Workshop

Biostatistics in relationship testing
ESWG-meeting 2014

Daniel Kling

Norwegian Institute of Public Health and Norwegian University of Life Sciences, Norway

Andreas Tillmar

National Board of Forensic Medicine, Sweden

STATISTICAL METHODS IN RELATIONSHIP TESTING

PhD student Daniel Kling[1] and forensic geneticist Andreas Tillmar[2], PhD

[1]*Department of Family Genetics, Norwegian Institute of Public Health, Oslo, Norway*

[2]*Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden*

09.00-10.30	**Session 1** Introduction on how to calculate the LR in general How to calculate LR for a simple paternity case How to account for mutations, silent alleles, linkage and co-ancestry
10.30-11.00	***Coffee Break***
11.00-12.30	**Session 2** Training examples under supervision
12.30-13.45	***Lunch Break***
13.45-15.15	Review of tricky markers in ESWG Paper Challenge (2010-2014)
15.15-15.45	***Coffee Break***
15.45-17.00	**Session 3** How to calculate LR for X-chromosomal markers FamLinkX

http://www.famlink.se/BiostatWorkshopESWG2014.html

Outline of this first session

- Likelihood ratio principle
- Theoretical: paternity issues
  - Calculation of LR “by hand”.
  - Accounting for:
    - Mutations
    - Null/silent alleles
    - Linkage and linkage disequilibrium (allelic association)
    - Co-ancestry

DNA-testing to solve relationship issues

Is Mike the father of Brian?

DNA-data

	vWA	D12S391	D21S11	….
Mike:	12,13	22,22	31,32.2	….
Mother of Brian:	12,14	23,23.2	30,30	….
Brian:	13,14	22,23	30,31	….

Can we use the information from the DNA-data to answer the question?

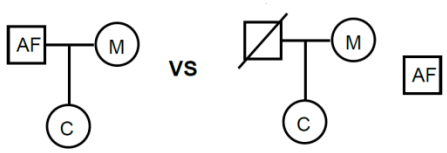
YES!

We estimate the "weight of evidence" by biostatistical calculations.

We will follow the recommendations given by Gjertson et al., 2007

"ISFG: Recommendations on biostatistics in paternity testing."

Calculating the "Weight of evidence" in a case with DNA-data



H_1: The alleged father is the father of the child

H_2: The alleged father is unrelated to the child

We wish to find the probability for H_1 to be true, given the evidence

= **$Pr(H_1|E)$**

Via Baye's theorem:

$$\frac{Pr(H_1 \mid E)}{Pr(H_2 \mid E)} = \frac{Pr(E \mid H_1)}{Pr(E \mid H_2)} \frac{Pr(H_1)}{Pr(H_2)}$$

Posterior odds | LR (or PI) | Prior odds

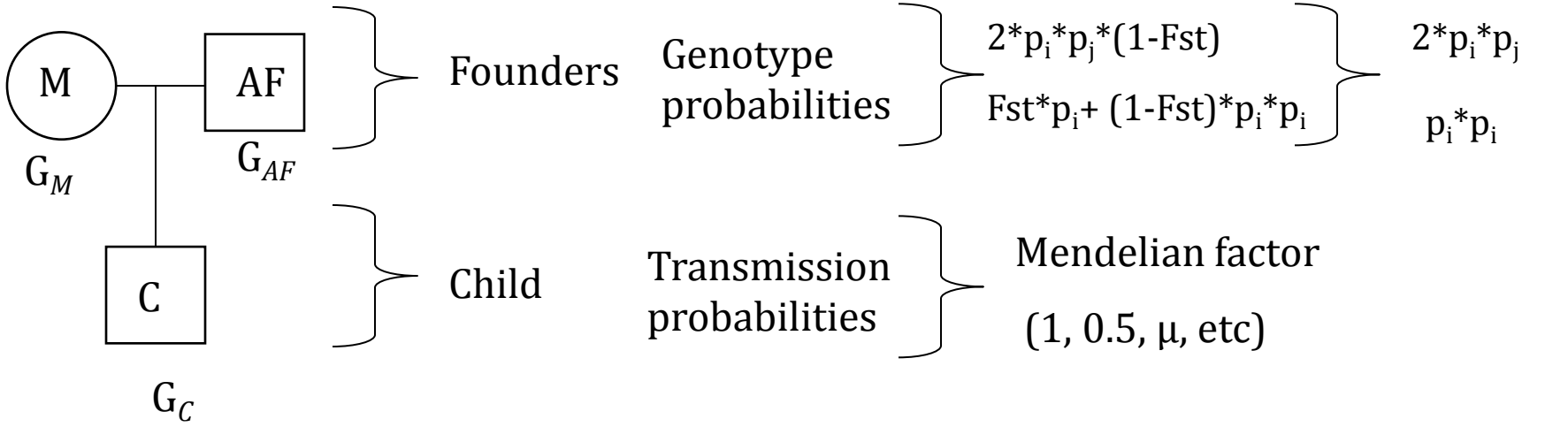
Today we focus on the likelihood ratio (LR)

