X-chromosomal markers in Forensic Genetics

GHEP 2025 Virtual workshop series,
March 10,17 and 24th

Daniel Kling and Andreas Tillmar

Teachers

Daniel Kling, PhD

- Forensic Expert
- National Board of Forensic Medicine, Sweden
- Worked in the field for 13 years
- Developer of Familias, FamLink and FamLinkX
- Applied biostatistics, relationship inference, genetic genealogy

Andreas Tillmar, PhD

- Forensic geneticist & Associate professor
- National Board of Forensic Medicine, Sweden and Linköping University, Sweden
- Worked in the field for over 15 years
- Technical leadership mixed with R&D
- Applied biostatistics, relationship inference, population genetics, genetic genealogy.
- Lead author of the ISFG Commission on X-chromosomal testing

Presentations, exercises etc are available at https://familias.name/GHEP2025/

Session 1 – Basics (March 10)

16:00	Introduction
16:15-17:00	Basics of kinship testing and the utility of X-chromosomal markers
17:00-17:10	Short break
17:10-18:00	Software: FamLinkX
18:05-18:40	Exercises
18:40-19:00	Summary

Session 2 – Advanced (March 17)

16:00	Introduction
16:15-17:00	Advanced theory
17:00-17:10	Short break
17:10-18:00	Haplotypes and databases
18:05-18:40	Exercises
18:40-19:00	Summary

Session 3 – Applications and examples (March 24)

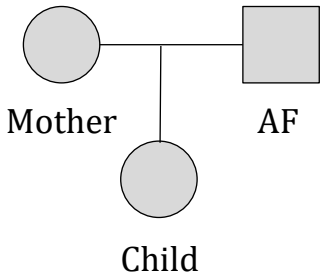
16:00	Introduction
16:15-17:00	Summary of theory and some more advanced topics
17:00-17:10	Short break
17:10-18:00	Examples
18:05-18:40	Exercises
18:40-19:00	Summary

Solving relationship issues with DNA data

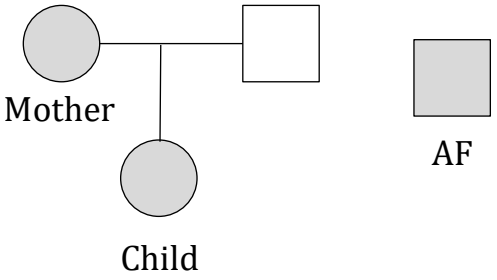
Legal situations: (e.g.) paternity, immigration, missing person identification, criminal acts (incest, human trafficking), investigative leads and more

Example 1 ”Simple” question

H_1: AF is the father of the child



H_2: AF is unrelated to the child

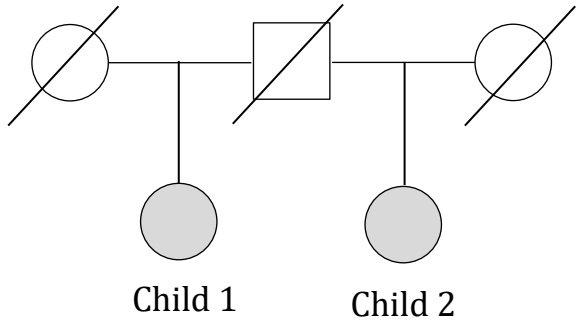


Question: Is AF the biological father of the child?

Genetic data: 15-21 autosomal STRs

Example 2 ”More complex” question

H_1: Child 1 and 2 have the same father



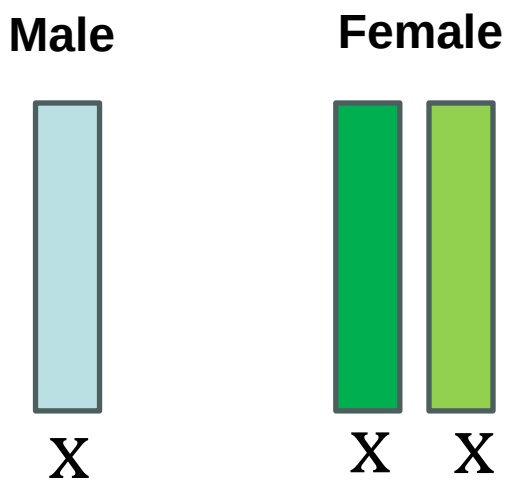
H_2: Child 1 and 2 are unrelated

Question: Are child 1 and child 2 paternal half-sibs, or unrelated

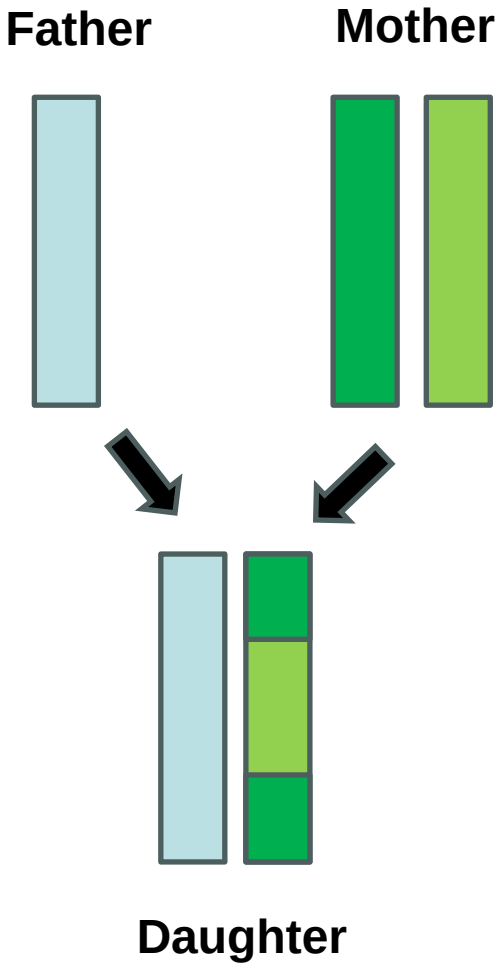
Genetic data: 4-12 X-STRs

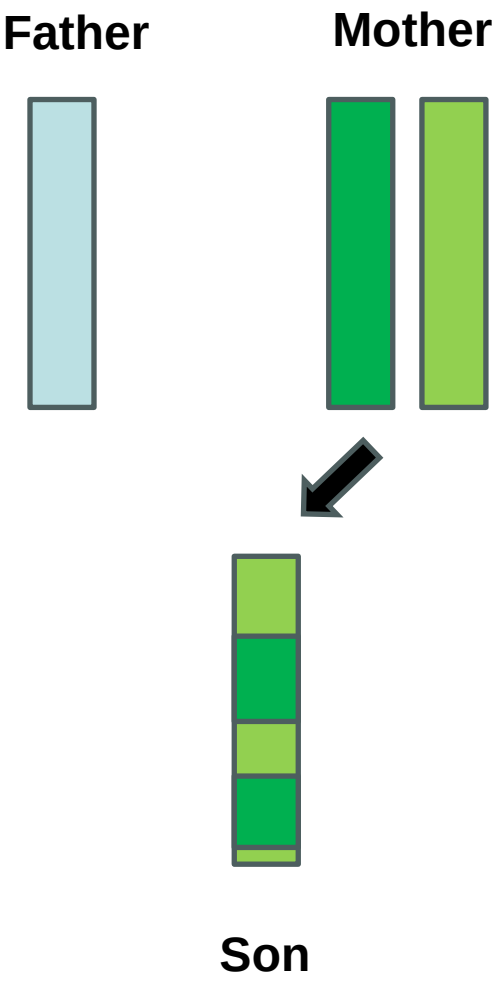
X chromosome in humans

- A female has two X chromosomes
- A male has one X chromosome
- In rare occasions other variations may exist,
  XXY (Klinefelter), X0 (Turner), XXX (Triple X), XYY



