2.3.1 *ibd* OF MORE THAN TWO GENES:

Label $2k$ genes of k individuals sucessively, giving each the label previously assigned to genes to which it is *ibd*, and otherwise the next available integer.

$1\,2\;\;1\,3\;\;4\,4\;\;1\,5$: the paternal genes of individuals 1,2,4 are *ibd* and the two genes of individual 3 are *ibd*.
Reduce to genotypically equivalent classes of states:
$1\,2\;\;1\,3\;\;4\,4\;\;1\,5 \equiv 1\,2\;\;3\,1\;\;4\,4\;\;1\,5 \equiv 1\,2\;\;3\,1\;\;4\,4\;\;5\,1 \equiv$
$1\,2\;\;1\,3\;\;4\,4\;\;5\,1 \equiv 1\,2\;\;2\,3\;\;4\,4\;\;2\,5 \equiv 1\,2\;\;3\,2\;\;4\,4\;\;2\,5 \equiv$
$1\,2\;\;3\,2\;\;4\,4\;\;5\,2 \equiv 1\,2\;\;2\,3\;\;4\,4\;\;5\,2$
Note that when the two genes of the first individual are interchanged, we must relabel the genes $1 \leftrightarrow 2$, to obtain a legal state label.

The case of 4 genes of two individuals is shown in the Table: there are 15 states and 9 state classes.

For 12 genes in 6 individuals there are more than 4 million states, but only about 198,000 state classes.

2.3.2 Table for two individuals:

ibd pattern B_1 $p\ m$	*ibd* pattern B_2 $p\ m$	*ibd* label	*ibd* group	state description: individuals autozygous	state description: genes shared
● ●	● ●	1 1 1 1	1 1 1 1	B_1, B_2	4 genes *ibd*
● ●	● ○	1 1 1 2	1 1 1 2	B_1	3 genes *ibd*
● ●	○ ●	1 1 2 1			
● ○	● ●	1 2 1 1	1 2 1 1	B_2	3 genes *ibd*
● ○	○ ○	1 2 2 2			
● ●	○ ○	1 1 2 2	1 1 2 2	B_1, B_2	none
● ●	○ †	1 1 2 3	1 1 2 3	B_1	none
● ○	† †	1 2 3 3	1 2 3 3	B_2	none
● ○	● ○	1 2 1 2	1 2 1 2	none	2 genes
● ○	○ ●	1 2 2 1			shared
● ○	● †	1 2 1 3	1 2 1 3	none	1 gene
● ○	† ●	1 2 3 1			shared
● ○	○ †	1 2 2 3			
● ○	† ○	1 2 3 2			
● ○	† ⋆	1 2 3 4	1 2 3 4	none	none

2.3.3 *ibd* OF TWO NON-INBRED RELATIVES:

For two non-inbred relatives, 7 states, 3 classes, 2 probs

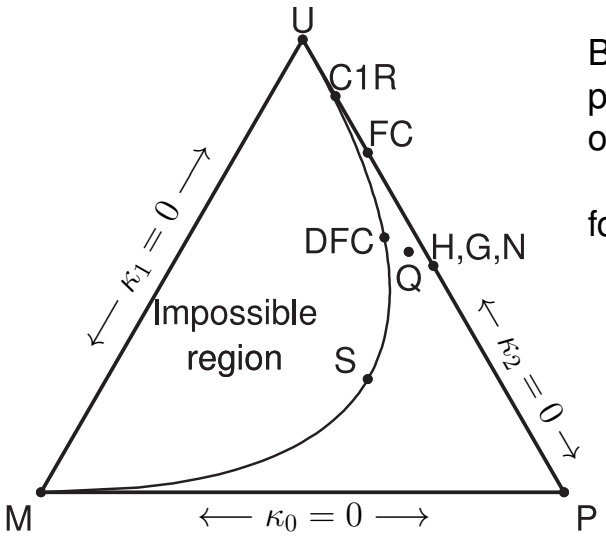
$$\kappa_i = \Pr(i \text{ genes } \textit{ibd}), \quad \kappa_2 + \kappa_1 + \kappa_0 = 1$$
$$\psi = \frac{1}{2}\kappa_2 + \frac{1}{4}\kappa_1 + 0\kappa_0 = \frac{1}{4}(2\kappa_2 + \kappa_1)$$
$$\text{If } \kappa_2 = 0, \quad \kappa_1 = 4\psi.$$