$$\frac{Pr(DNA \mid H_1)}{Pr(DNA \mid H_2)}$$

$$\text{LR} = \frac{\Pr(DNA \mid H_1)}{\Pr(DNA \mid H_2)}$$

$$\Pr(DNA \mid H_1)$$

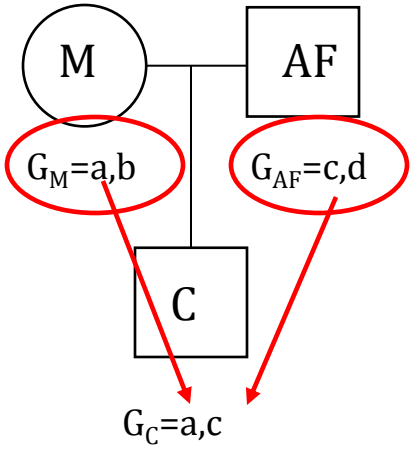
Assuming HWE



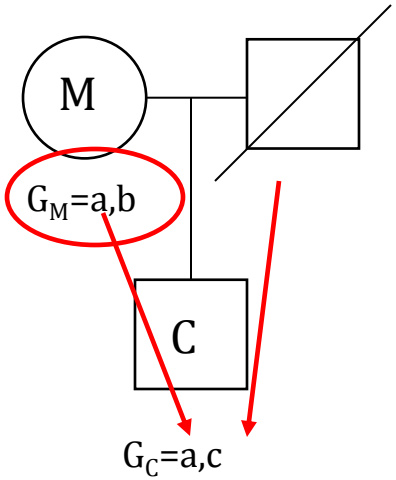
$$\Pr(DNA \mid H_1) = \Pr(G_C, G_{AF}, G_M \mid H_1) = \Pr(G_C \mid G_{AF}, G_M, H_1) \cdot \Pr(G_{AF}, G_M \mid H_1)$$

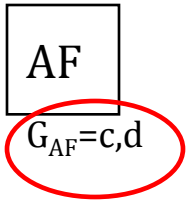
Paternity trio – An example

$$\Pr(DNA \mid H_1) = \Pr(G_C, G_{AF}, G_M \mid H_1) = \Pr(G_{AF}, G_M \mid H_1) \cdot \Pr(G_C \mid G_{AF}, G_M, H_1)$$



$$\Pr(DNA \mid H_1) = 2 \cdot p_a \cdot p_b \cdot 2 \cdot p_c \cdot p_d \cdot 0.5 \cdot 0.5$$





$$\Pr(DNA \mid H_2) \quad 2 \cdot p_a \cdot p_b \cdot 2 \cdot p_c \cdot p_d \cdot 0.5 \cdot p_c$$

Paternity trio – An example

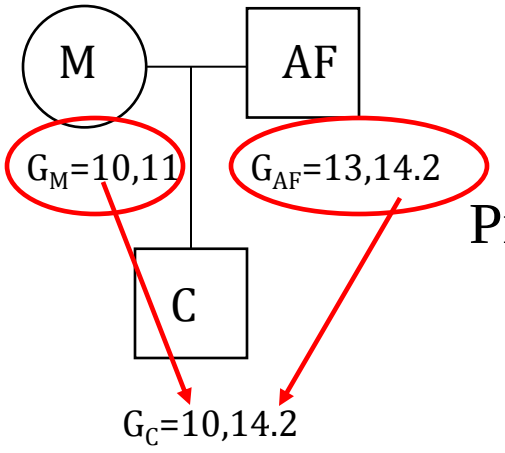
$$\text{LR} = \frac{\Pr(DNA \mid H_1) = 2 \cdot p_a \cdot p_b \cdot\ ] 2 \cdot p_c \cdot p_d \cdot\ ] 0.5 \cdot\ ] 0.5}{\Pr(DNA \mid H_2) = 2 \cdot p_a \cdot p_b \cdot\ ] 2 \cdot p_c \cdot p_d \cdot\ ] 0.5 \cdot\ ] p_c} =$$

$$= \frac{0.5}{p_c}$$

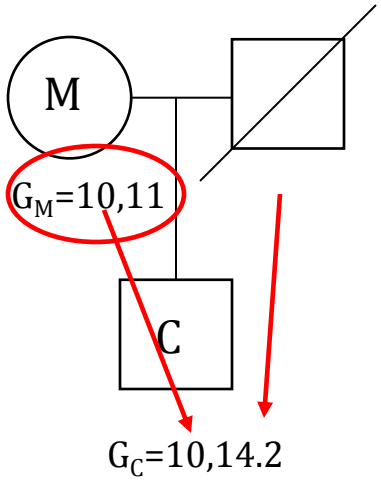
Probability to observe a certain allele in the population (e.g. population frequency)

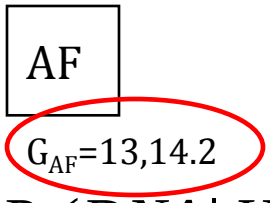
$$p_C \approx \frac{x_c + 1}{N + 1}$$

Paternity trio – real data



$$\Pr(DNA|H_1) = 2 \cdot p_{10} \cdot p_{11} \cdot 2 \cdot p_{13} \cdot p_{14.2} \cdot 0.5 \cdot 0.5$$

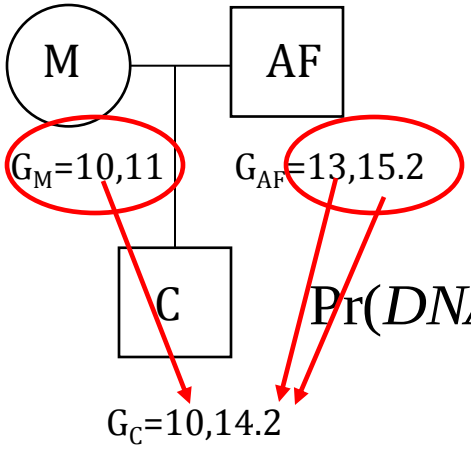




$$\Pr(DNA|H_2) \quad 2 \cdot p_{10} \cdot p_{11} \cdot 2 \cdot p_{13} \cdot p_{14.2} \cdot 0.5 \cdot p_{14.2}$$

$$\text{LR} = \frac{0.5}{p_{14.2}}$$

Paternity trio - Mutation



Mutation?