X-chromosomal inheritance pattern

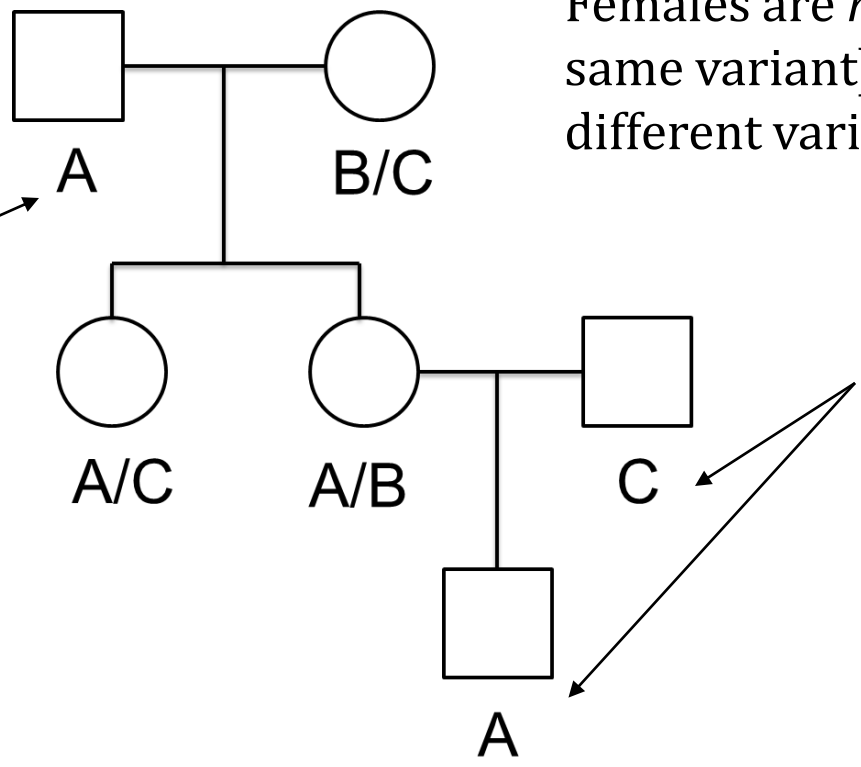




Inheritance pattern (one X locus)

Males are *hemizygous*

Females are *homozygous* (two alleles of the same variant) *or heterozygous* (two alleles of different variants)



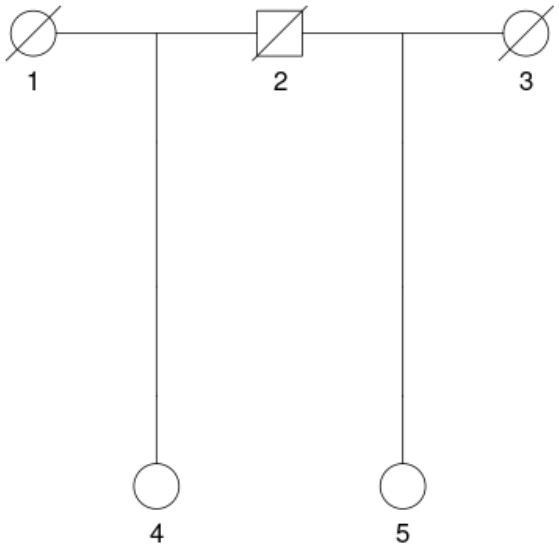
Forced transmission from father to daughter

No transmission from father to son

Inheritance pattern makes X-chromosomal analysis more (or less) informative compared with autosomal DNA analysis

Generally more informative

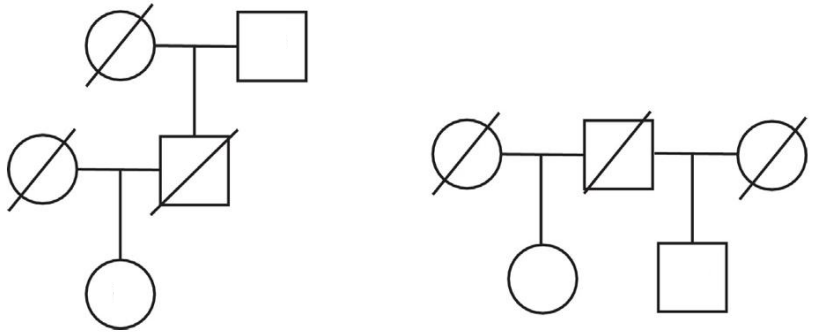
- Paternal half-sisters vs unrelated
- Paternal grandmother/granddaughter vs unrelated
- For many pedigress, the exclusion probability is not null



See Pinto et al., 2011

Generally less informative

- Father/son vs unrelated
- Paternal grandfather/grandson vs unrelated
- Paternal halfbrothers vs unrelated

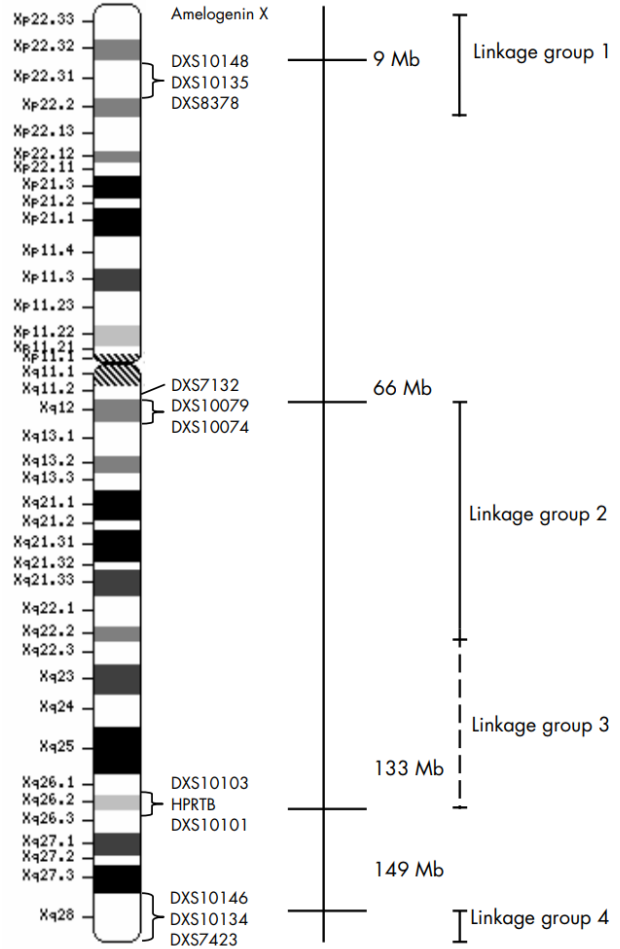


Tillmar et al., 2017

Two common X-chromosomal marker panels

- STRs (short tandem repeats)
- **"Decaplex"**
  - 10 X STRs, in genetic linkage but mostly not in linkage disequilibrium (LD, allelic association).
  - Developed by GEP-ISFG (Gusmao et al., 2009)

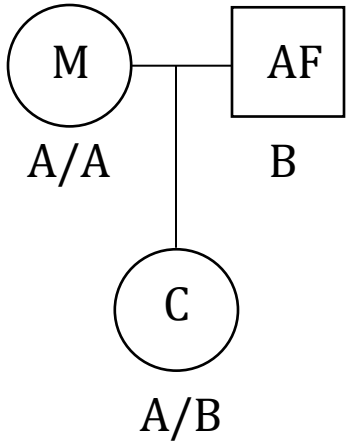
- **Argus X-12QS**
  - 12 X STRs, in four "linkage groups", in genetic linkage but mostly not in linkage disequilibrium (LD, allelic association).
  - Investigator Argus X-12 QS (Qiagen)



Basic notations:

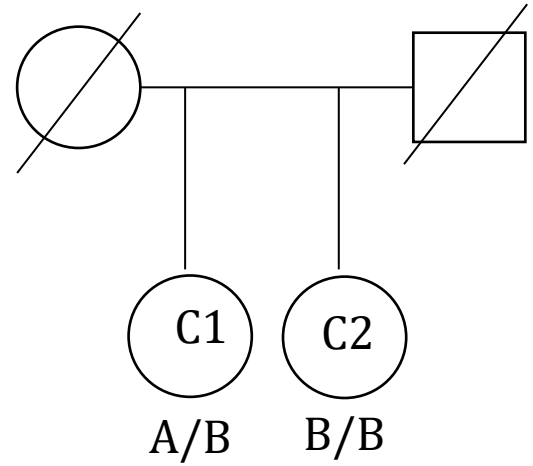
Genetic markers, alleles and pedigrees

- Consider one X-chromosomal genetic marker (e.g. STR) with possible alleles A, B and C
- Consider a family "trio" (mother, alleged father and a male child)



In this pedigree:

- Circles represent females
- Squares represent males
- M is the biological mother of C
- AF is the biological father of C
- M and AF are unrelated
- The genotype of the mother is A/A (homozygous)
- The genotype of the alleged father is A (hemizygous)
- The genotype of the child is A/B (heterozygous)

