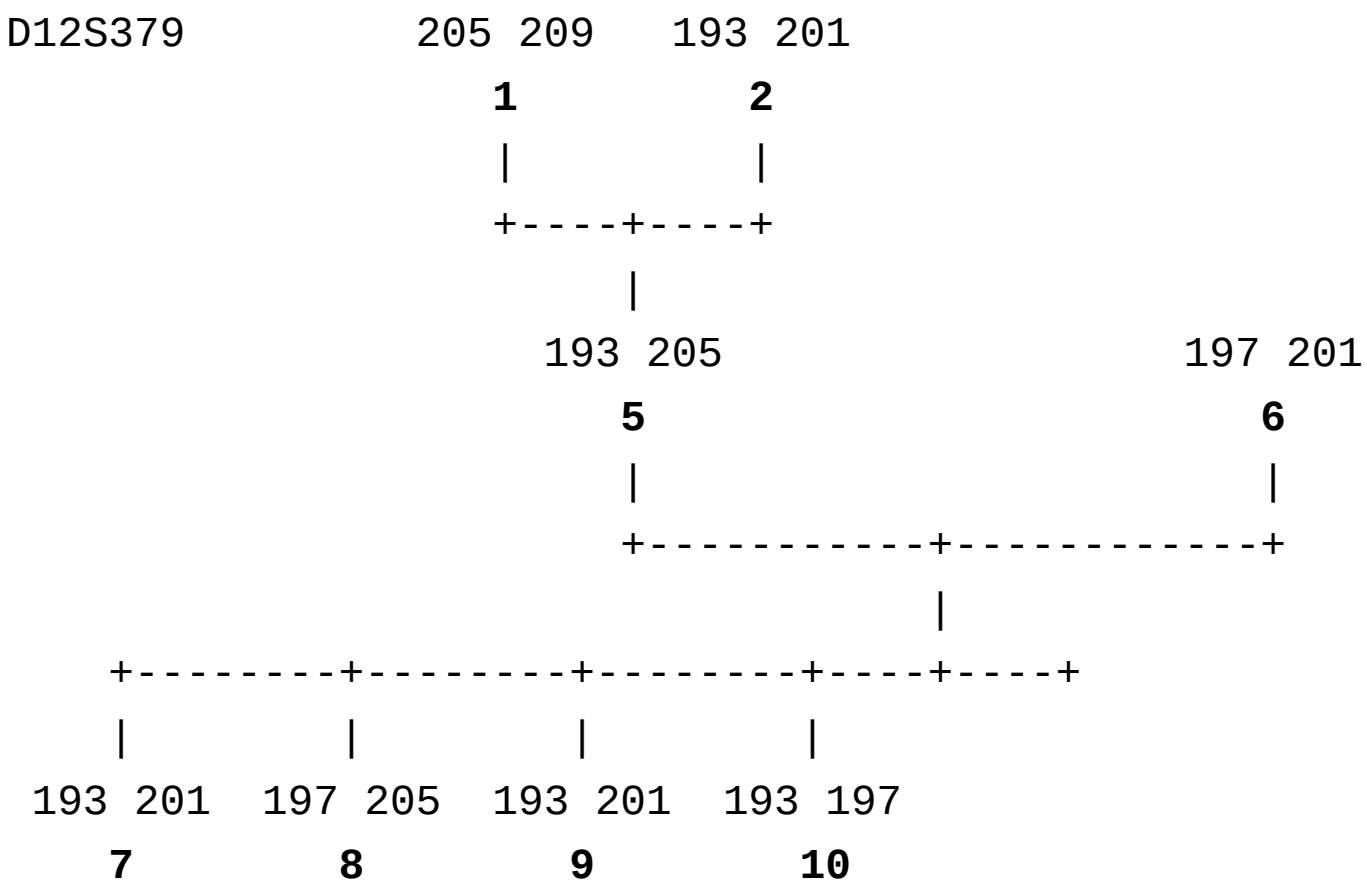


Here are some examples. Returning to the Amish pedigree above, individuals **2** and **7** both carry a 193 and a 201 (repeat) allele at the D12S379 locus.

QIMR

Therefore they share two alleles identical by state (*ibs*). However, the 201 allele was not transmitted from grandparent **2** to grandchild **7**, so they share only the 193 allele identical by descent (*ibd*). Grandchild **7** shares no alleles *ibs* or *ibd* with his/her grandparents **1** and **3** at D12S379.

Examples of IBD and IBS 2



For the first four siblings in the third generation, the *ibd* sharing is the same as the *ibs* sharing.