$$\Pr(DNA \mid H_1) = 2 \cdot p_{10} \cdot p_{11} \cdot 2 \cdot p_{13} \cdot p_{15.2} \cdot 0.5 \cdot$$

Mutations

"The possibility of mutation shall be taken into account whenever a genetic inconsistency is observed"

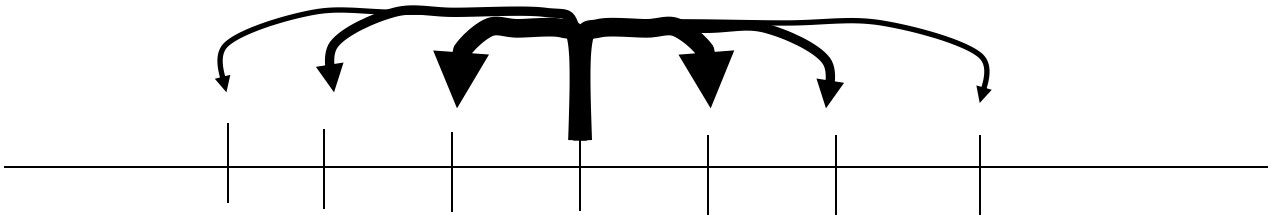
Brinkmann et al (1998) found 23 mutations in 10,844 parent/child offsprings. Out of these 22 were single step and 1 were two-step mutations.

AABB summary report covering ~4000 mutations showed approx 90-95% single step, 5-10% two-step and 1% or less more than two steps.

Note that these are the *observed* mutation rate, not the *actual* mutation rate (hidden mutations, assumption of mutation with the smallest step.

Mutation rate depend on marker, sex (female/male), age of individual, allele size

Mutations cont



Allele:	9	10	11	12	13	14	15
	-3	-2	-1	0	+1	+2	+3
				$1-\mu_{tot}$			
	μ_{-3}	μ_{-2}	μ_{-1}		μ_{+1}	μ_{+2}	μ_{+3}

$$\mu_{tot}=\mu_{+1}+\mu_{-1}+\mu_{+2}+\mu_{-2+}\ldots..$$

Approaches to calculate LR accounting for mutations:

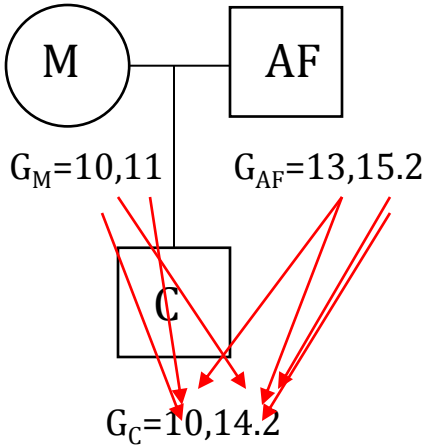
LR~(μ_{tot}*adj_steps)/p(paternal allele**)** *e.g "stepwise mutation model"*

LR~μ/A (A=average PE)

LR~μ

LR~μ/p(paternal allele)

Paternity trio - Mutation



$$(mut_{13->14.2} + mut_{15.2->14.2})$$

Mutation model decreasing with range

$$mut_{15.2->14.2}$$

Total mutation rate

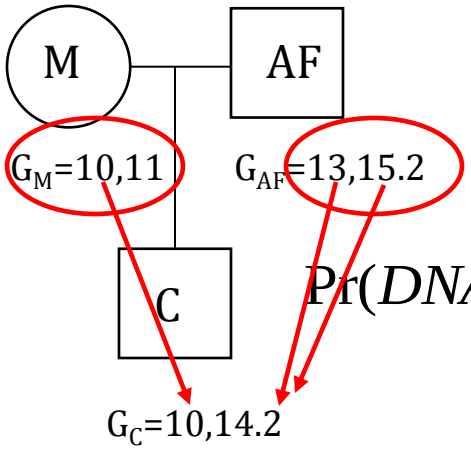
50% are loss of fragment size

$$mut_{15.2->14.2} = \{ \text{“Stepwise decreasing with range”} \} = 0.5 \cdot \mu_{Tot} \cdot 0.9 \cdot 0.5$$