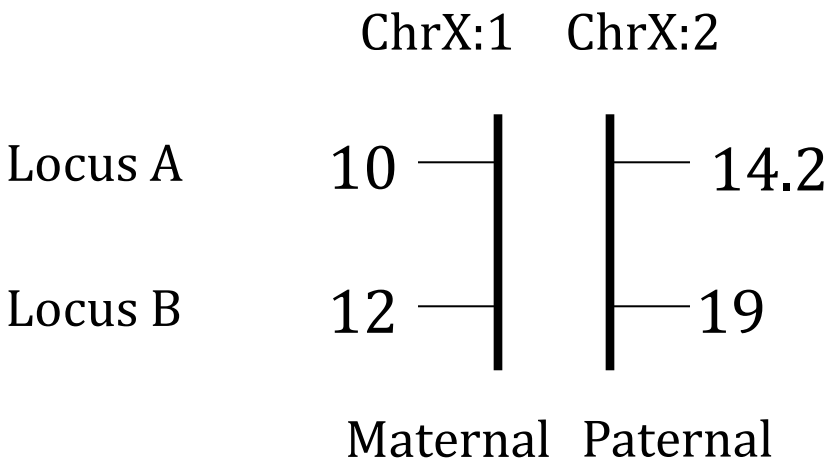


C1 and C2 are full siblings, DNA data is available only for C1 and C2

Basic notations:

Allele, haplotype, genotype, diplotype

Female

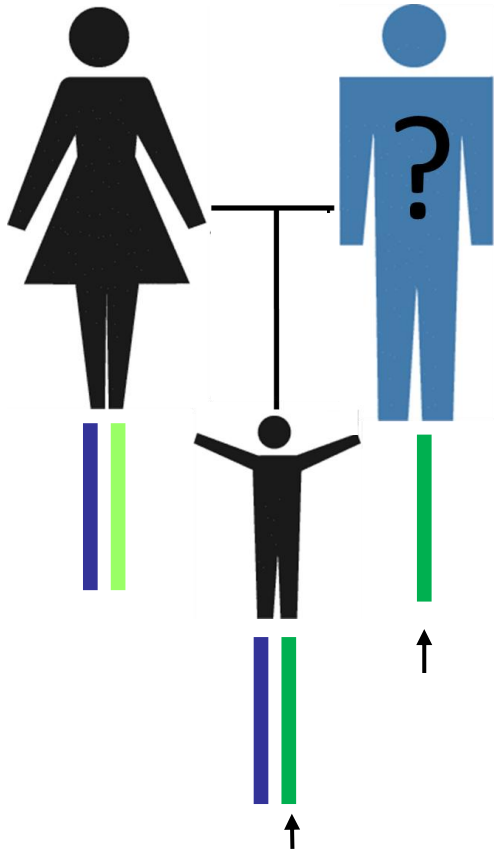


- 10 is an *allele*
- 10/14.2 (or 10,14.2) is a *genotype*
- 10_12 is a *haplotype*
- 10_12/14.2_19 is a *diplotype*
  (or 10_12|14.2_19)
  (or 10|14.2
  12|19)

Infer genetic relationship from DNA data – two key components

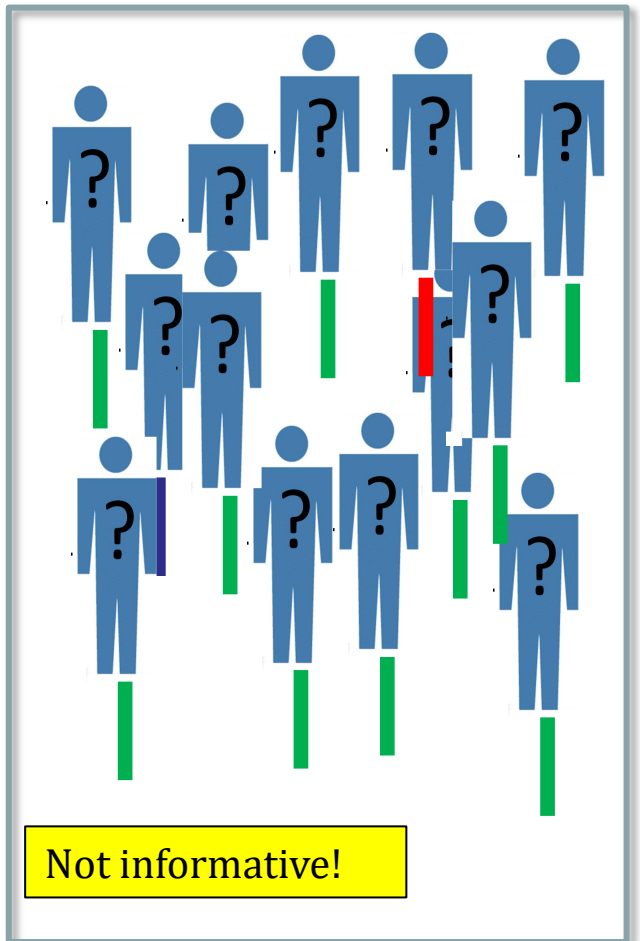
Inheritance patterns (mendelian principles)

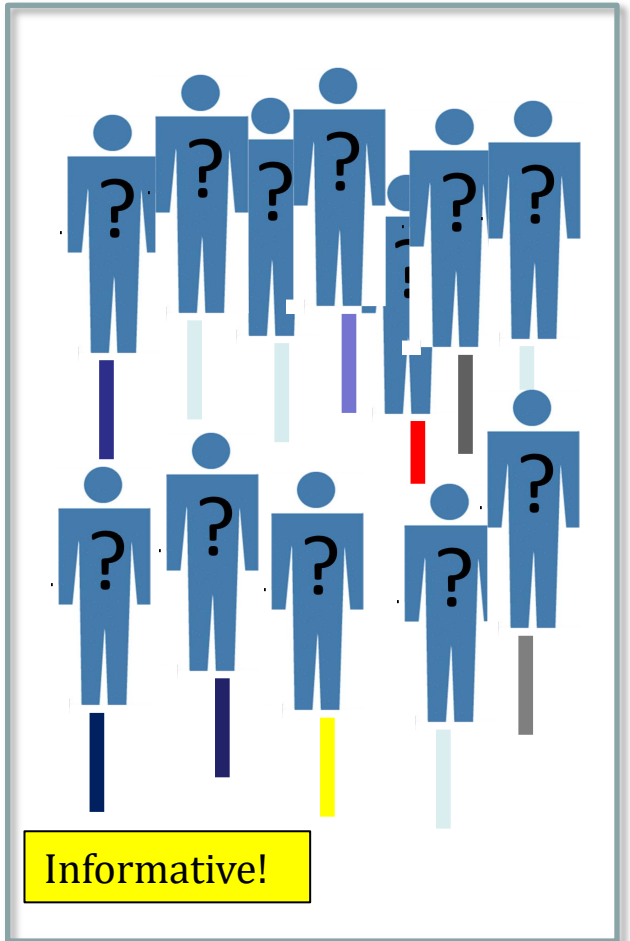
How DNA segments are segregated from parents to child



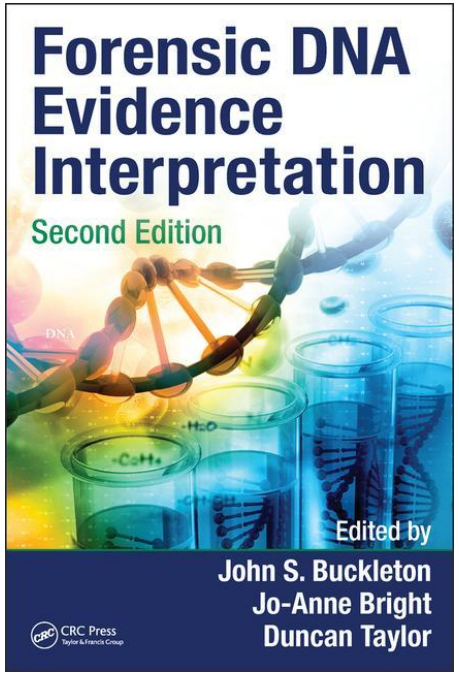
Population genetics

How common are shared alleles among ”unrelated” individuals?