Pairwise relationship	κ_0	κ_1	κ_2	ψ
Unrelated (U)	1.00	0	0	0
Parent-offspring (P)	0	1.00	0	0.25
Monozygous twin (M)	0	0	1.00	0.50
Half sib (H),grandad (G),aunt (N)	0.50	0.50	0.00	0.125
Full Sib (S)	0.25	0.50	0.25	0.25
First cousin (FC)	0.75	0.25	0	0.0625
Double first cousin (DFC)	0.5625	0.375	0.0625	0.125
QHFC (Q)	0.5312	0.4375	0.0312	0.125

2.3.4 THE RELATIONSHIP TRIANGLE:

• Three numbers that sum to 1 ($\kappa_2 + \kappa_1 + \kappa_0 = 1$) can be represented as a point in an equilateral triangle of unit height.

• κ_i is the perpendicular distance from the side $\kappa_i = 0$, opposite the vertex $\kappa_i = 1$, $i = 0, 1, 2$



By direct computation of expressions using the equations on next page we can show
$\kappa_1^2 \geq 4\kappa_0\kappa_2$
for all real relationships.

2.3.5 COMPUTATION OF *ibd* PROBABILITIES:

- The following equations relate ψ and κ_i, $i = 0, 1, 2$.

$$\begin{aligned}
\psi &= (1/2)\kappa_2 + (1/4)\kappa_1 \;=\; (1/4)(1 + \kappa_2 - \kappa_0) \\
\psi &= (1/4)(\psi_{mm} + \psi_{mf} + \psi_{fm} + \psi_{ff}) \\
\kappa_2 &= \psi_{mm}\psi_{ff} + \psi_{mf}\psi_{fm} \quad \text{— new equation} \\
\kappa_1 &= 4\psi - 2\kappa_2, \quad \kappa_0 \;=\; 1 - \kappa_1 - \kappa_2
\end{aligned}$$

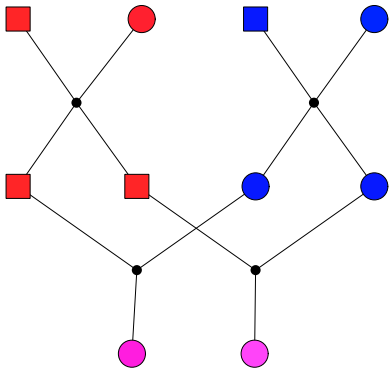
- Example: double first cousins:
  $\psi_{mm} = \psi_{ff} = 1/4$ and $\psi_{mf} = \psi_{fm} = 0$ or vv.
  $\kappa_2 = 1/16$, $\psi = 1/8$, so $\kappa_1 = 3/8$, $\kappa_0 = 9/16$.

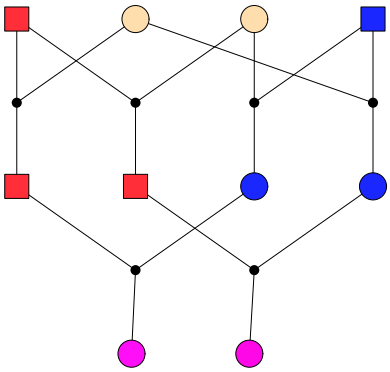
- The inequality:

$$\begin{aligned}
4\kappa_2 &= 4\psi_{mm}\psi_{ff} + 4\psi_{mf}\psi_{fm} \;\le\; (\psi_{mm} + \psi_{ff})^2 + (\psi_{mf} + \psi_{fm})^2 \\
&\le (\psi_{mm} + \psi_{ff} + \psi_{mf} + \psi_{fm})^2 \;=\; (4\psi)^2 \;=\; (2\kappa_2 + \kappa_1)^2 \\
4\kappa_2 &\le 4\kappa_2(\kappa_2 + \kappa_1) + \kappa_1^2 \quad \text{or} \quad 4\kappa_2(1 - \kappa_2 - \kappa_1) \le \kappa_1^2 \\
\text{So} \quad & 4\kappa_2\kappa_0 \;\le\; \kappa_1^2
\end{aligned}$$

with equality if and only if $\psi_{mm} = \psi_{ff}$ and $\psi_{mf} = \psi_{fm} = 0$ or vv.