QIMR

	Individual **7**	Individual **8**	Individual **9**	Individual **10**
Individual **7**	-	0%	100%	50%
Individual **8**	0/2	-	0%	50%
Individual **9**	2/2	0/2	-	50%
Individual **10**	1/2	1/2	1/2	-

Estimating IBD for sib-pairs

Mating Type	Sib pair	Population frequency*	*ibd*=0%	*ibd*=50%	*ibd*=100%	Mean *ibd*
aa x aa	aa, aa	a^4	1/4	1/2	1/4	50%
aa x bb	ab, ab	$2a^2b^2$	1/4	1/2	1/4	50%
aa x ab	aa, aa	a^3b	0	1/2	1/2	75%
	aa, ab	$2a^3b$	1/2	1/2	0	25%
	ab, ab	a^3b	0	1/2	1/2	75%

Mating Type	Sib pair	Population frequency*	*ibd*=0%	*ibd*=50%	*ibd*=100%	Mean *ibd*
aa x bc	ab, ab *or* ac, ac	a^2bc	0	1/2	1/2	75%
	ab, ac	$2a^2bc$	1/2	1/2	0	25%
ab x ab	aa, aa *or* bb, bb	$a^2b^2/4$	0	0	1	100%
	aa, bb	$a^2b^2/2$	1	0	0	0%
	aa, ab *or* bb, ab	a^2b^2	0	1	50%	0
	ab, ab	a^2b^2	1/2	0	1/2	50%

Mating Type	Sib pair	Population frequency*	*ibd*=0%	*ibd*=50%	*ibd*=100%	Mean *ibd*
ab x ac	aa, aa	$a^2bc/2$	0	0	1	100%
	aa, ab *or* aa, ac	a^2bc	0	1	0	50%
	aa, bc	a^2bc	1	0	0	0%
	ab, ab *etc*	$a^2bc/2$	0	0	1	100%
	ab, ac	a^2bc	1	0	0	0%
	ab, bc	a^2bc	0	1	0	50%
	ac, bc	a^2bc	0	1	0	50%

Mating Type	Sib pair	Population frequency*	*ibd*=0%	*ibd*=50%	*ibd*=100%	Mean *ibd*
ab x cd	ac, ac *etc*	abcd/2	0	0	1	100%
	ac, ad *etc*	abcd	0	1	0	50%
	ac, bd *or* ad, bc	abcd	1	0	0	0%

* Population frequency of that type of family in the population assuming random mating and HWE. Each letter represents the population frequency of that allele in the general population.

Affected sib pairs with untyped parents

If a disease occurs late in life, both parents of an ASP are likely to be dead.

We can still work out the *ibd* probabilities for the sibs. If the marker is multiallelic, and the pair are *a*/*b* and *c*/*d*, for example, they must also be *ibd*=0. If we know the marker allele frequencies, and assume panmixia, HWE etc, we can obtain the expected *ibd*s by adding up the probabilities under each possible mating type that could give rise to that pair,

Sib pair	Population frequency*	*ibd*=0%	*ibd*=50%	*ibd*=100%	Mean *ibd*
aa, aa	$a^2(1+a)^2/4$	$a^2/(1+a)^2$	$2a/(1+a)^2$	$1/(1+a)^2$	$1/(1+a)$
aa, bb	$a^2b^2/2$	1	0	0	0
aa, ab	$a^2b(1+a)$	$a/(1+a)$	$1/(1+a)$	0	$1/(2+2a)$
aa, ac	a^2jk	1	0	0	0

Sib pair	Population frequency*	*ibd*=0%	*ibd*=50%	*ibd*=100%	Mean *ibd*
ab, ab	ab(1+a+b+2ab)/2	2ab/(1+a+b+2ab)	(a+b)/(1+a+b+2ab)	1/(1+a+b+2ab)	(2+a+b)/(2+2a+2b+4ab)
ab,ac	abc(1+2a)	2a/(1+2a)	1/(1+2a)	0	1/(2+4a)
ab,cd	2abcd	1	0	0	0

* Population frequency of that type of family in the population assuming random mating and HWE. Each letter represents the population frequency of that allele in the general population.

For example, if the *a* allele has a population frequency of 0.5, an ASP with genotypes *a*/*a* and *a*/*b* will contribute one-third of an observation to the *ibd*=0 cell, and two-thirds to the *ibd*=50% cell. The expected counts and the chi-square will be worked out in the usual way.

Faraway's improved (UMP) affected sib pair linkage test

We can therefore calculate *ibd* sharing for a sib-pair, or indeed any other kind of relative pair. If there is no inbreeding in the families sampled, the only kind of relative pair that can share more than one allele *ibd* (50% *ibd* sharing) is the sib pair (and MZ twins, but these contain **no linkage information**).

Using *ibd* sharing as the measure of similarity, there are actually three simple chisquare tests suggested for affected sib pair data in the following table.

	Identity by descent allele sharing			Total
	ibd=100%	ibd=50%	ibd=0%	
Observed Count	O_2	O_1	O_0	N
Expected Count	N/4	N/2	N/4	N

Note that there are "fractional" contributions from less informative families. For example, an ASP with genotypes *a*/*a* and *a*/*b* arising from the backcross *a*/*a* x *a*/*b* mating will contribute one-half of an observation to the *ibd*=0 cell, and one-half to the *ibd*=50% cell.