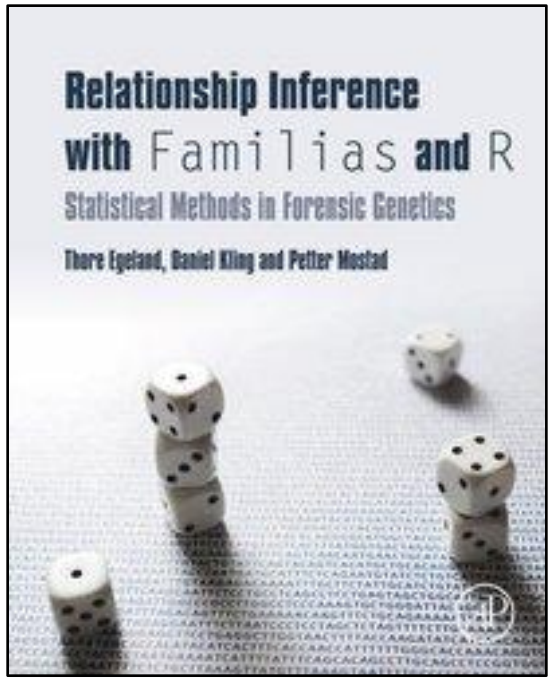




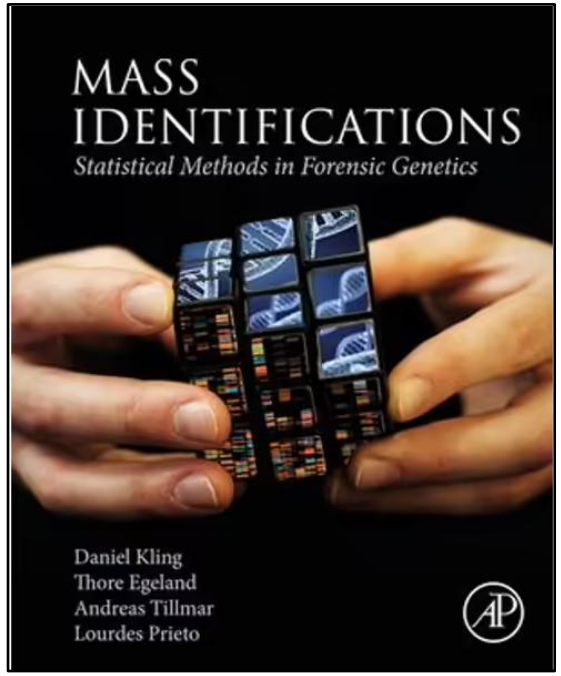
Books (example)



Forensic DNA Evidence Interpretation
John S. Buckleton, Jo-Anne Bright, Duncan Taylor



Relationship Inference with Familias and R Statistical Methods in Forensic Genetics
Thore Egeland, Daniel Kling, Petter Mostad



Mass Identifications Statistical Methods in Forensic Genetics
Daniel Kling, Thore Egeland, Andreas Tillmar, Lourdes Prieto

– Focus of this first presentation -
Assessing the weight of evidence

- Framework for the statistical interpretation
- Likelihood ratio principle
  - Basic principles (paternity/kinship)
  - Accounting for:
    - Mutations
    - Linkage and linkage disequilibrium
- Presenting the evidence
- The goal is to gain an understanding of how it works, in order to understand ***what*** affects the evidential weight and ***how*** it affects.

All from an X-chromosomal perspective!

Framework

- The goal is to provide the evidential weight, of the DNA findings, to the court (or other decision makers).
- The interpretation should be based on a solid framework, scientific principles, thoroughly tested methods
- Different approaches exist and have been used throughput the years
  - See next slides
- Importantly, we (forensic geneticists) provide the evidence and an interpretation, we **DO NOT** make decisions

Different approaches to infer relationship based on DNA data

- **Likelihood based approaches**
  - Calculates the likelihood of observed DNA data given pre-defined hypotheses.
  - Likelihood ratio principle, Paternity/Kinship/Sibling Index, full Bayesian approach

- **Exploratory approaches**
  - Quantifies allele sharing properties
  - Kinship coefficient, Relationship coefficient

- **Segment approach**
  - Measures the lengths of shared chromosomal segments
  - Most common in genetic genealogy applications

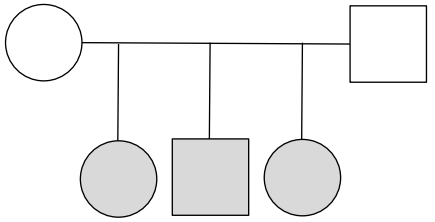
Require many DNA markers (>10,000) and will not be covered in this workshop

Different frameworks have historically been used for paternity and relationship testing

- Exclusion probability
  - *”The probability to exclude a sibship between Adam and Boris has been calculated to 99.9%”*
- Likelihood ratio principle
  - *“The observed DNA profiles are 1739 times more likely if Adam and Boris are related as half sibling than if Adam and Boris are unrelated”*
- Relationship probability
  - *”The posterior probability of Adam being a half sibling of Boris has been estimated to 99.99%”*

Likelihood based approaches - Basics

- Relationships are estimated based on specified alternatives (sources and hypotheses/pedigrees).
  - Can handle two or more hypotheses, and an ”unlimited” number of individuals.



- From each hypothesis/pedigree, a likelihood or conditional probability of observing the genotype results is computed based on Mendelian principles and underlying allele/ haplotype frequencies.
- This approach is current golden standard for classic forensic testing and the recommended approach.
  - ISFG (Gjertson et al., 2007, Tillmar et al., 2017, Roewer et al., 2020)
  - ENFSI (ENFSI Guideline for Evaluative Reporting in Forensic Science, 2016)
  - AABB (Guidance for Standards for Relationship Testing Laboratories, 2016)
  - SWGDAM (Recommendations of the SWGDAM Ad Hoc Working Group on Genotyping Results Reported as Likelihood Ratios)
- Via the Bayesian principle it is possible to combined evidence from multiple sources and convert prior information into posterior likelihoods and/or probabilities to aid decision making.

DNA testing to solve relationship issues

Is John the father of Mia?

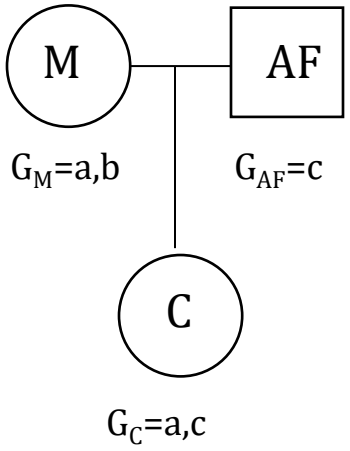
DNA data

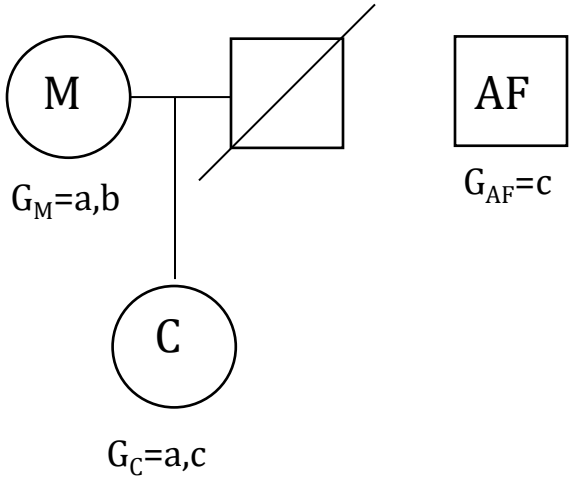
	DXS10101	DXS10301	HPRTB	….
Mia:	12,13	22,22	31,32.2	….
Mother of Mia:	12,14	22,23.2	31,30	….
John:	13	22	32.2	….

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

We will follow the recommendations given by:
Gjertson et al., 2007 "ISFG: Recommendations on biostatistics in paternity testing."
and
Tillmar et al., 2017 "DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis."

Paternity trio – An example





- Which pedigree/hypothesis explains the observed DNA data the best?
  - We calculate and compare these probabilities, and can get an estimate of the strength of the DNA evidence, supporting either hypothesis

Genotype/diplotype, allele/haplotype frequencies

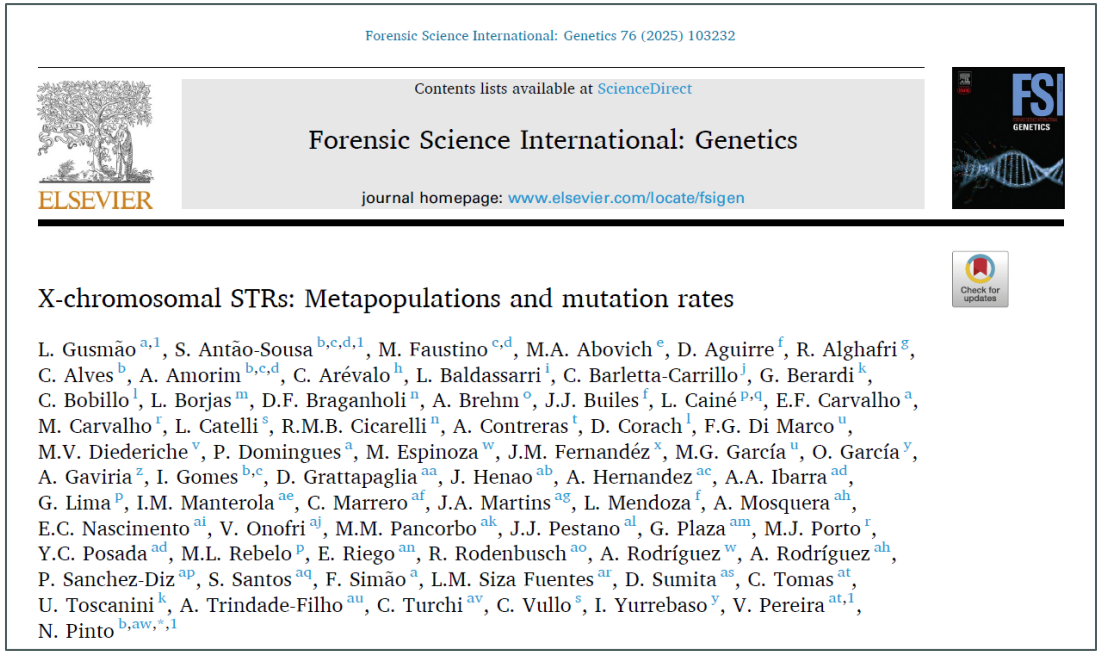
- We need **genotype/diplotype frequencies vs allele frequencies/ haplotype.**
  - Genotype/diplotype frequencies could be estimated from observed population genotype/diplotype reference data
    (requires large number of reference individuals)