2.3.6 DOUBLE COUSINS & QUAD HALF COUSINS:



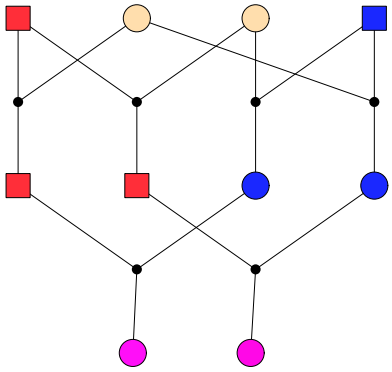


- Each shares 1/4 of her maternal and of her paternal genome *ibd* with the other individual (on average).

- For QHFC, each of the mom and dad of each individual is related to *both* the mom and the dad of the *other* individual, but mom is not related to dad.

- For DFC, probability of sharing maternal *and* paternal genome *ibd* with the other individual is $(1/4) \times (1/4) = 1/16$.
For QHFC this is 1/32.

2.3.7 EXAMPLE OF QUAD HALF FIRST COUSINS:



Each of the mother and the father of each child is related to both the mother and the father of the other. But, for each child, the mother is not related to the father.

Then all four of $\psi(M_1, M_2)$, $\psi(F_1, F_2)$, $\psi(M_1, F_2)$ and $\psi(F_1, M_2)$ are non-zero without the children being inbred.

For QHFC,
$\psi(M_1, M_2) = \psi(F_1, F_2) = \psi(M_1, F_2) = \psi(F_1, M_2) = 1/8$
so $\kappa_2 = 1/32$, $\psi = 1/8$, $\kappa_1 = 4\psi - 2\kappa_2 = 7/16$,
$\kappa_0 = 1 - \kappa_2 - \kappa_1 = 17/32$

2.4.1 DATA ON NON-INBRED RELATIVES:

• IDEA: given relationship $\mathcal{R}$, $\Pr(\mathbf{Y}|\mathcal{R}) = \sum_{\mathbf{J}} \Pr(\mathbf{Y} \mid \mathbf{J}) \Pr(\mathbf{J}|\mathcal{R})$ where $\mathbf{J}$ are all possible relevant patterns of *ibd*.

• EXAMPLE: one individual; 2 genes; 2 states—*ibd* or not;

$$\begin{aligned} \mathbf{J} &= (I, N), \quad \Pr(I) = f, \ \Pr(N) = 1 - f, \\ \Pr(AA) &= \Pr(AA|I)f + \Pr(AA|N)(1 - f) \\ &= qf + q^2(1 - f) = q^2 + fq(1 - q) \end{aligned}$$

• EXAMPLE: two non-inbred individuals; 3 states — 2, 1, or 0 *ibd*
$\mathbf{Y} = (G_1, G_2)$ = data on B_1, B_2. $\mathcal{R}$ = relationship: $\Pr(\mathbf{Y}|\mathcal{R})$

$$\begin{aligned} &= \kappa_0(\mathcal{R})\Pr(\mathbf{Y}|J_0) + \kappa_1(\mathcal{R})\Pr(\mathbf{Y}|J_1) + \kappa_2(\mathcal{R})\Pr(\mathbf{Y}|J_2) \\ &= \kappa_0(\mathcal{R})\Pr(\mathbf{Y}|\text{Unrel}) + \kappa_1(\mathcal{R})\Pr(\mathbf{Y}|\text{Par} - \text{offsp}) \\ &\quad + \kappa_2(\mathcal{R})\Pr(\mathbf{Y}|\text{MZ} - \text{twins}) \\ &= \kappa_0(\mathcal{R})\Pr(G_1)\Pr(G_2) + \kappa_1(\mathcal{R})\Pr(G_1)\Pr(kid = G_2|par = G_1) \\ &\quad + \kappa_2(\mathcal{R})\Pr(G_1)I(G_2 \equiv G_1) \end{aligned}$$

2.4.2 PARENT-OFFSPRING PROBABILITIES:

• Offspring should share allele with parent; provided there are no typing errors.