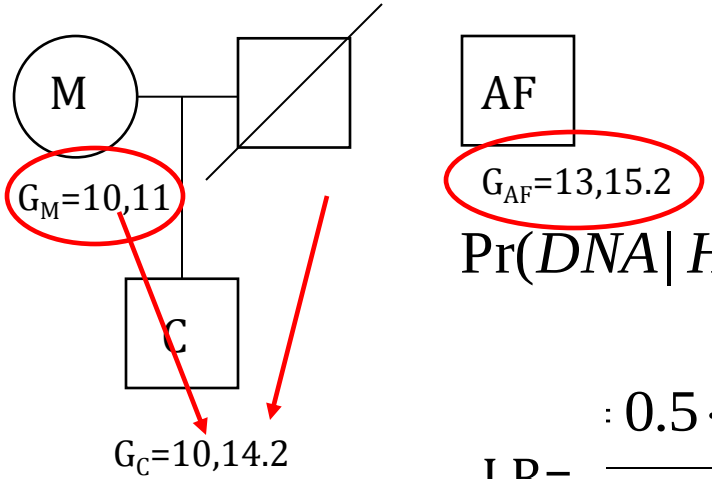
50% for 15.2 to be transmitted

90% of mutations are 1 step

Paternity trio - Mutation



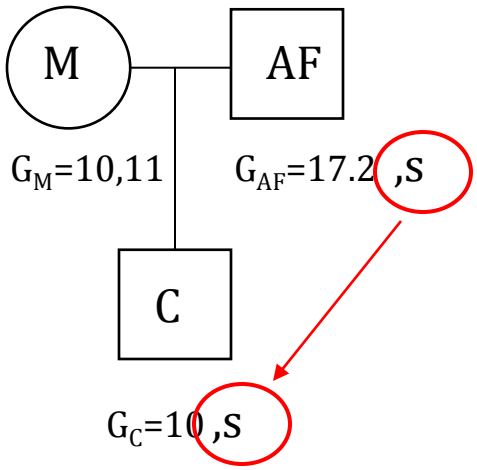
$$\Pr(DNA|H_1) = 2 \cdot p_{10} \cdot p_{11} \cdot 2 \cdot p_{13} \cdot p_{15.2} \cdot 0.5 \cdot \underbrace{(mut_{13->14.2} + mut_{15.2->14.2})}_{\approx 0.5 \cdot \mu_{Tot} \cdot 0.9 \cdot 0.5}$$



$$\Pr(DNA|H_2) \approx 2 \cdot p_{10} \cdot p_{11} \cdot 2 \cdot p_{13} \cdot p_{15.2} \cdot 0.5 \cdot p_{14.2}$$

$$LR = \frac{\approx 0.5 \cdot \mu_{Tot} \cdot 0.9 \cdot 0.5}{p_{14.2}}$$

Paternity trio – Silent allele



G_{AF}=17.2,17.2

G_{AF}=17.2, s

G_C=10,10

G_C=10,s

Null/silent allele

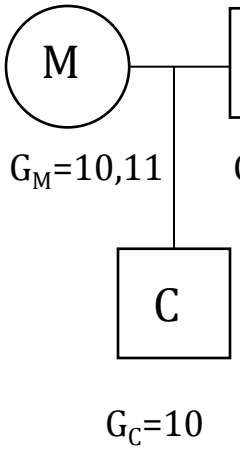
"STR null alleles mainly occur when a primer in the PCR reaction fails to hybridize to the template DNA..."

"Laboratories should try to resolve this situation with different pairs of STR primers"

"…no data from which to estimate the null allele frequency exists, consider a generous bracket of plausible values for the frequency and compute a corresponding range of values for the PI. If the resulting range of case-specific PI values are all large (as large as a laboratory's threshold value for issuing non-exclusion paternity reports), then report that the PI is ''at least greater than the smallest value'' in the range"

Impact of null/silent allele may vary on the pedigree and the genotype constellations

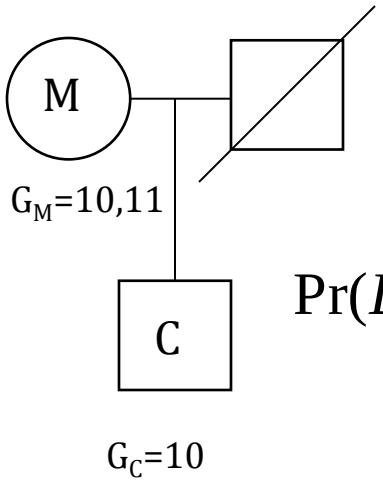
Paternity trio – Silent allele



G_{AF}=17.2,17.2 G_C=10,10

G_{AF}=17.2, s G_C=10,s

$$\Pr(DNA|H_1) = 2 \cdot p_{10} \cdot p_{11} \cdot \ 2 \cdot p_{17.2} \cdot p_s \cdot 0.5 \cdot 0.5$$



$$\Pr(DNA|H_2) \ 2 \cdot p_{10} \cdot p_{11} \cdot (2 \cdot p_{17.2} \cdot p_s + p_{17.2} \cdot p_{17.2}) \cdot 0.5 \cdot (p_{10} + p_s)$$

$$\text{LR} = \frac{p_{17.2} \cdot p_s}{(2 \cdot p_{17.2} \cdot p_s + p_{17.2} \cdot p_{17.2}) \cdot \ (p_{10} + p_s)}$$