  or

  - Genotype/diplotype frequencies could be estimate from population allele/haplotype frequencies (assuming Hardy-Weinberg Equilibrium)
- Allele/haplotype frequencies should be estimated from a case-independent set of reference individuals from the population of interest.
- The size of such reference database depends on the expected number of alleles/haplotypes, homogeneity/heterogeneity of the population (200-500 individuals could be sufficient, but could require more).
- The population should reflect the population of interest in the case.

Allele/haplotype frequencies

- Gusmao et al., 2025



Forensic Science International: Genetics 76 (2025) 103232

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen

X-chromosomal STRs: Metapopulations and mutation rates

L. Gusmão [a,1], S. Antão-Sousa [b,c,d,1], M. Faustino [c,d], M.A. Abovich [e], D. Aguirre [f], R. Alghafri [g], C. Alves [b], A. Amorim [b,c,d], C. Arévalo [h], L. Baldassarri [i], C. Barletta-Carrillo [j], G. Berardi [k], C. Bobillo [l], L. Borjas [m], D.F. Braganholi [n], A. Brehm [o], J.J. Builes [f], L. Cainé [p,q], E.F. Carvalho [a], M. Carvalho [r], L. Catelli [s], R.M.B. Cicarelli [n], A. Contreras [t], D. Corach [l], F.G. Di Marco [u], M.V. Diederiche [v], P. Domingues [a], M. Espinoza [w], J.M. Fernandéz [x], M.G. García [u], O. García [y], A. Gaviria [z], I. Gomes [b,c], D. Grattapaglia [aa], J. Henao [ab], A. Hernandez [ac], A.A. Ibarra [ad], G. Lima [p], I.M. Manterola [ae], C. Marrero [af], J.A. Martins [ag], L. Mendoza [f], A. Mosquera [ah], E.C. Nascimento [ai], V. Onofri [aj], M.M. Pancorbo [ak], J.J. Pestano [al], G. Plaza [am], M.J. Porto [r], Y.C. Posada [ad], M.L. Rebelo [p], E. Riego [an], R. Rodenbusch [ao], A. Rodríguez [w], A. Rodríguez [ah], P. Sanchez-Diz [ap], S. Santos [aq], F. Simão [a], L.M. Siza Fuentes [ar], D. Sumita [as], C. Tomas [at], U. Toscanini [k], A. Trindade-Filho [au], C. Turchi [av], C. Vullo [s], I. Yurrebaso [y], V. Pereira [at,1], N. Pinto [b,aw,*,1]

- https://famlink.se/fx_databases.html

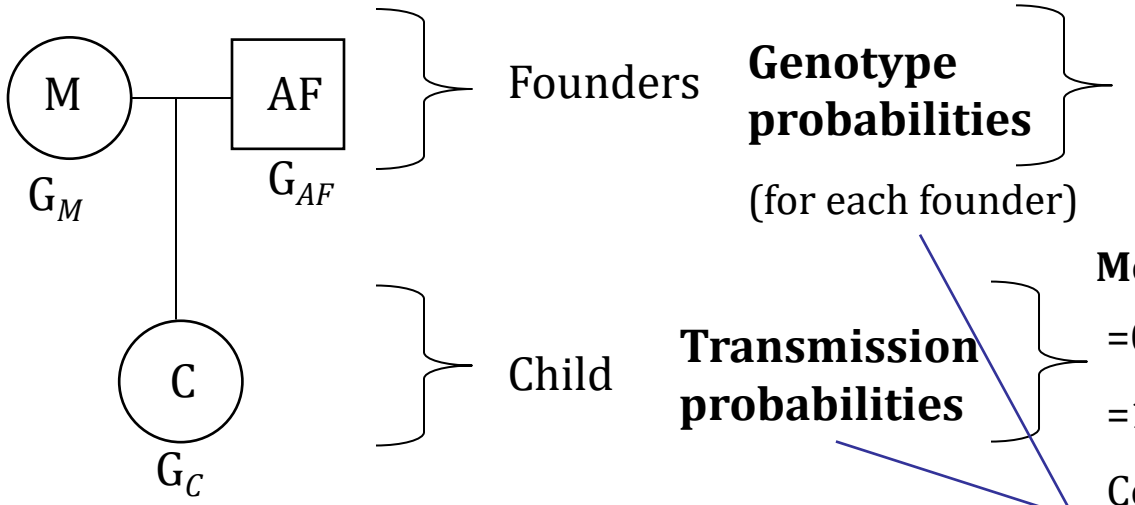
Subpopulation correction

"If a significant degree of substructuring is known to be present in a population, algorithms that take substructuring into consideration shall be used." (Gjertson et al., 2007)

For X loci, see Ayres et al., (2005) *Calculating the exclusion probability and paternity index for X-chromosomal loci in the presence of substructure.* FSI 149:201–203

- What is subpopulation effects?
  - First developed by Wright in (1965)
  - Balding & Nichols (1994)
- How to estimate it?
  - Reasonable values (0-0.05, less than 0.001)
- Sampling formula (Fst = θ)

LR=



Genotype probabilities

(for each founder)

Assuming Hardy-Weinberg equilibrium (HWE) and no substructure

p_i: allele frequency for allele *i*

$2*p_i*p_j$	*Heterozygous*
p_i*p_i	*Homozygous*
p_i	*Hemizygous*

Transmission probabilities

Mendelian factor

=0.5, if parent is heterozygous

=1, if parent is homozygous or hemizygous

Could also be μ, r etc

Paternity trio – A simple example with one X-chromosomal marker

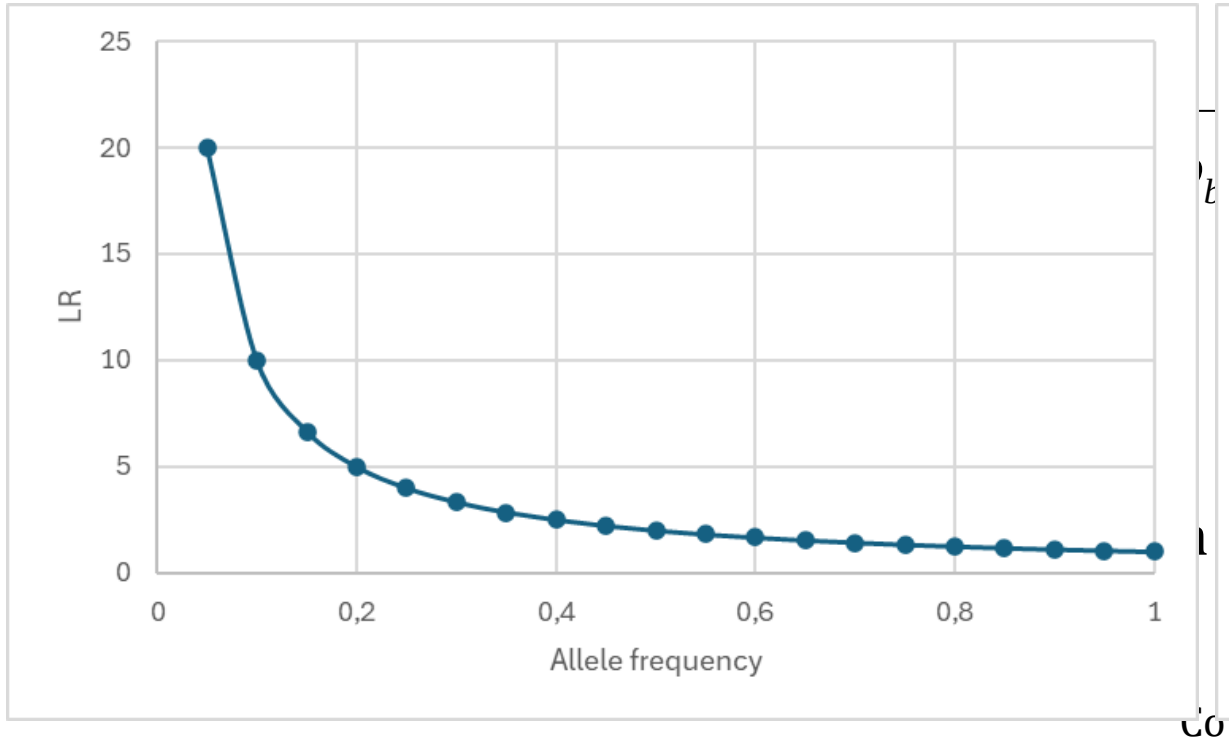
$$LR = \frac{2 \cdot p_a \cdot p_b \quad p_c \quad 0.5 \quad 1}{= \quad 2 \cdot p_a \cdot p_b \quad p_c \quad 0.5 \quad p_c} = \frac{1}{p_c}$$