• Probabilities $\Pr(\text{child}|\ \text{parent})$: any number of alleles

parent genotype		Child's genotype A_iA_i	A_iA_j	A_iA_k	A_jA_k
A_iA_i	p_i^2	p_i	p_j	p_k	0
A_iA_j	$2p_ip_j$	$\frac{1}{2}p_i$	$\frac{1}{2}(p_i+p_j)$	$\frac{1}{2}p_k$	$\frac{1}{2}p_k$

• For markers with just 2 alleles:

parent geno.	$\Pr(\text{parent, child})$. child genotype AA	AB	BB	Data count child geno. AA	AB	BB
AA	q^3	$q^2(1-q)$	0	n_{00}	n_{01}	0
AB	$q^2(1-q)$	$q(1-q)$	$q(1-q)^2$	n_{10}	n_{11}	n_{12}
BB	0	$q(1-q)^2$	$(1-q)^3$	0	n_{21}	n_{22}

2.4.3 ESTIMATING q FROM DATA ON RELATIVES:

For simplicity we consider just mother-baby pairs and assume HWE.

$$\begin{aligned}
\ell &= \sum_{(i,j)} n_{ij} \log \Pr(G_i, G_j) \\
&= n_{00}\log(q^3) + n_{01}\log(q^2(1-q)) + n_{10}\log(q^2(1-q)) \\
&\quad + n_{11}\log(q(1-q)) + n_{12}\log(q(1-q)^2) \\
&\quad + n_{21}\log(q(1-q)^2) + n_{22}\log((1-q)^3) \\
&= (3n_{00} + 2(n_{01}+n_{10}) + n_{11} + n_{12} + n_{21})\log q + \\
&\quad (3n_{22} + 2(n_{21}+n_{12}) + n_{11} + n_{10} + n_{01})\log(1-q) \\
&= m_A \log q + m_B \log(1-q)
\end{aligned}$$

The MLE of q is $m_A/(m_A+m_B)$, where $(m_A+m_B) = 3n - n_{11}$ and $m_A = (3n_{00} + 2(n_{01}+n_{10}) + n_{11} + n_{12} + n_{21})$.

2.4.4 ALTERNATIVES TO THE MLE:

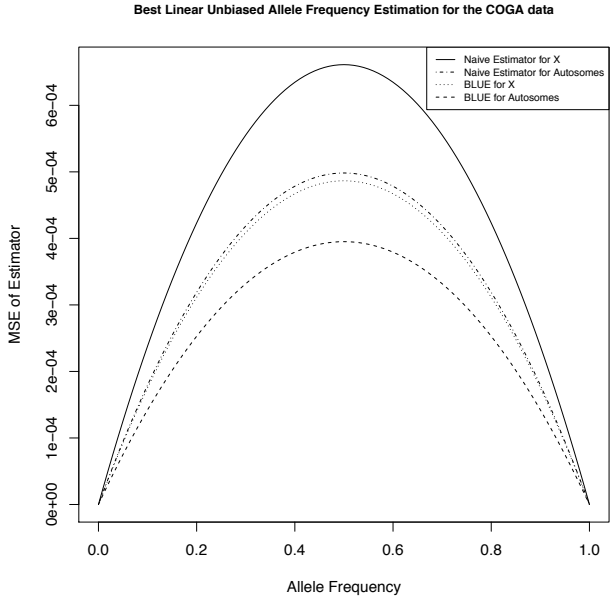
The MLE is "best", but there are simpler estimators that are not so bad.