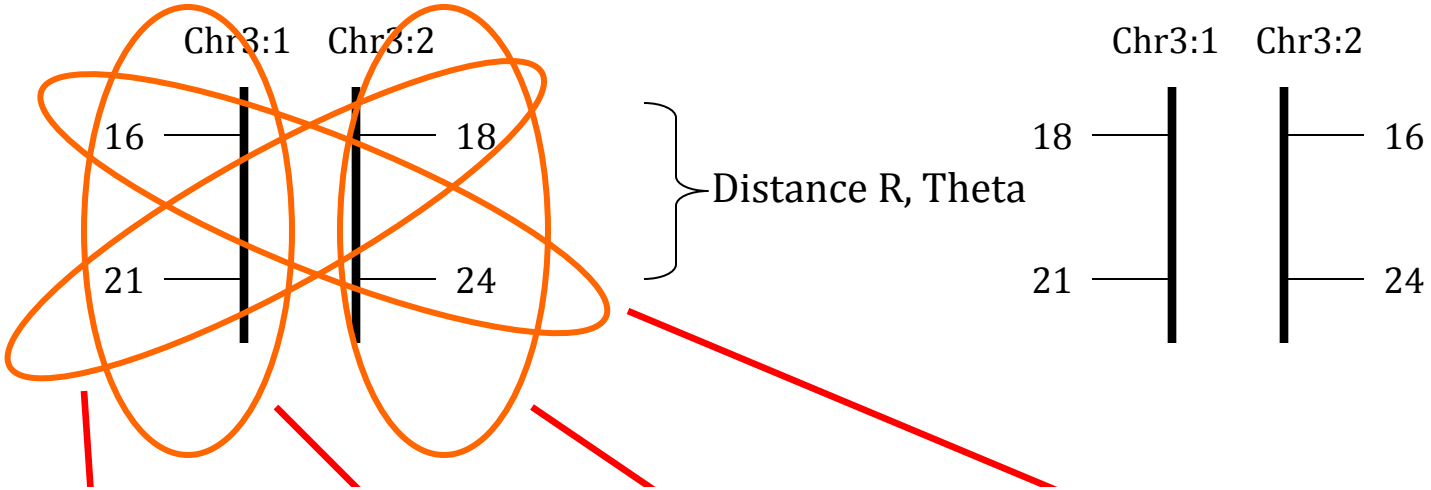
Linkage and Linkage disequilibrium

- Linkage
  - Can be described as the co-segregation of closely located loci within a family or pedigree.
  - **Effects the transmission probabilities!**

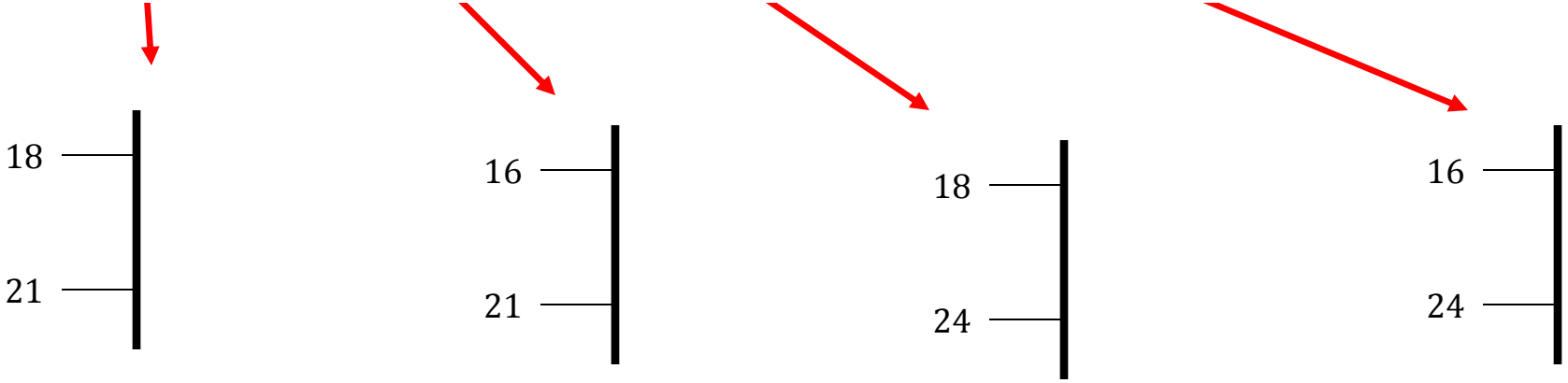
- Linkage disequilibrium (LD)
  - Allelic association.
  - Two alleles (at two different markers) which is observed more often/less often than can be expected.
  - **Effects the founder genotype probabilities, not the transmission probabilities!**

	Marker 1	Marker 2
Mother:	16,18	21,24

If **LD**, different probabilities for each phase



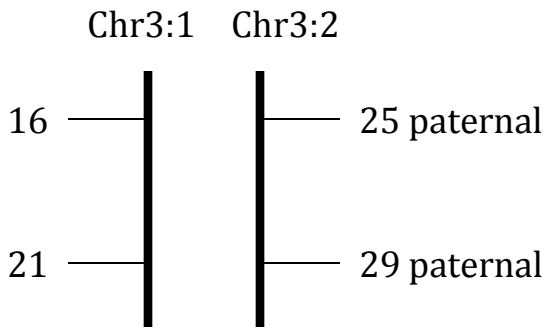
If **linkage**, different probability of transmission