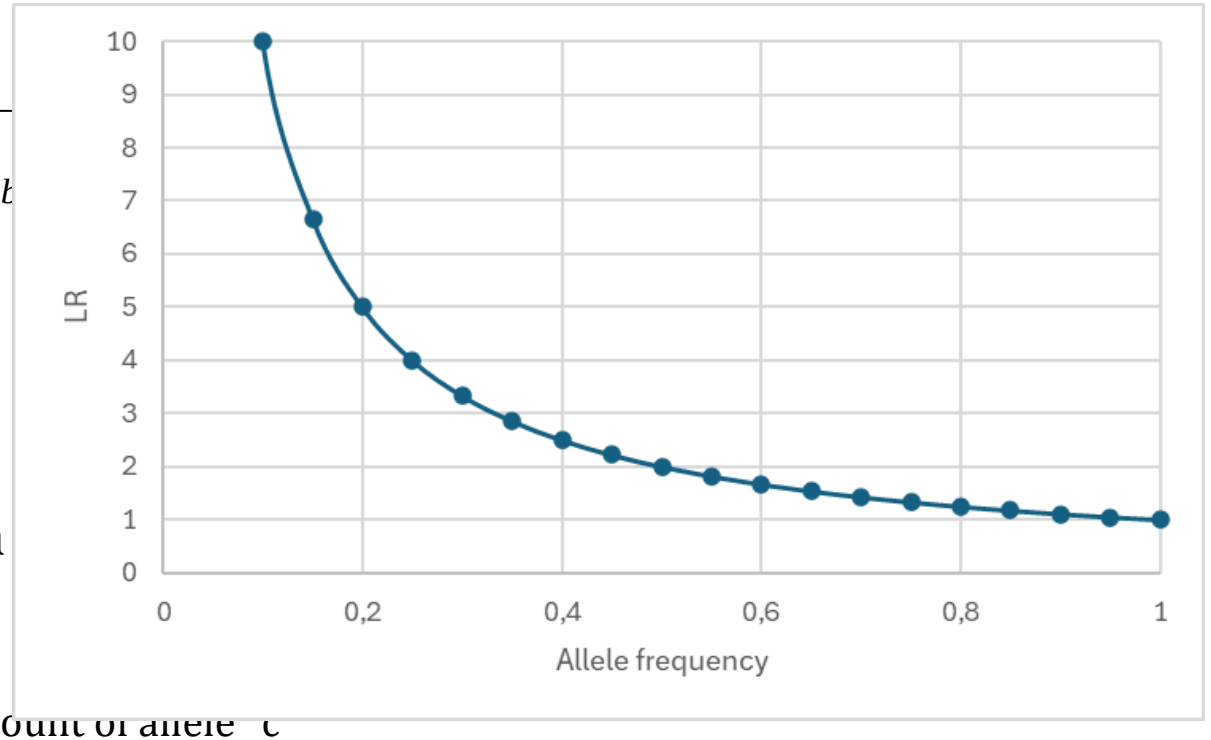
P_c : Probability to observe allele “c” in the population (e.g. population frequency)

Count of allele ”c”

Total number of observed alleles in the population database

Paternity trio – A simple example with one X-chromosomal marker

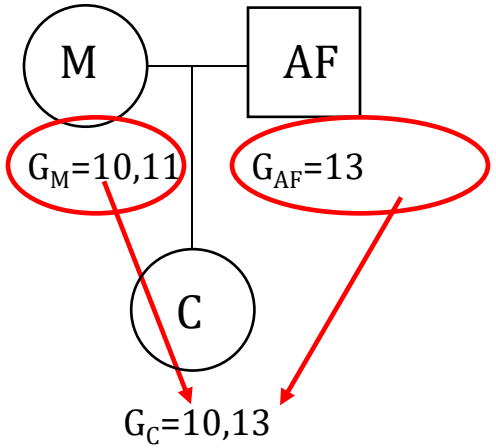




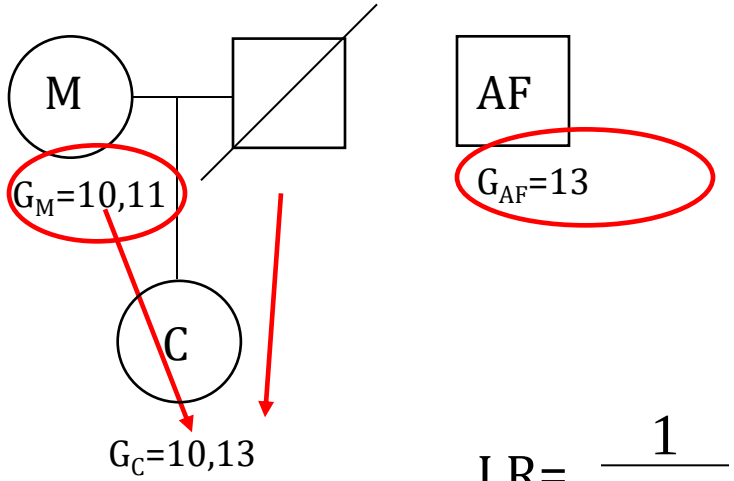
Count of allele "t"

Total number of observed alleles in the population database

Paternity trio – real X-STR data



$$2 \cdot p_{10} \cdot p_{11} \cdot p_{13} \quad 0.5 \quad 1$$



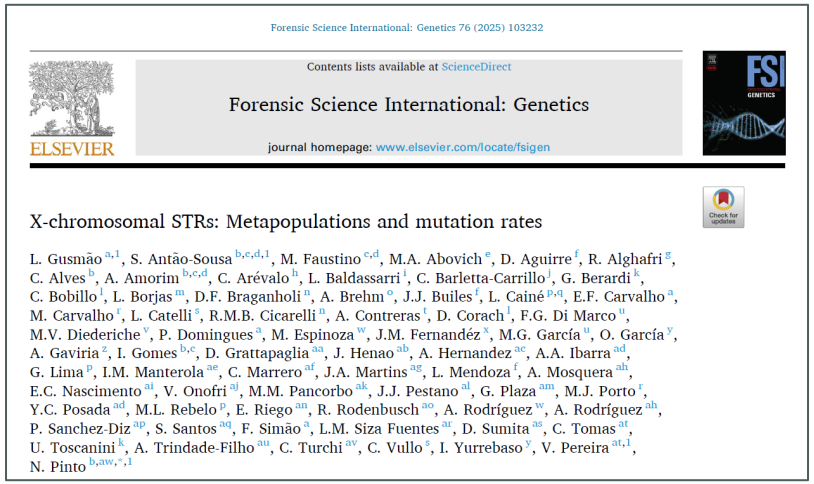
$$2 \cdot p_{10} \cdot p_{11} \cdot p_{13} \quad 0.5 \cdot p_{13}$$

$$\text{LR} = \frac{1}{p_{13}}$$

STR mutations

"The possibility of mutation shall be taken into account whenever a genetic inconsistency is observed" (Gjertson et al., 2007)

- 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.
- Gusmao et al (2025)