(a) Use only founders (the moms):
estimate q by $(2n_{AA}+n_{AB})/2n$ where n_{AA} is number of AA moms, and n_{AB} is number of AB moms. ($n_{AA} = n_{00} + n_{01}$).

(b) Use everyone, disregarding relationship:
estimate q by $(2m_{AA} + m_{AB})/4n$, where m_{AA} is total number of AA individuals, and m_{AB} is total number of AB individuals. ($m_{AA} = 2n_{00} + n_{01} + n_{10}$).

These are both unbiased estimators, but asymptotically the MLE has smaller variance.

2.4.5 EFFECTS OF RELATEDNESS IN ESTIMATING q:



- Results due to Tim Thornton.
- COGA data set; $\sim$ 1500 individuals, in many pedigrees, $\sim$ 1000 observed.
- For the naive estimators we count alleles.
- For independent alleles the variance is $\sim q(1-q)/n$. All these curves are very close to $q(1-q)/n$ for some n, and we can think of n as the "effective sample size". (Larger variance $\equiv$ smaller n)

- Naive estimator; For X: eff-n =375 (125 female, 125 male)
For autosomal; eff-n = 500 (250 people) (Factor of 1/.75 = 1.33)
- For BLUE ($\sim$ MLE): For X: eff-n = 515 (approx) For autosomes: eff-n = 680 (Factor of 1/.75 = 1.33 for going from naive to BLUE)

2.4.6 SPECIFYING INHERITANCE:

• *Segregation* of genes is fully specified by *meiosis indicators*

$$\begin{aligned} S_i &= 0 \quad \text{if gene is parent's maternal gene} \\ &= 1 \quad \text{if gene is parent's paternal gene} \end{aligned}$$

where $i = 1, ..., m$ indexes the meioses.

• Mendel's First Law is S_i are independent with

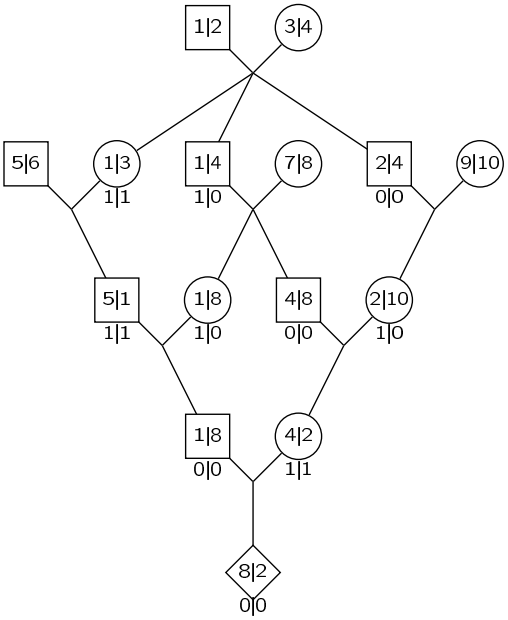
$$\Pr(S_i = 0) \;=\; \Pr(S_i = 1) \;=\; \frac{1}{2}.$$

• *ibd* state $\mathbf{J}$ at a locus is a function of the $\{S_i\}$ at that locus.

• If $\{S_i\}$ are known, then we know which founder genomes (FGL) descend to each individual.

2.4.7 Example showing descent of FGL:

• Consider the following segregation pattern of genes:



• Label the founder genomes.
• Use the $\{S_i\}$ to trace descent of FGL.
• Same FGL implies *ibd*.
• Example: The final individual and his maternal grandfather share FGL 8 – not by direct descent, but because both receive DNA from the founder who carries FGL 8.