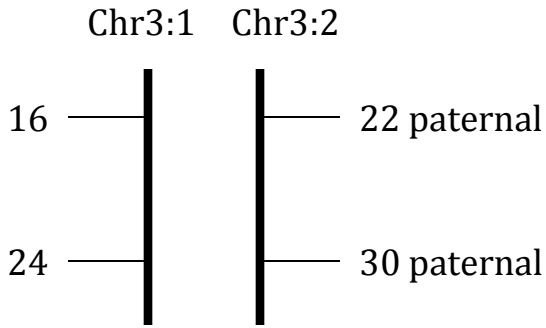


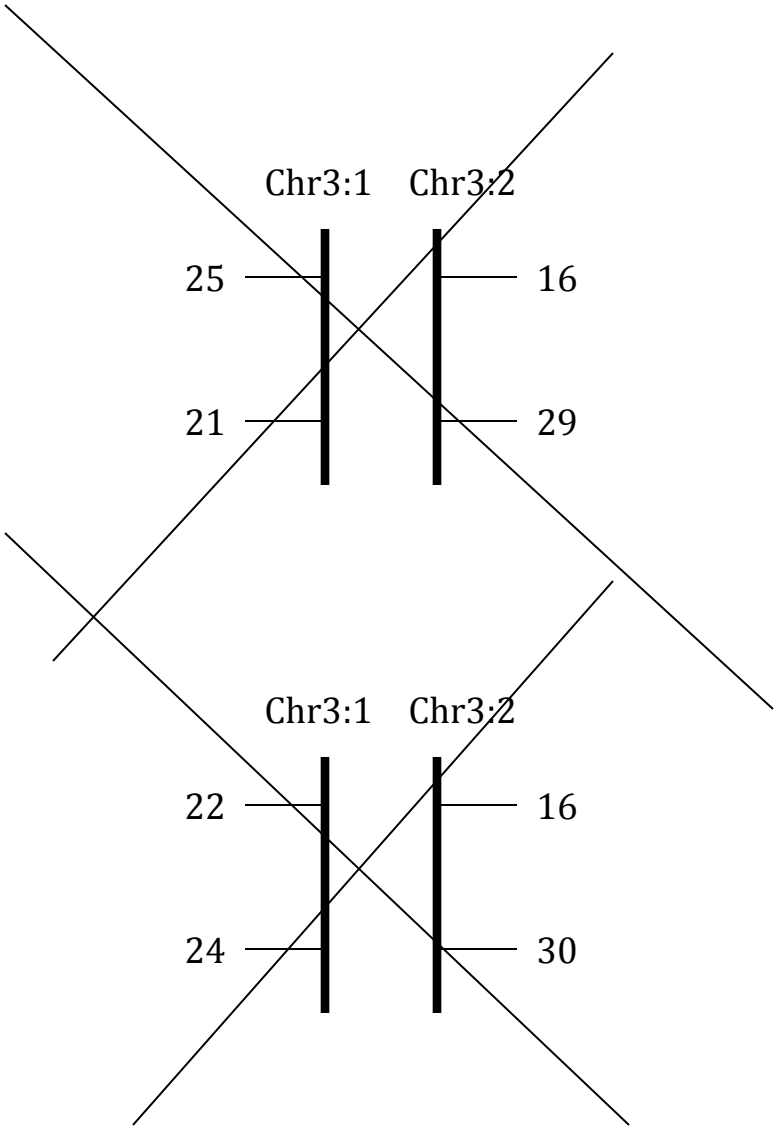
Mother: 16,18 21,24

Child1: 16,25 21,29

Child2: 16,22 24,30







Linkage and LD when it comes to forensic DNA-markers?

- 14 pair of autosomal STRs < 50cM
  - Phillips et al 2012
  - vWA-D12S391 (around 10cM, i.e. 10% recomb rate)
- Linkage will have an effect on LR for some pedigrees
  - Gill et al 2012, O'Connor & Tillmar 2012, Kling et al, 2012.
  - Incest.
  - Sibling, uncle-nephew.
  - Impact of linkage will depend on
    - Pedigree
    - Markers typed
    - Individuals' DNA data
- LE for vWA-D12S391!
  - Gill et al 2012, and O'Connor et al 2011.

Linkage cont.

- Assuming LE (no allelic association)
  - "linkage have an impact…"
- Thus…
  - For standard paternity cases (father vs unrelated), linkage has **no impact** on LR
  - BUT for paternity vs uncle (or other close related alternative), linkage has **impact** on LR

Presenting the evidence

- Likelihood ratio (LR) or Paternity index (PI)
- Posterior probability (must set priors)

$$W_i = \frac{\pi_i L_i}{\sum_{j=1}^{n} \pi_j L_j}$$

L_i=Likelihood under hypothesis *i*
π_i=Prior probability for hypothesis *i*

Simplifies to LR/LR+1,
assuming two hypotheses with equal priors

Subpopulation correction

- *What is subpopulation effects?*
  - *First developed by Wright in (1965)*
  - *Balding (1994)*
- *Hardy Weinberg equilibrium*
- *How to estimate it?*
  - *Reasonable values (0.01-0.05)*
- *Sampling formula (Fst = θ)*

$$p'(a_i) = \frac{Fst * n_{a_i} + (1 - Fst) * p(a_i)}{1 + (N_{Obs} - 1) * Fst}$$

Example 1. Genotype probabilities

- *HWE*
  - *Homozygouts: p(A)p(A)*
  - *Heterozygots: 2p(A)p(B)*
- *HWD*
  - *Homozygouts: Fst*p(A) + (1-Fst)*p(A)p(A)*
  - *Heterozygouts: (1-Fst)*p(A)p(B)*

$$P(A,A) = P(\text{Sampling two A:s}) = \frac{\theta * 0 + (1-\theta)\,p(A)}{1+(0-1)\theta} * \frac{\theta * 1 + (1-\theta)\,p(A)}{1+(1-1)\theta} = \theta\,p(A) + (1-\theta)\,p(A)^2$$

$$P(A,B) = P(\text{Sampling one A and one B}) = \frac{\theta * 0 + (1-\theta)\,p(A)}{1+(0-1)\theta} * \frac{\theta * 0 + (1-\theta)\,p(B)}{1+(1-1)\theta} = (1-\theta)\,p(A)\,p(B)$$

Example 2. Random Match Prob.

- H_1: The profiles G1 and G2 are from the same individual
- H_2: The profiles G1 and G2 are from different individuals

Let G1=A,A and G2=A,A

- HWE: $1/RMP = \frac{P(Data \mid H_1)}{P(Data \mid H_2)} = \frac{P(G1)}{P(G1)*P(G2)} = \frac{1}{p(A)^2}$

- HWD:

$$1/RMP = \frac{P(Data \mid H_1)}{P(Data \mid H_2)} = \frac{P(G1)}{P(G1,G2)} = \langle Sampling \rangle = \frac{P(\text{Sampling two A:s})}{P(\text{Sampling four A:s})} =$$

$$= \frac{(1+\theta)(1+2\theta)}{[2\theta+(1-\theta)p(A)][3\theta+(1-\theta)p(A)]}$$

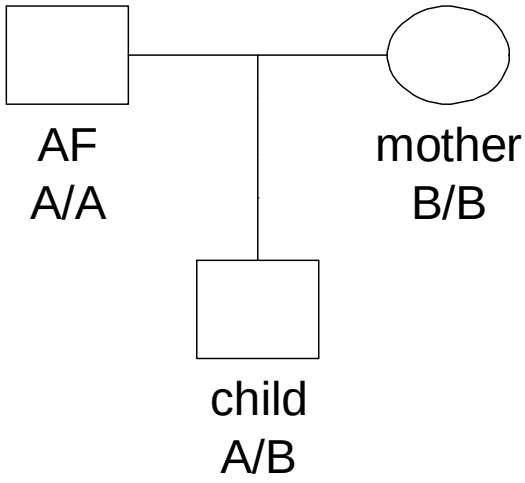
Example 3. Paternity

H_1: The alleged father (AF) is the real father

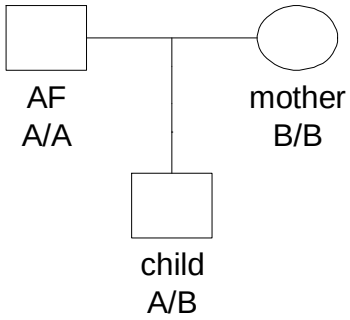
H_2: AF and the child are unrelated.

p(A) = 0.05

Standard paternity case. First marker



Standard paternity case. First marker



$$LR = \frac{\text{probability of data } \textit{given AF} \text{ father}}{\text{probability of data } \textit{given AF} \text{ unrelated}}$$

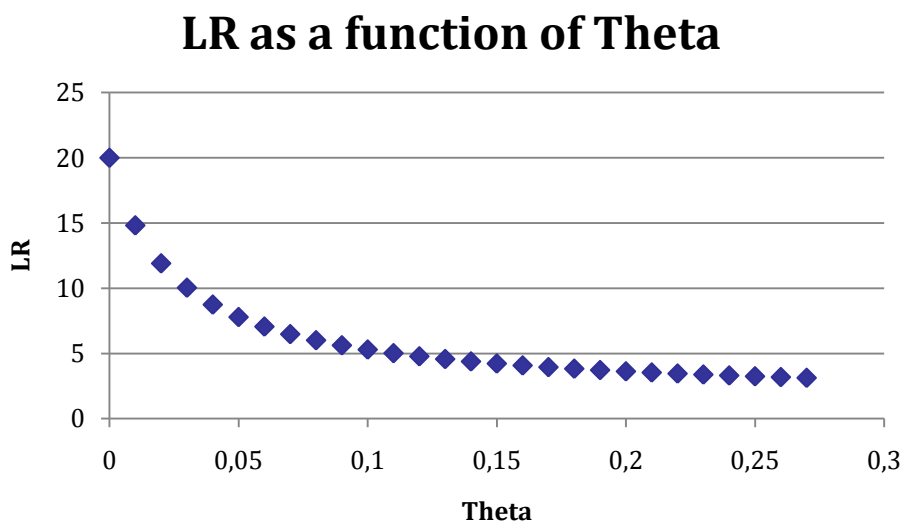
$$= \frac{P(child \mid mother, AF)}{P(child \mid mother)} = \langle \theta = 0 \rangle = \frac{1}{p_A} = \frac{1}{0.05} = 20.$$

$$LR = \frac{P(mother)P(AF)P(child \mid mother, AF)}{P(AF)P(mother)P(child \mid mother)} = \langle \theta \neq 0 \rangle = \frac{P(\text{Sampling two A:s and two B:s})}{P(\text{Sampling three A:s and two B:s})} =$$

$$= \frac{1}{P(\text{Sampling the third A})} = \frac{1}{\frac{2\theta + (1-\theta)p_A}{1+(4-1)\theta}} = \frac{1+3\theta}{2\theta + (1-\theta)p_A}$$

$$LR = \frac{P(mother)P(AF)P(child \mid mother, AF)}{P(AF)P(mother)P(child \mid mother)} = \langle \theta \neq 0 \rangle = \frac{P(\text{Sampling two A:s and two B:s})}{P(\text{Sampling three A:s and two B:s})} =$$

$$= \frac{1}{P(\text{Sampling the third A})} = \frac{1}{\frac{2\theta + (1-\theta)p_A}{1+(4-1)\theta}} = \frac{1+3\theta}{2\theta + (1-\theta)p_A}$$