Faraway's improved (UMP) affected sib pair linkage test

We have already seen the overall best simple test, which is usually called the "mean" test,

Mean test = $2/N\ (2O_2+O_1-N)^2$

The other tests are superior only if the trait has particular mode of inheritance, such as a simple Mendelian recessive. The two-degree-of-freedom "genotypic" test is,

$X_{2^2} = [O_2 - N/4]^2 / [N/4] + [O_1 - N/2]^2 / [N/2] + [O_0 - N/4]^2 / [N/4]$

and the "two-allele" test is simply,

$X_{1^2} = [O_2 - N/4]^2 / [N/4]$

The “Possible Triangle” for IBD sharing

Faraway (1992) showed that a combination of these different tests is the theoretically best test against a genetic alternative hypothesis.

Observed identity by descent*	Value of composite statistic
$2p_2+p_1 > 1, p_1 > 1/2$	mean test
$3p_1/2 + p_2 < 1, p_2 > 1/4$	two-allele test
$2p_2+p_1 < 1, p_2 < 1/4$	Not consistent with genetic cause
Otherwise	2 d.f. chi-square

Here p_2,p_1,p_0 is the observed proportion of pairs sharing two, one, zero alleles ibd. Unfortunately, since one has to choose a different test for each situation, a correct P-value can no longer be looked up in the conventional chi-square table. For example, if your sample has 150 ASPs, the critical chi-square value for a one-tailed P=0.05 is not 2.71, but 3.42.

The “Possible Triangle” for IBD sharing

An equivalent test to this is the “MLS” ASP test, implemented in programs such as Genehunter, ASPEX and GAS.

MERLIN offers the mean test, parameterised as the Kong and Cox score test.

Other types of relative pair

We can easily construct similar tests for other types of relative pair. For example, if we have a set of families containing an affected individual and their affected grandparent, or two affected half-sibs, the expected *ibd* is 25% (or half an allele). The observed value will either be one or zero alleles shared *ibd*. For this case, we can use an approximate chi-square, or exact binomial test on the observed counts. Because there are more "intervening" relatives between the members of the grandparent-grandchild pair, there is more room for ambiguous cases to arise (the connecting parent needs to be heterozygous, *and* the grandparental contributions need to be identifiable ie different grandparental genotypes).

One type of affected relative pair linkage analysis is the Kong and Cox scoring approach. This is a maximum likelihood based approach, and is available in programs such as MERLIN.

Multipoint estimation of identity-by-descent sharing

Programs such as Allegro, Genehunter, Loki, MENDEL, MERLIN and SIMWALK2 use maximum likelihood approaches to improve the estimation of *ibd* probabilities when genotypes at multiple linked markers are available.

As in the case of multipoint linkage analysis, the *ibd* probabilities for all pairs of relatives in a pedigree can be evaluated at any location between (or indeed outside) the set of genotyped markers. One will usually evaluate *ibd* at the location of the markers themselves (where there is often maximal information), or on a fixed grid (every 1, 2, 5 or 10 cM along the map).

Risch's parameterisation for *ibd* based ASP analysis

One will often encounter the results and notation derived in Risch [1990], a paper that summarizes much earlier work on ASP analysis. The expected values under specific genetic hypotheses were quite complicated using V_A, V_D and *R*. Risch introduced some simpler formulae for the expected values.

The recurrence risk is the probability a family member will be affected (for a dichotomous trait) given that a specified relative is affected. For example, for a rare fully penetrant recessive gene (f_2=1, f_1=0, f_0=0), the recurrence risk to a sibling will be approximately 25%. James (1971) had shown that the recurrence risk was,

$$RecR = R + (k_1 V_A + k_2 V_D)/R$$

where k_1 and k_2 are kinship coefficients as before.

Risch's parameterisation for *ibd* based ASP analysis

If we define the *Population Relative Risk* (PRR) as *RecR*/*R*, then the expected *ibd* under a specific genetic hypothesis for a specific type of relative pair is,

	Identity by descent allele sharing		
	ibd=100%	ibd=50%	ibd=0%
Expected Prop	k_2PRR_{MZ}/PRR	k_1PRR_{PO}/PRR	k_0/PRR

PRR_{MZ} is the PRR for a monozygotic or identical twin of an affected individual, and PRR_{PO} is the PRR for the child of an affected parent. Therefore, if descriptive data about a trait is available, we can work out firstly how many families we will need in our study to get a significant chi-square (the power of the study), as well as detecting if a trait locus linked to our marker explains all the cases of disease in the population.

ASP Exclusion mapping

A third, related use is to perform *exclusion mapping*. If we specify *R*, PRR_{MZ} and PRR_{PO} we can test whether our observed *ibd* counts are significantly different from what they would be if the trait locus was close to our marker locus. If the chi-square is large enough, we can *exclude* the trait from being in that chromosomal region. This allows us to quantify how "non-significant" a small ASP chi-square value is, since a small chi-square can either arise from having a small study (not very powerful) or from the trait and marker locus being unlinked.