Forensic Science International: Genetics 76 (2025) 103232

Contents lists available at ScienceDirect

Forensic Science International: Genetics

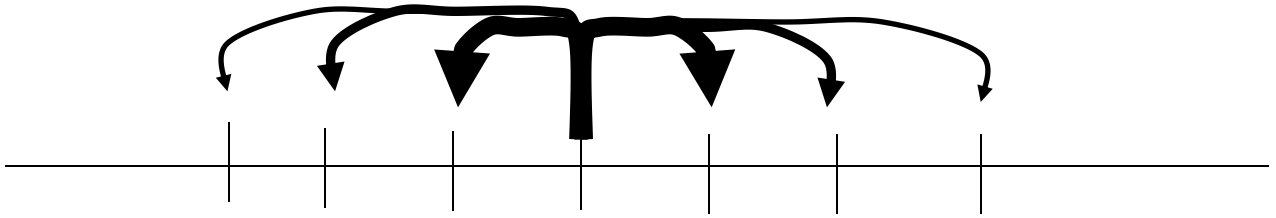
journal homepage: www.elsevier.com/locate/fsigen

X-chromosomal STRs: Metapopulations and mutation rates

L. Gusmão [a,1], S. Antão-Sousa [b,c,d,1], M. Faustino [c,d], M.A. Abovich [e], D. Aguirre [f], R. Alghafri [g], C. Alves [b], A. Amorim [b,c,d], C. Arévalo [h], L. Baldassarri [i], C. Barletta-Carrillo [j], G. Berardi [k], C. Bobillo [l], L. Borjas [m], D.F. Braganholi [n], A. Brehm [o], J.J. Builes [f], L. Cainé [p,q], E.F. Carvalho [a], M. Carvalho [r], L. Catelli [s], R.M.B. Cicarelli [n], A. Contreras [t], D. Corach [i], F.G. Di Marco [u], M.V. Diederiche [v], P. Domingues [a], M. Espinoza [w], J.M. Fernandéz [x], M.G. García [u], O. García [y], A. Gaviria [z], I. Gomes [b,c], D. Grattapaglia [aa], J. Henao [ab], A. Hernandez [ac], A.A. Ibarra [ad], G. Lima [p], I.M. Manterola [ae], C. Marrero [af], J.A. Martins [ag], L. Mendoza [f], A. Mosquera [ah], E.C. Nascimento [ai], V. Onofri [aj], M.M. Pancorbo [ak], J.J. Pestano [al], G. Plaza [am], M.J. Porto [r], Y.C. Posada [ad], M.L. Rebelo [p], E. Riego [an], R. Rodenbusch [ao], A. Rodríguez [w], A. Rodríguez [ah], P. Sanchez-Diz [ap], S. Santos [aq], F. Simão [a], L.M. Siza Fuentes [ar], D. Sumita [as], C. Tomas [at], U. Toscanini [k], A. Trindade-Filho [au], C. Turchi [av], C. Vullo [s], I. Yurrebaso [y], V. Pereira [at,1], N. Pinto [b,aw,*,1]

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

Models for STR mutations



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}

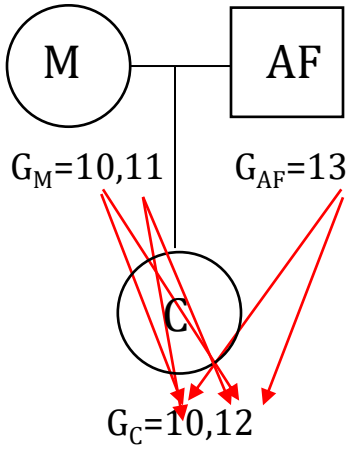
$$\mu_{tot}=\mu_{+1}+\mu_{-1}+\mu_{+2}+\mu_{-2}+\ldots$$

Different approaches to calculate LR accounting for mutations exist.

The most used one *"Stepwise mutation model"*

LR ~ (μ_{tot}*adj_steps)/p(paternal allele)

Paternity trio - Mutation



$(mut_{13\rightarrow12})$

Mutation model decreasing with range

A software like FamLinkX will consider all mutation possibilities (if the mutation rate is set to >0)

Total mutation rate for the locus

50% are loss of fragment size

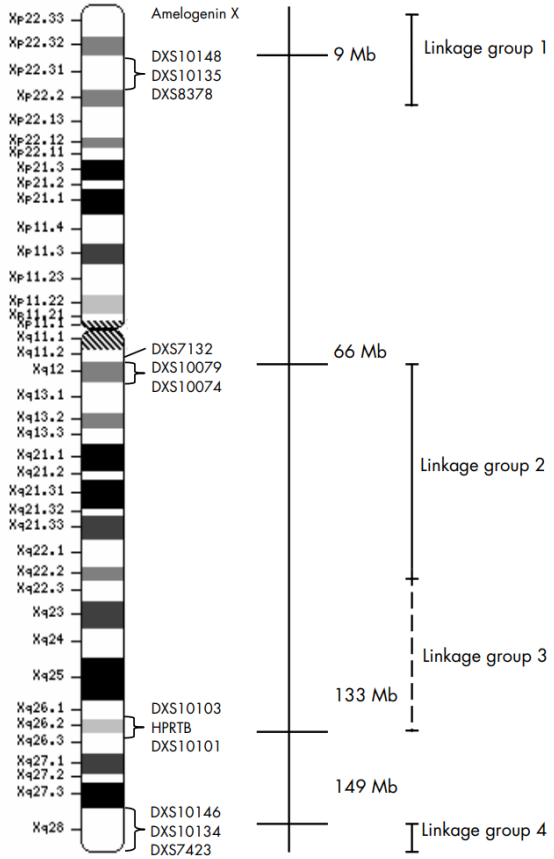
$$mut_{13->12} = \{ \text{“Stepwise decreasing with range”} \} = 1 \cdot \mu_{Tot} \cdot 0.9 \cdot 0.5$$

100% for 13 to be the parental allele

90% of mutations are 1-step

Linkage and Linkage disequilibrium (LD)

Argus X12 QS

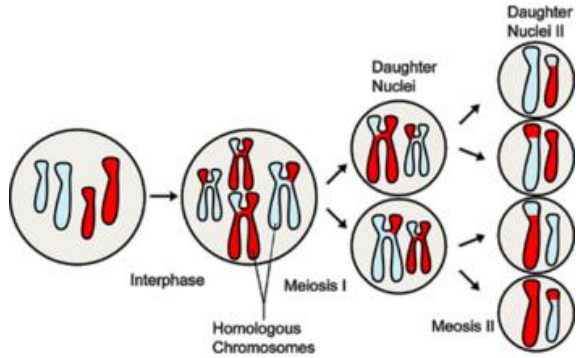


1. LD within each linkage group => Haplotypes instead of alleles

2. Co-segregation of markers from different linkage groups (i.e. linkage groups are linked) => LR per marker/linkage group **cannot** be multiplied into a combined LR

Linkage and Linkage disequilibrium

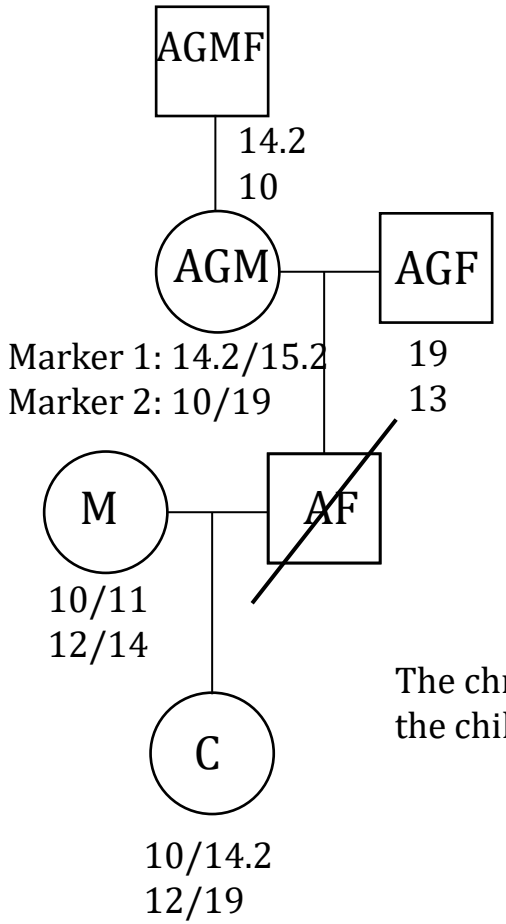
- Linkage
  - Can be described as the co-segregation of closely located loci within a family or pedigree.
  - **Effects the transmission probabilities!**



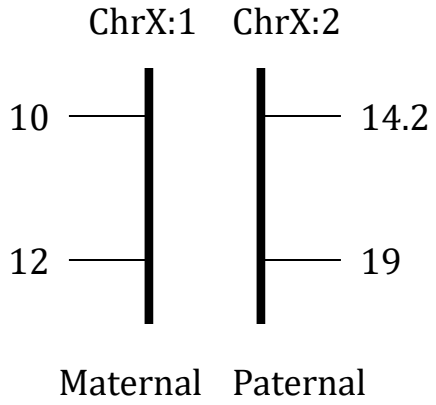
- Linkage disequilibrium (LD)
  - Allelic association.
  - Two alleles (at two different markers) which is observed more often/less often than can be expected.
  - **Effects the founder genotype probabilities, not the transmission probabilities!**

Linkage and how it impacts the LR

Marker 1 and 2 are located on the same chromosome (chr X)



The chromosomal phase of the child can be inferred



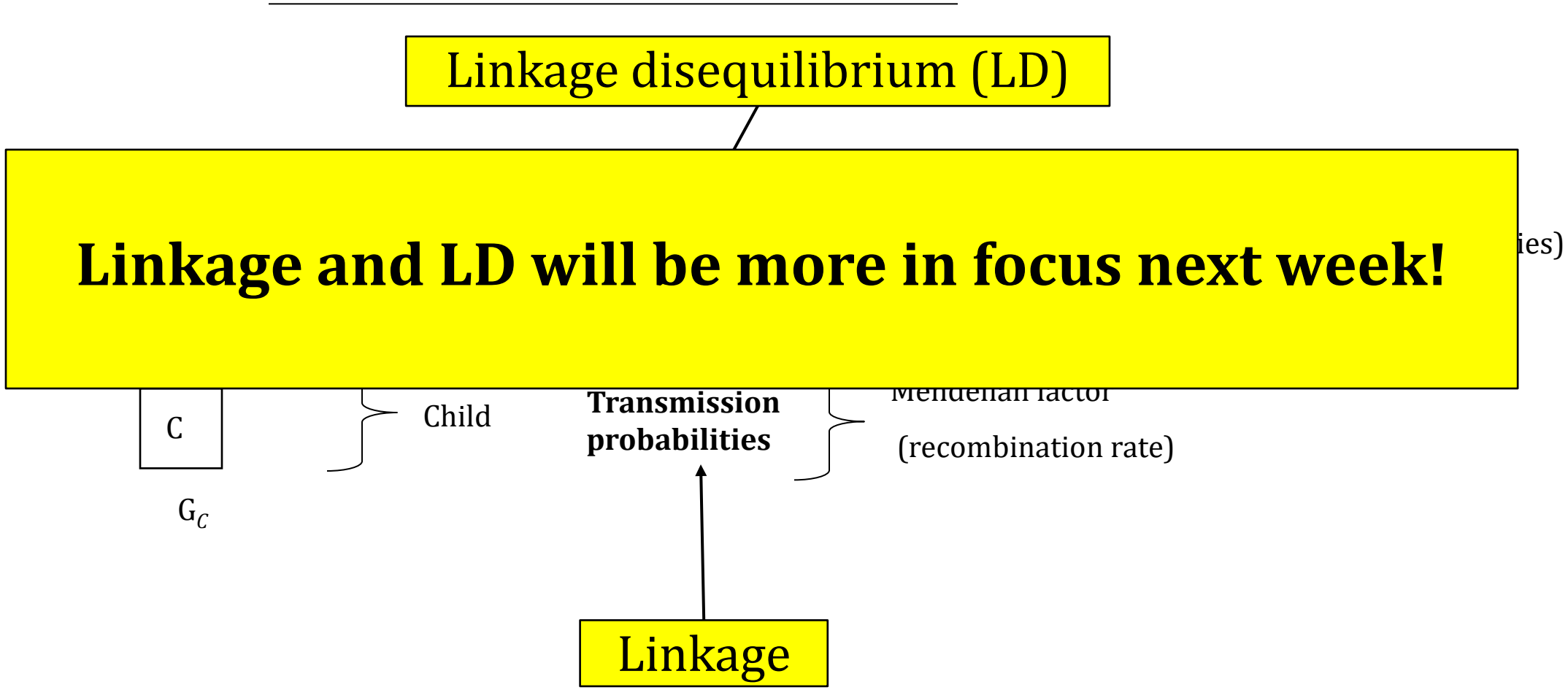
A recombination must have occurred at AGM to explain the data (given this pedigree).

The probability of the observed DNA data (given this pedigree) depends on the recombination rate between marker 1 and marker 2!

E.g. if this recombination rate is very low, the probability is very low. Also, if the recombination rate is 0 (very very close markers), the observed data is not possible (given this pedigree)

Ignoring genetic linkage may result in false LRs

LR=


Linkage disequilibrium (LD)
Linkage and LD will be more in focus next week!
ies)
C
Child
Transmission probabilities
Mendelian factor
(recombination rate)
G_C
Linkage

Presenting the evidence

- Important that we can present the evidential weight to the decision makers!
- Likelihood ratio (LR) (or Paternity index (PI))
- Posterior probability (must set priors!)

45 report LR(PI) and/or 36 paternity probability (*ESWG questionnaire, proficiency test 2017*)

Gjertson et al., 2007

Simplifies to LR/LR+1 if we have two hypotheses with equal priors

An example

- H1: The tested female children has the same father
- H2: : The tested female children has different fathers
- Likelihood, Pr(DNA|H1)=0.0123
- Likelihood, Pr(DNA|H2)=0.0010
- **LR=12.3** (…"the observed DNA data is 12.3 times more likely if the tested children has the same father, than if the tested children has different fathers")
- Prior probability: 0.8 (H1), 0.2 (H2)

$$W_i = \frac{\pi_i L_i}{\sum_{j=1}^{n} \pi_j L_j}$$ *Gjertson et al., 2007*

- **Posterior probability, Pr(H1|Evidence):** 0.8*0.0123/(0.8*0.0123+0.2*0.001)=**0.98008** (…"the posterior probability that the tested children has the same father has been estimated to 98.0 %.")

ENFSI Guideline for Evaluative Reporting in Forensic Science

Hummel's chart		Equivalent in terms of Likelihood	
Vaterschafts-wahrscheinlichkeit	verbales Prädikat	verbal predicate	likeli ratio
(99,9)	Vaterschaft praktisch erwiesen	Paternity practically proven	
99,8			399
	Vaterschaft höschst wahrscheinlich	highly likely	
99			99
	Vaterschaft sehr wahrscheinlich	very likely	
95			19
	Vaterschaft wahrscheinlich	likely	
90			9
(50)	(Ohne Prädikat)	(no verbal predicate)	
10			1/9
	Vaterschaft unwahrscheinlich	unlikely	
5			1/19
	Vaterschaft sehr unwahrscheinlich	very unlikely	
1			1/99
	Vaterschaft höchst unwahrscheinlich	highly unlikely	
0,2			1/39
(0,1)	Vaterschaft praktisch ausgeschlossen	practically excluded	

Hummel 1981 Biomathematical evidence of paternity, (Beweis der Vaterschaft) Berlin, Sprir

Values* of likelihood ratio	Verbal equivalent (two options of phrasing are suggested)
1	The forensic findings do not support one proposition over the other. The forensic findings provide no assistance in addressing the issue.
2 - 10	The forensic findings provide weak support** for the first proposition relative to the alternative. The forensic findings are slightly more probable given one proposition relative to the other.
10 - 100	...provide moderate support for the first proposition rather than the alternative …are more probable given…proposition...than proposition...
100 - 1000	...provide moderately strong suppor tfor the first proposition rather than the alternative …are appreciably more probable given… proposition...than proposition...
1000 - 10,000	...provide strong support for the first proposition rather than the alternative …are much more probable given… proposition...than proposition...
10,000 - 1,000,000	...provide very strong support for the first proposition rather than the alternative …are far more probable given… proposition...than proposition...
1,000,000 and above	...provide extremely strong support for the first proposition rather than the alternative …are exceedingly more probable given… proposition...than proposition...

Setting the hypotheses?

- Who?
  - Court/customer to hinder bias
- How many hypotheses?
  - Only the relevant to hinder ”dilution”
    - 99% uncle vs 1% unrelated (if 2 hypotheses)
    - 33% uncle vs 33% half sibs vs 33% grandfather vs 1 % unrelated. (if 4 hypotheses)
- Always report assumptions
  - ”Full siblings” vs ”unrelated”, not ”Siblings” vs ”not siblings”, not only “Siblings”

Authentic case example 1

-DNA profiles from alleged father and child
- No genetic inconsistencies

Father vs unrelated : LR=75 654 625 i.e. >99.999% father

Uncle vs unrelated: LR=23 975 i.e. >99.99% uncle

Father vs uncle: LR=3 155 i.e. >99.9% father

Authentic case example 2

Full siblings?

DNA profiles from individuals A and B

Full siblings vs unrelated : LR=132 i.e. >99% full siblings

Full siblings vs half siblings: LR=0.01 i.e. <1% full siblings

Choice of hypotheses matters!

Exercises

- Separate file (and solutions)