Example 4. Siblings

H0: Two persons are full siblings

H1: The same two persons are unrelated

- *The two persons are homozygous A,A (p(A)=0.05)*
- *We use the "IBD method"*

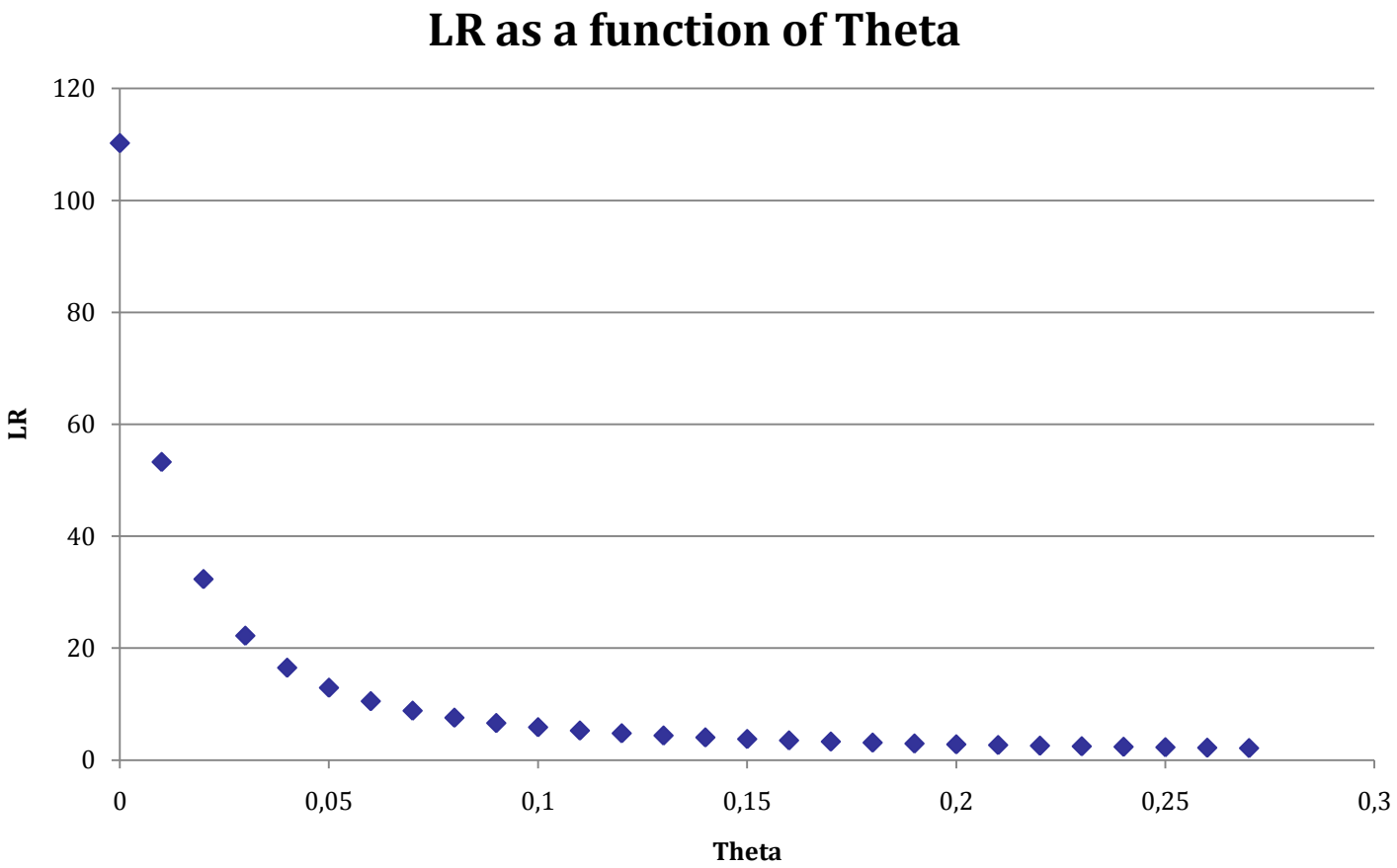
$$LR = \frac{P(Data \mid H0)}{P(Data \mid H1)} = \langle \theta \neq 0 \rangle =$$

$$= \frac{P(\text{Sampling two A:s})*P(\text{IBD=2}|\text{H0})+P(\text{Sampling three A:s})*P(\text{IBD=1}|\text{H0})+P(\text{Sampling four A:s})*P(\text{IBD=0}|\text{H0})}{P(\text{Sampling four A:s})} =$$

$$= \frac{0.25 + 0.5P(\text{Sampling the third A}) + 0.25P(\text{Sampling the last two A:s})}{P(\text{Sampling the last two A:s})} =$$

$$= \frac{0.25 + 0.5\frac{2\theta + (1-\theta)p_A}{1+(2-1)\theta} + 0.25\frac{2\theta + (1-\theta)p_A}{1+(2-1)\theta} * \frac{3\theta + (1-\theta)p_A}{1+(3-1)\theta}}{\frac{2\theta + (1-\theta)p_A}{1+(2-1)\theta} * \frac{3\theta + (1-\theta)p_A}{1+(3-1)\theta}}$$

Example 4. Siblings



Example 5. General relationships

- *We have N number of founder genotypes*
- *If Fst!=0 these are not independent*
- *We use the sampling formula to calculate updated allele frequencies*
- *It is more complex to do by hand*
  - In simple pairwise relationships use the IBD method (See Example 4)
  - Otherwise, use validated software!