2.4.8 The general formula for data probabilities:

$$\begin{aligned}\Pr(\mathbf{Y}) &= \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{S}) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{S}} \Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S})) \Pr(\mathbf{S}) \\ &= \sum_{\mathbf{J}} \Pr(\mathbf{Y} \mid \mathbf{J}) \Pr(\mathbf{J})\end{aligned}$$

$\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S}))$ is the sum over all possible assignments $\mathcal{A}$ of allelic types to non-*ibd* genes of the product of allele frequencies $q_{a(k)}$ of assigned alleles $a(k)$:

$$\Pr(\mathbf{Y} \mid \mathbf{J}(\mathbf{S})) = \sum_{\mathcal{A}} \prod_{k} q_{a(k)}.$$

• EXAMPLE: Mom-baby pairs: *ibd* state 1 2 1 3:
— Data AA, AB; 1 is A, 2 is A, 3 is B: prob $q^2(1-q)$
— Data AB, BB; 1 is B, 2 is A, 3 is B: prob $q(1-q)^2$
— Data AB, AB; 1 is A, 2 is B, 3 is B: prob $q(1-q)^2$
OR 1 is B, 2 is A, 3 is A: prob $q^2(1-q)$; sum $q(1-q)$.

2.4.9 EXAMPLE: DATA ON TWO INDIVIDUALS:

We know the relationship between two individuals, so can (we suppose) compute the probabilities $\Delta_1, \ldots, \Delta_9$ of the 9 *ibd* classes (groups of states). Suppose we observe the individuals to be AA and AC.

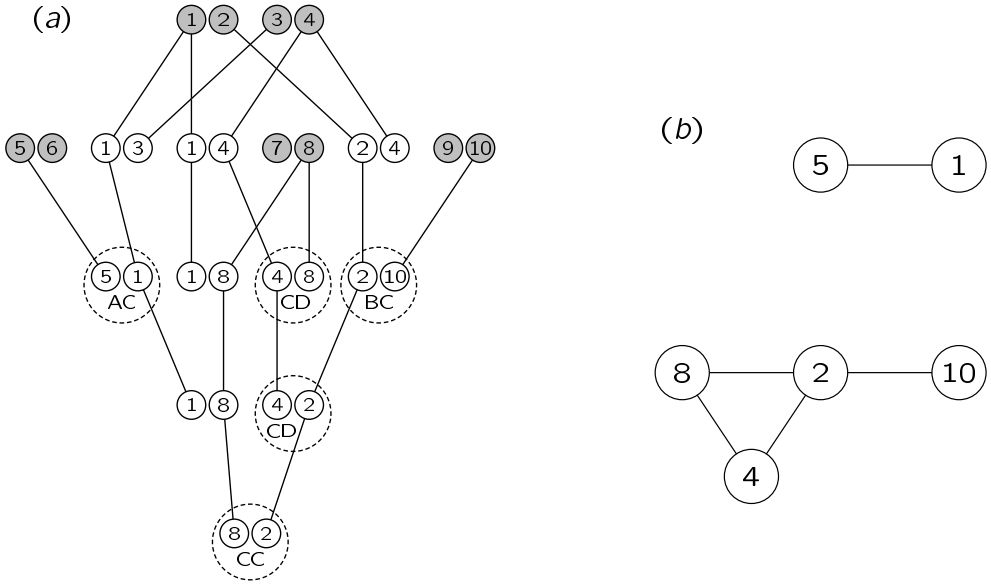
$P(\mathbf{J})$	$\mathbf{J}$	$P(AA, AC \mid \mathbf{J})$
Δ_1	1 1 1 1	0
Δ_2	1 1 1 2	$q_A q_C$
Δ_3	1 2 1 1	0
Δ_4	1 1 2 2	0
Δ_5	1 1 2 3	$q_A(2q_A q_C)$
Δ_6	1 2 3 3	0
Δ_7	1 2 1 2	0
Δ_8	1 2 1 3	$q_A q_A q_C$
Δ_9	1 2 3 4	$q_A^2.(2q_A q_C)$

Total probability of observing (AA, AC) is

$$P(AA, AC) = \Delta_2 q_A q_C + \Delta_5 2q_A^2 q_C + \Delta_8 q_A^2 q_C + \Delta_9 2q_A^3 q_C$$

2.4.10 Back to JV pedigree example :

• Given the particular descent pattern $\mathbf{S}$, consider the possible allelic types of these genes given the genotypes of 5 individuals shown:



• There are always 2, 1, or 0 possible assignments of allelic types to FGL nodes that are consistent with observed (no-error) genotypes.

Blank slide: