



Pedigrees and relatedness analysis in R

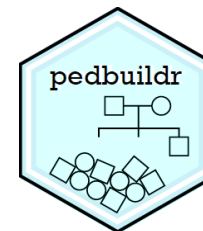
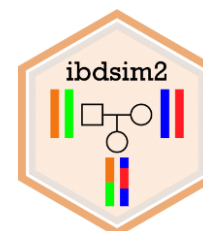
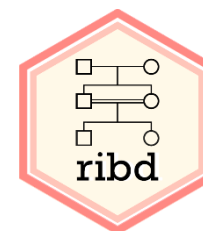
Zakopane - 23.05.2022

Magnus Dehli Vigeland



Plan

- Lecture
 - Part 1: Genetic relatedness
 - Part 2: Forensic applications
- Workshop:
 - Session 1: Pedigree analysis in R
 - Session 2: Forensic case studies

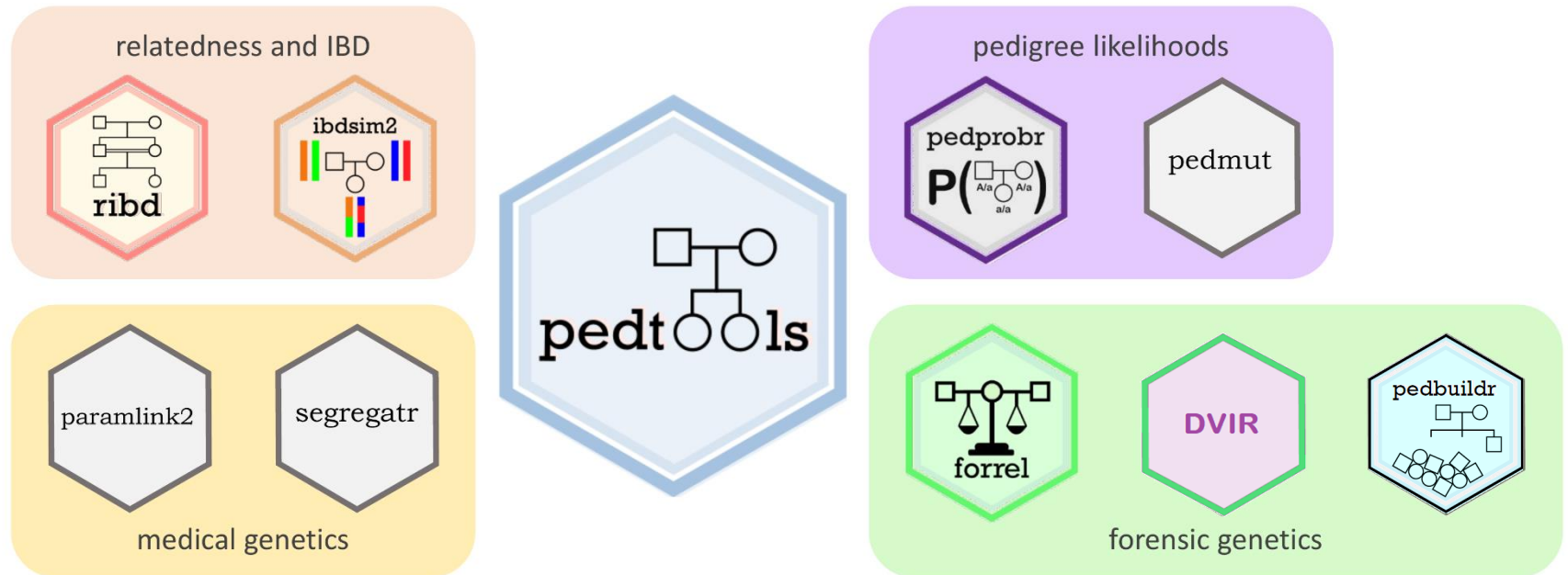


Lecture notes and exercise sets

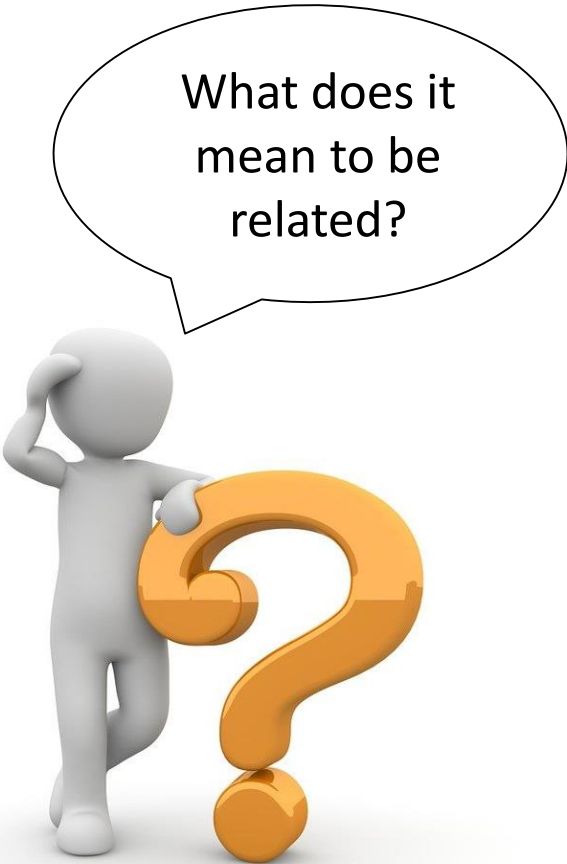
https://magnusdv.github.io/pedsuite/articles/web_only/courses.html

The ped suite

A collection of packages for pedigree analysis in R



Home page: <https://magnusdv.github.io/pedsuite/>

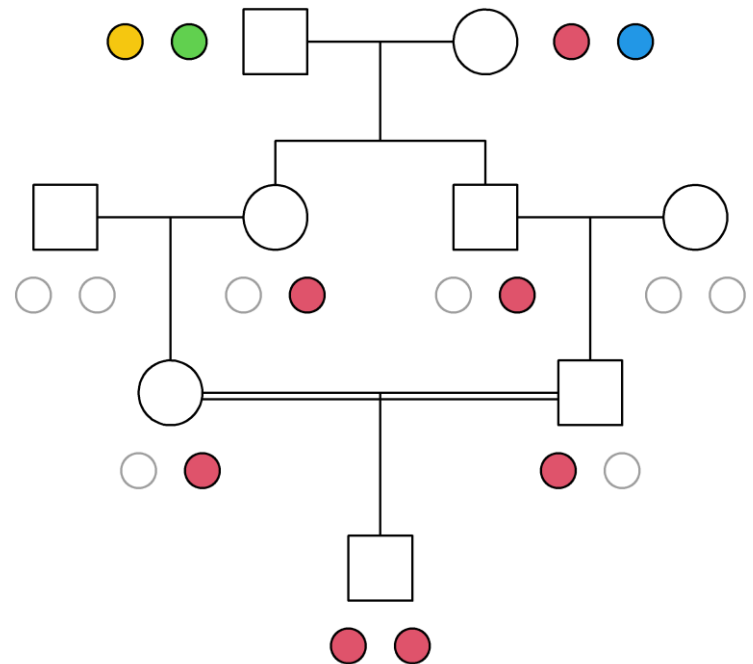


What does it
mean to be
related?

- Attempt 1
 - being connected through a pedigree
 - having a common ancestor...
not too far back
- Attempt 2 - genetic
 - sharing DNA
 - (more than unrelated people)
- To make this precise, we need some terminology!

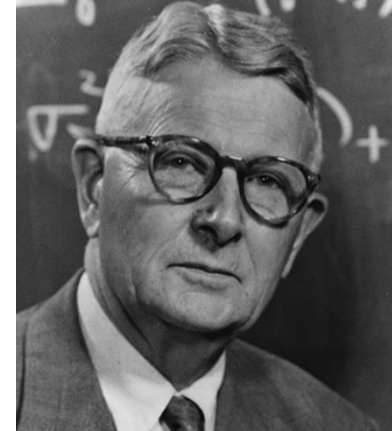
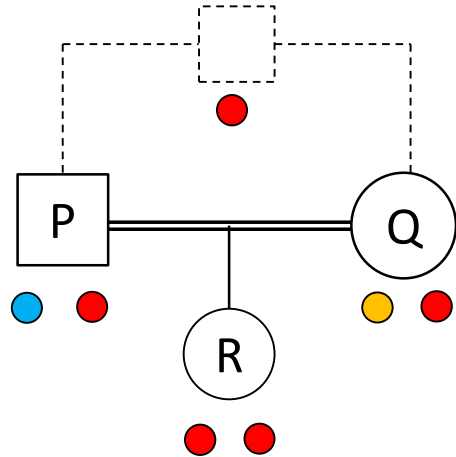
IBD and autozygosity

- IBD = **Identical by descent**
= identical alleles with a common origin **in the given pedigree**
- autozygous = homozygous + IBD



Inbreeding coefficient
 $f = P(\text{autozygosity})$

Coefficient of kinship/inbreeding



Sewall Wright
(1889 - 1988)

- Wright (1921): The **kinship coefficient** φ between P and Q

$$\varphi_{P,Q} = P(\text{random allele of P is IBD with random allele of Q})$$

$$= P(R \text{ is autozygous})$$

$$= f_R$$

Inbreeding coefficient of a child = kinship between parents

P and Q related

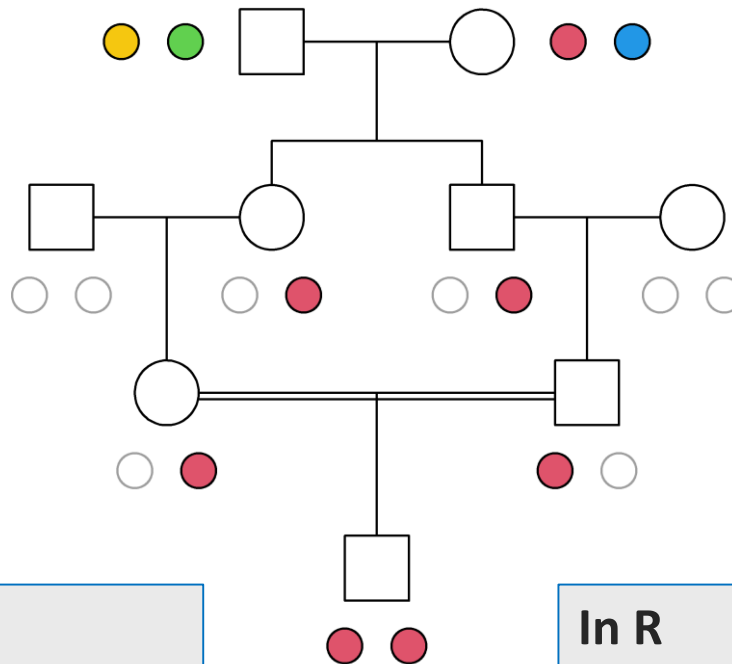


$$\varphi_{P,Q} > 0$$

Inbreeding coefficient: Example

Wright's path formula:

$$\varphi_{P,Q} = \sum_A \sum_v \left(\frac{1}{2}\right)^{|v|+1} (1 + f_A)$$



By hand

$$\begin{aligned} f &= P(\text{red circle autozygous}) \cdot 4 \\ &= 0.5^6 \cdot 4 \\ &= 1/16 \end{aligned}$$

↑
other
colors

In R

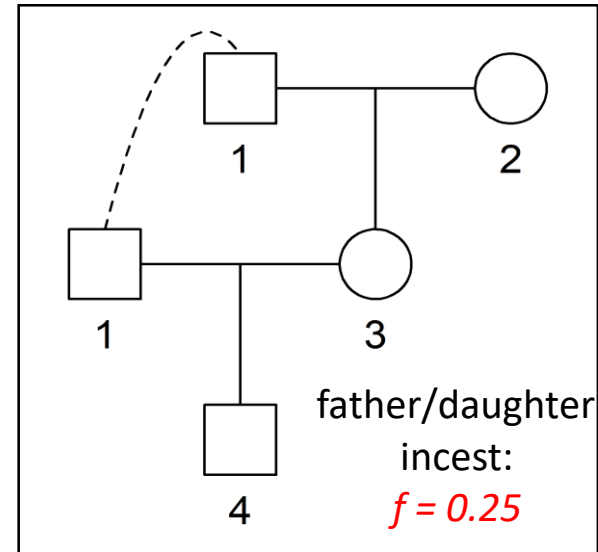
```
> library(ribd)

> x = cousinPed(1, child = T)
> inbreeding(x, ids = 9)

[1] 0.0625
```

More kinship & inbreeding coefficients

Relationship	Kinship	f (of child)
Parent-child	$1/4$	$1/4$
Full siblings	$1/4$	$1/4$
Half siblings	$1/8$	$1/8$
Grandparent-grandchild	$1/8$	$1/8$
Avuncular (uncle/aunt)	$1/8$	$1/8$
1st cousins	$1/16$	$1/16$
2nd cousins	$1/64$	$1/64$
3rd cousins	$1/256$	$1/256$



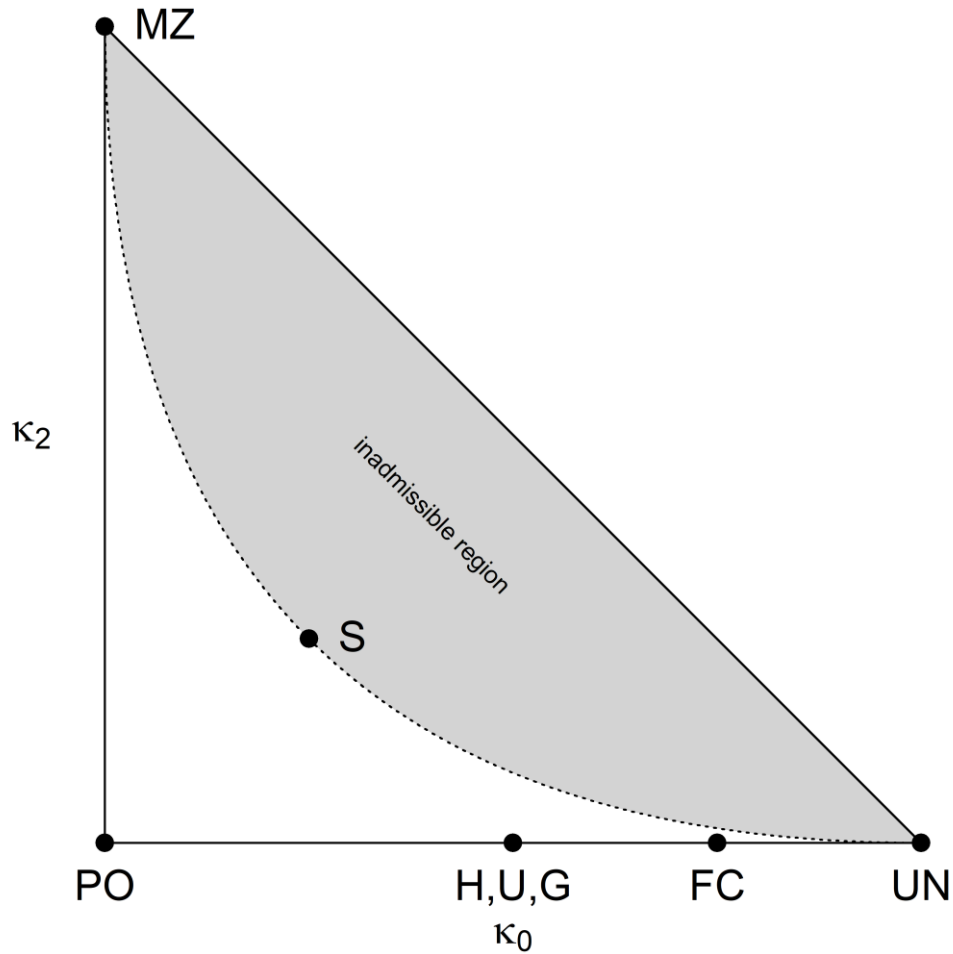
```

> x = nuclearPed(1, sex = 2)
> kinship(x, ids = c(1, 3))
[1] 0.25

> x = addchildren(x, 1, 3, nch = 1)
> inbreeding(x, 4)
[1] 0.25

```


The IBD triangle



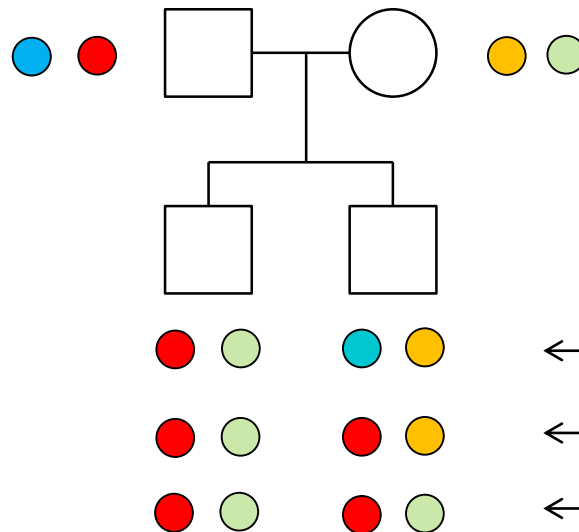
Charles Cotterman
(1914-1989)



Elisabeth Thompson
(1949 -)

IBD coefficients: Warm-up

- Summary so far:
 - Two individuals are related if they can have IBD alleles
 - Their kinship coefficient measures the amount of IBD sharing
- Natural generalisation:
 - How *many* alleles are IBD in each locus?



← no IBD sharing

← share ●

← share ● and ●

Humans are
diploid



IBD = 0, 1 or 2

IBD coefficients: Definition

- Given two (non-inbred) individuals

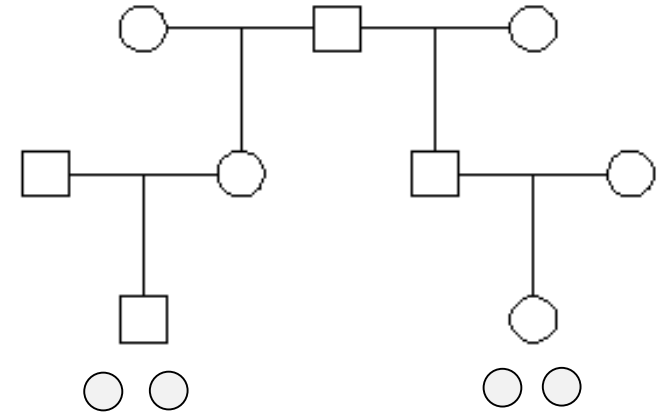
- For a random autosomal locus

$$\kappa_0 = P(0 \text{ alleles IBD})$$

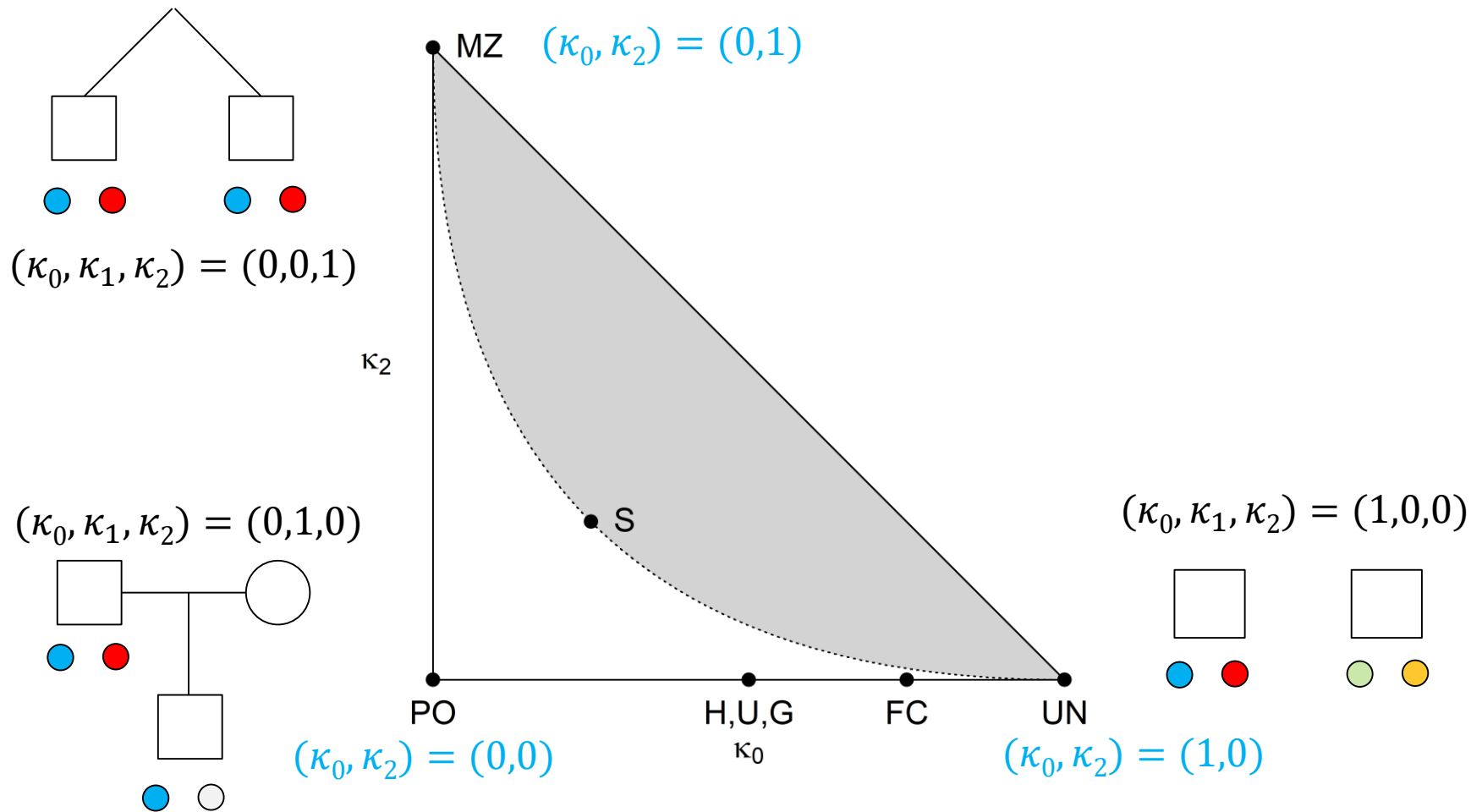
$$\kappa_1 = P(1 \text{ alleles IBD})$$

$$\kappa_2 = P(2 \text{ alleles IBD})$$

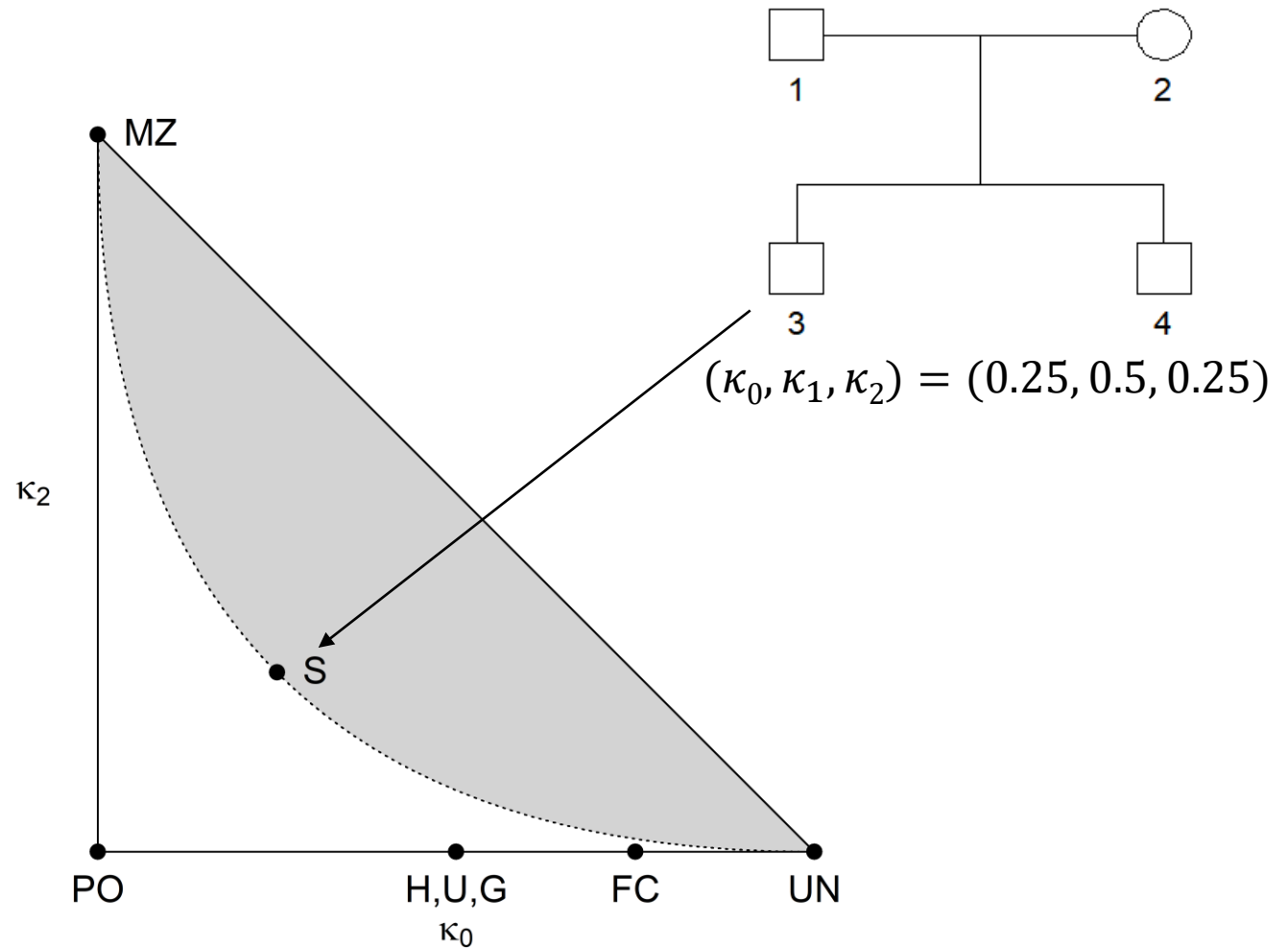
- We always have: $\kappa_0 + \kappa_1 + \kappa_2 = 1$



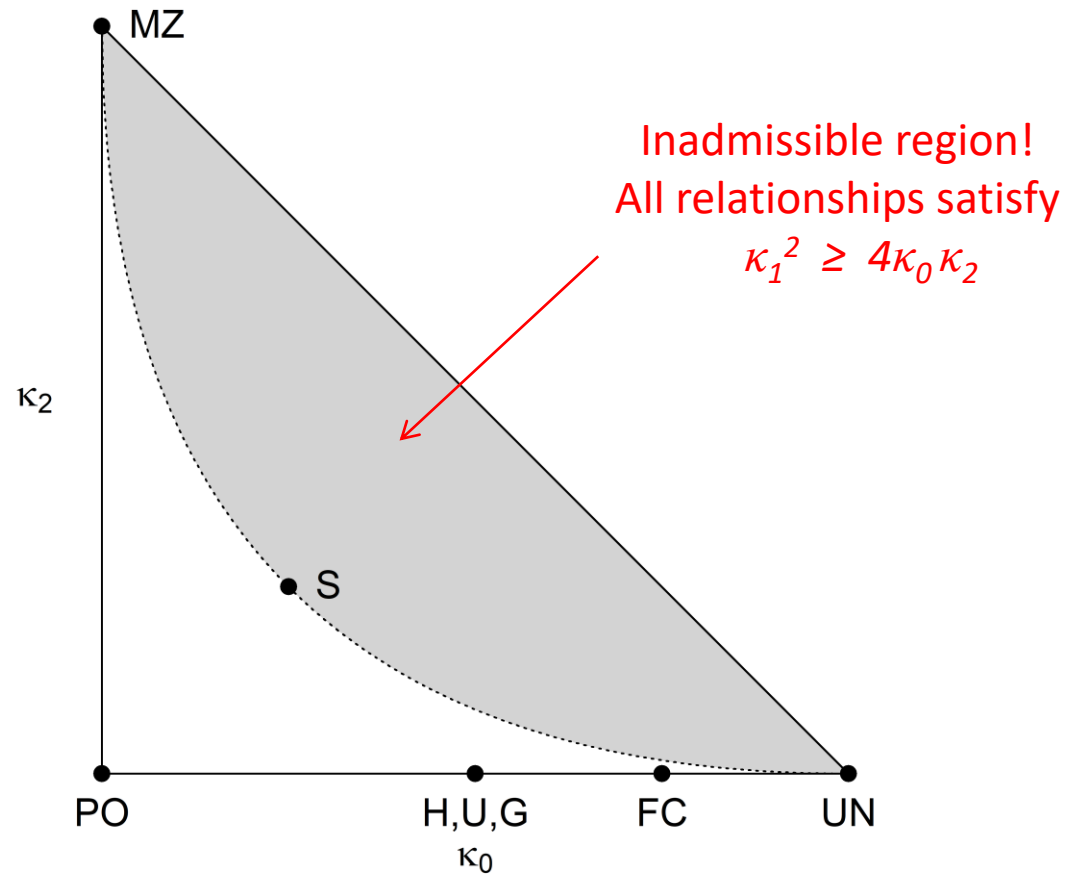
The relatedness triangle



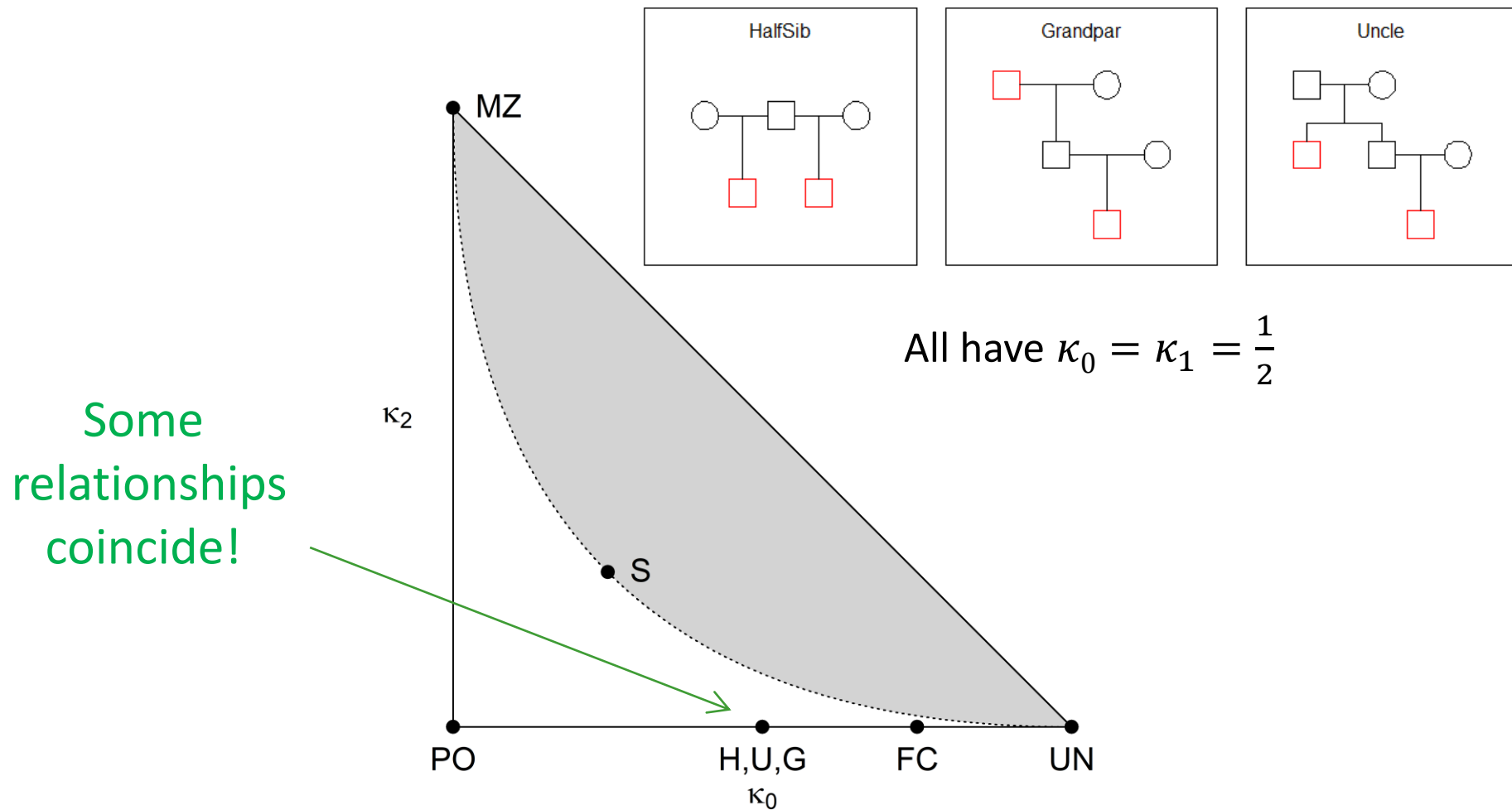
The relatedness triangle



The relatedness triangle

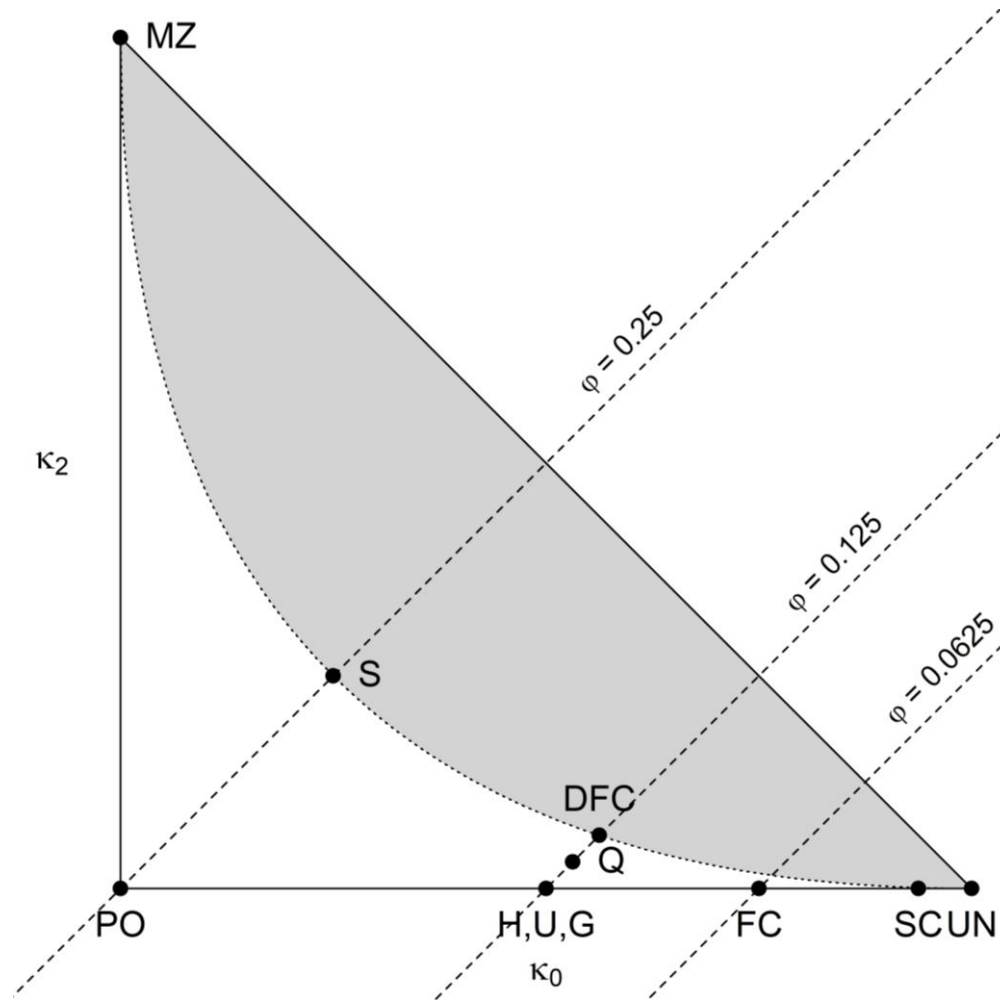


The relatedness triangle

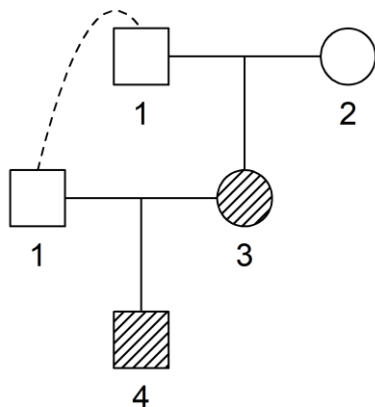


An important identity:

$$\varphi = \frac{1}{4}\kappa_1 + \frac{1}{2}\kappa_2$$



Reminder

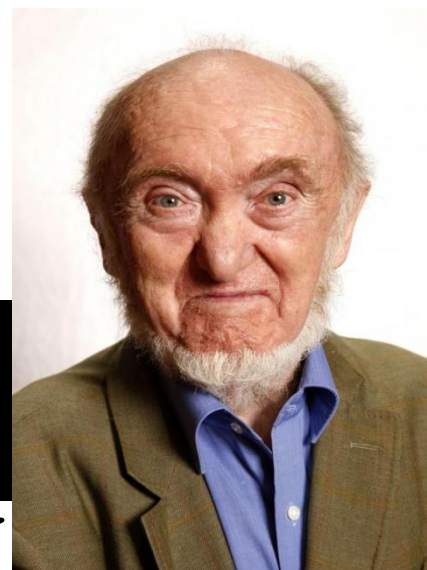


K -coefficients are only defined for non-inbred individuals.
For the whole story, we need 9 coefficients!

Jacquard's identity coefficients

Covered in Chapter 3 of
Pedigree analysis in R

Only for the
brave!



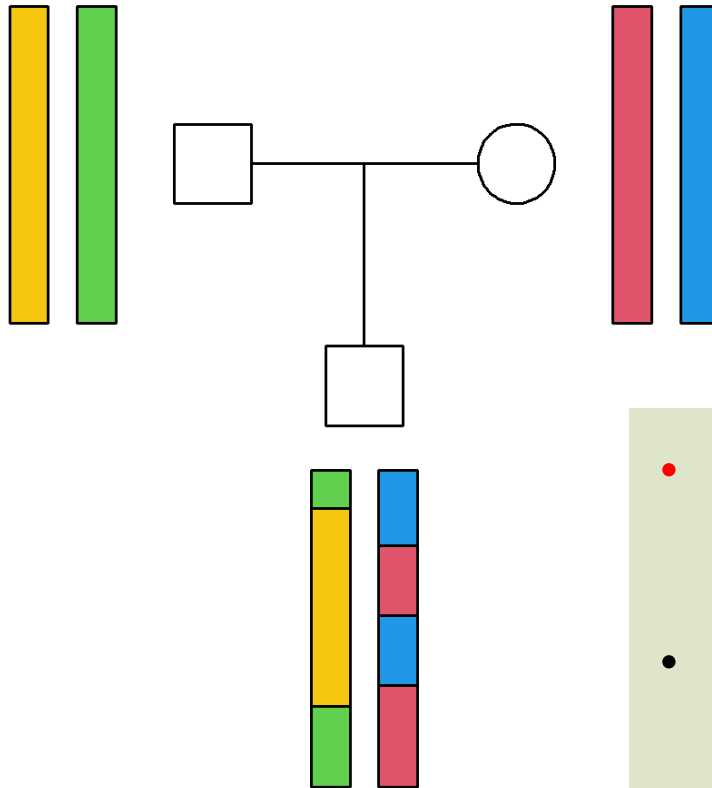
Albert Jacquard
(1925 - 2013)



Realised relatedness

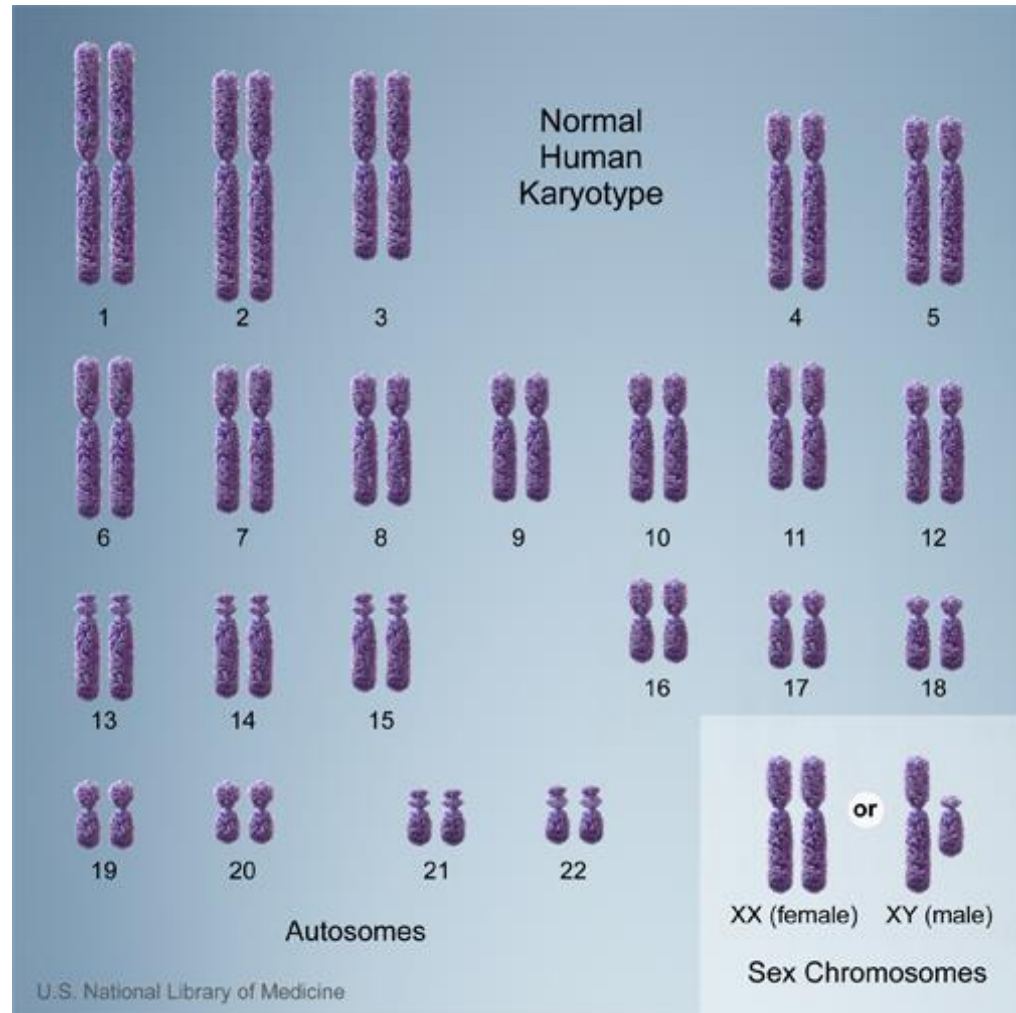
Why are some siblings more alike than others?

Recombination

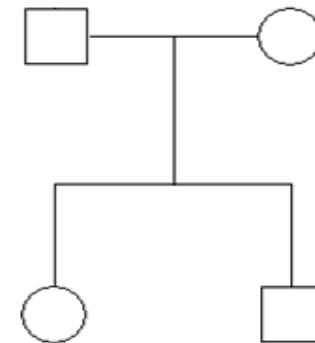
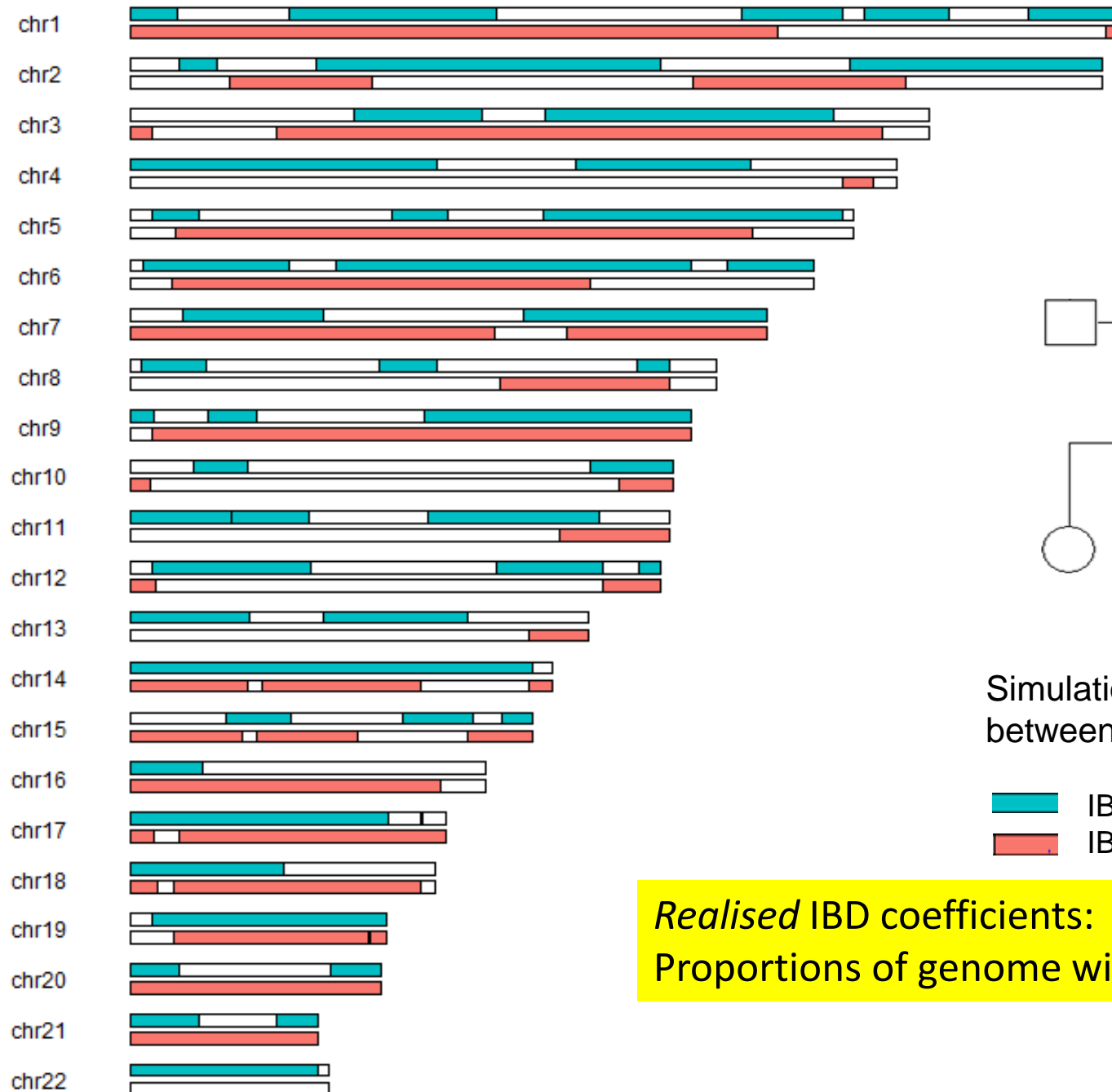


- **Genetic distance** between two loci:
= average # crossovers/ meiosis
- Units:
 - 1 Morgan (M) = 1 crossover per meiosis
 - 1 centiMorgan (cM) = 0.01 M
- The human genome: Ca 30 Morgan

Rule of thumb: One crossover per chromosome arm



IBD: Full siblings

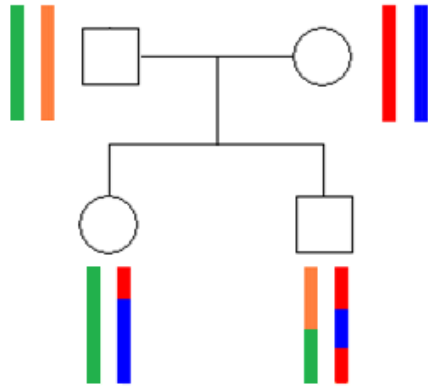


Simulation of IBD sharing
between full siblings

IBD from father
IBD from mother

Realised IBD coefficients:
Proportions of genome with IBD = 0, 1, 2

Variation in realised IBD coefficients

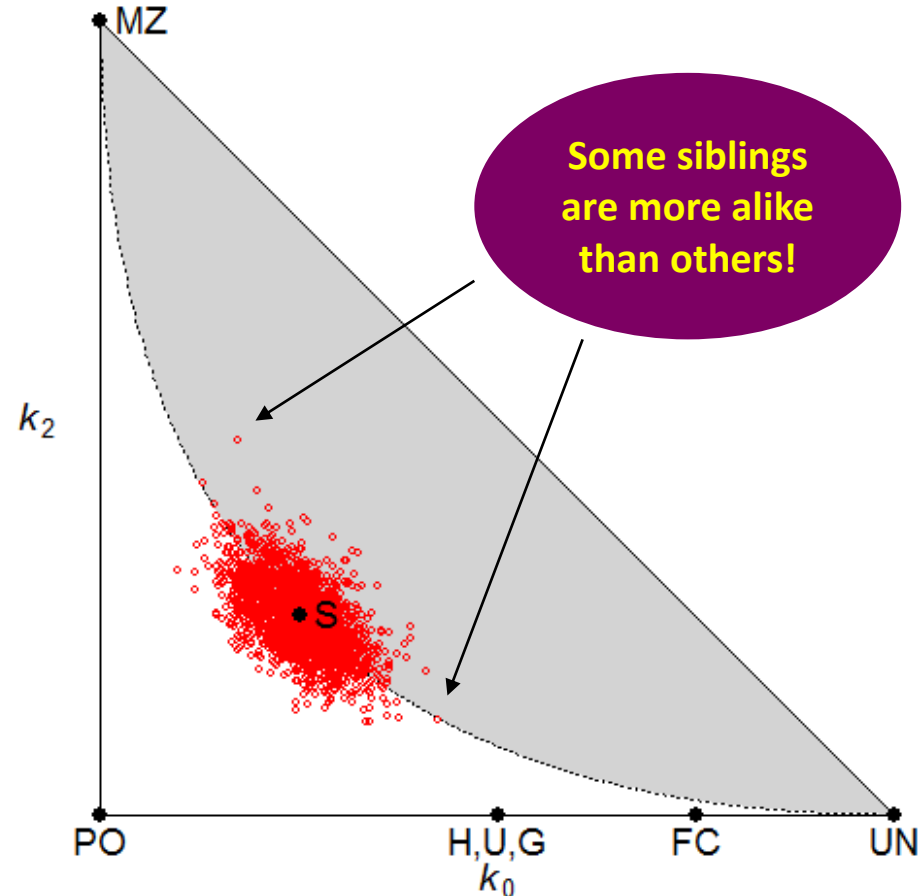


- 1000 simulations
- Decode recombination map
- All 22 human autosomes

```
> library(ibdsim2)

> x = nuclearPed(2)
> s = ibdsim(x, N = 1000)
> k = realisedKappa(s, ids = 3:4)

> ribd::showInTriangle(k)
```



Variation depends on the genome



Human:

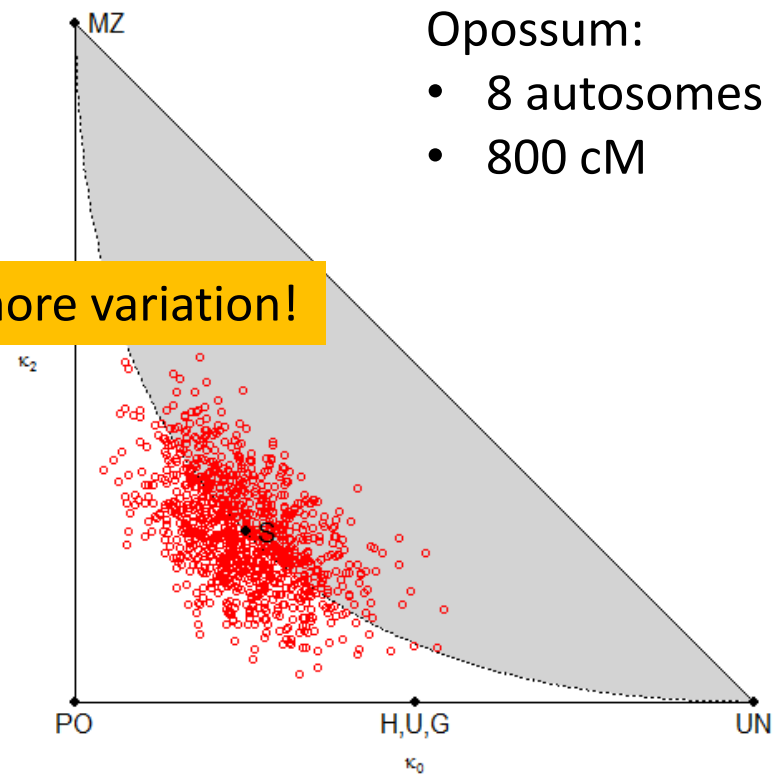
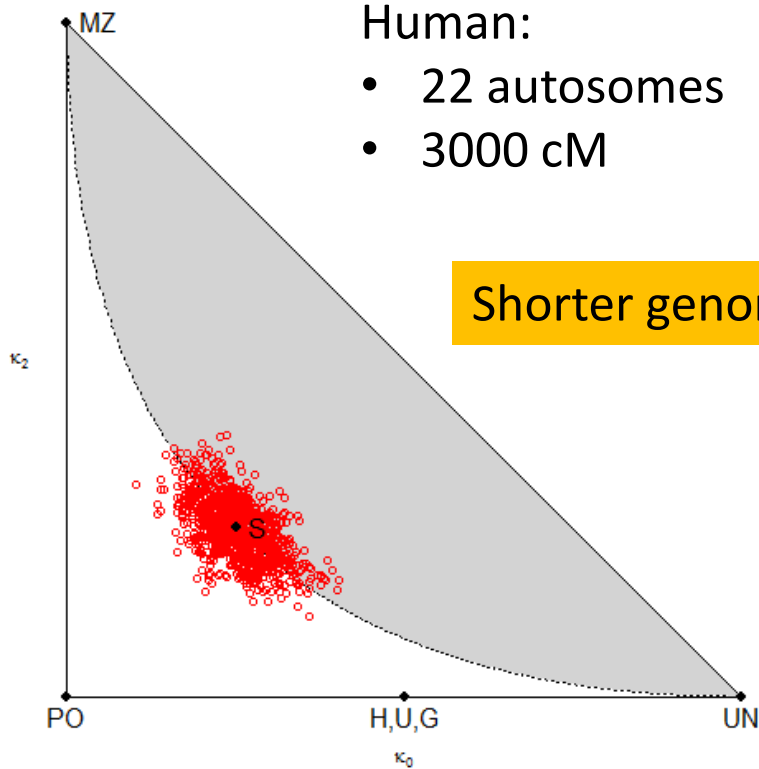
- 22 autosomes
- 3000 cM



Opossum:

- 8 autosomes
- 800 cM

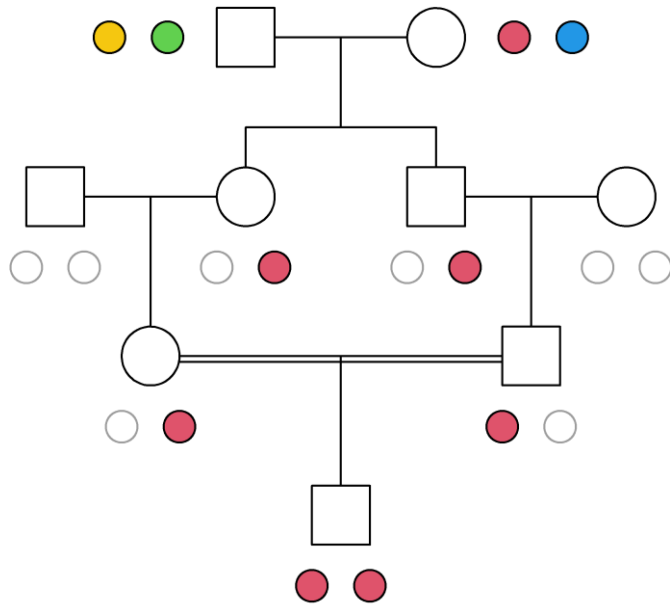
Shorter genome = more variation!



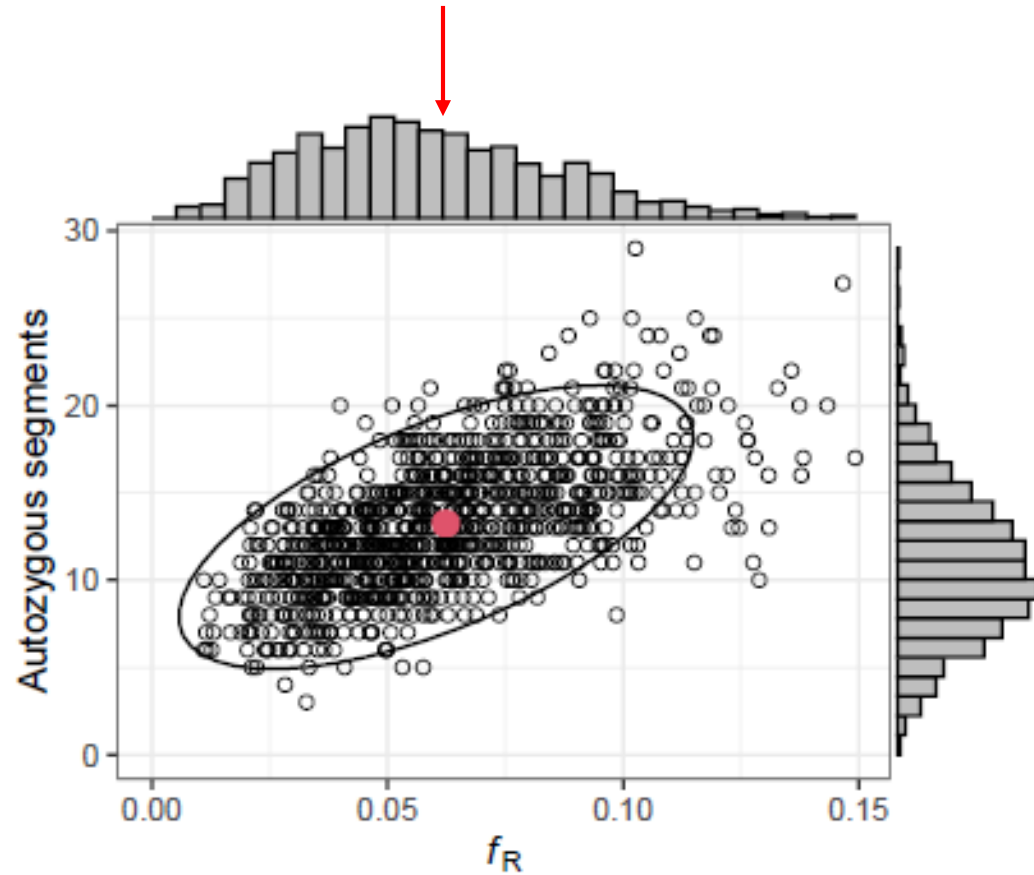
Variation in realised inbreeding



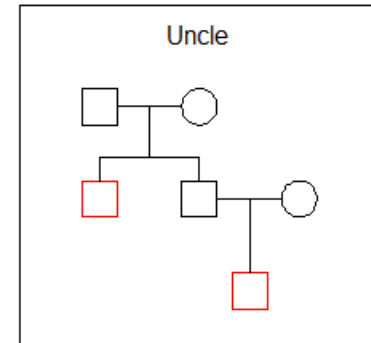
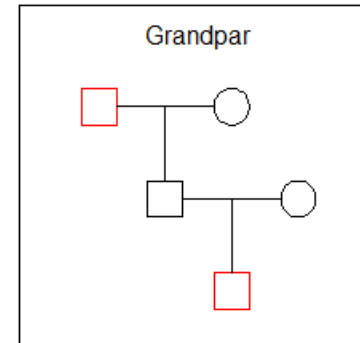
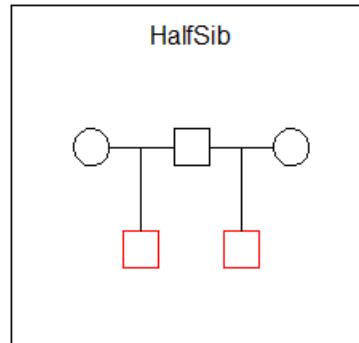
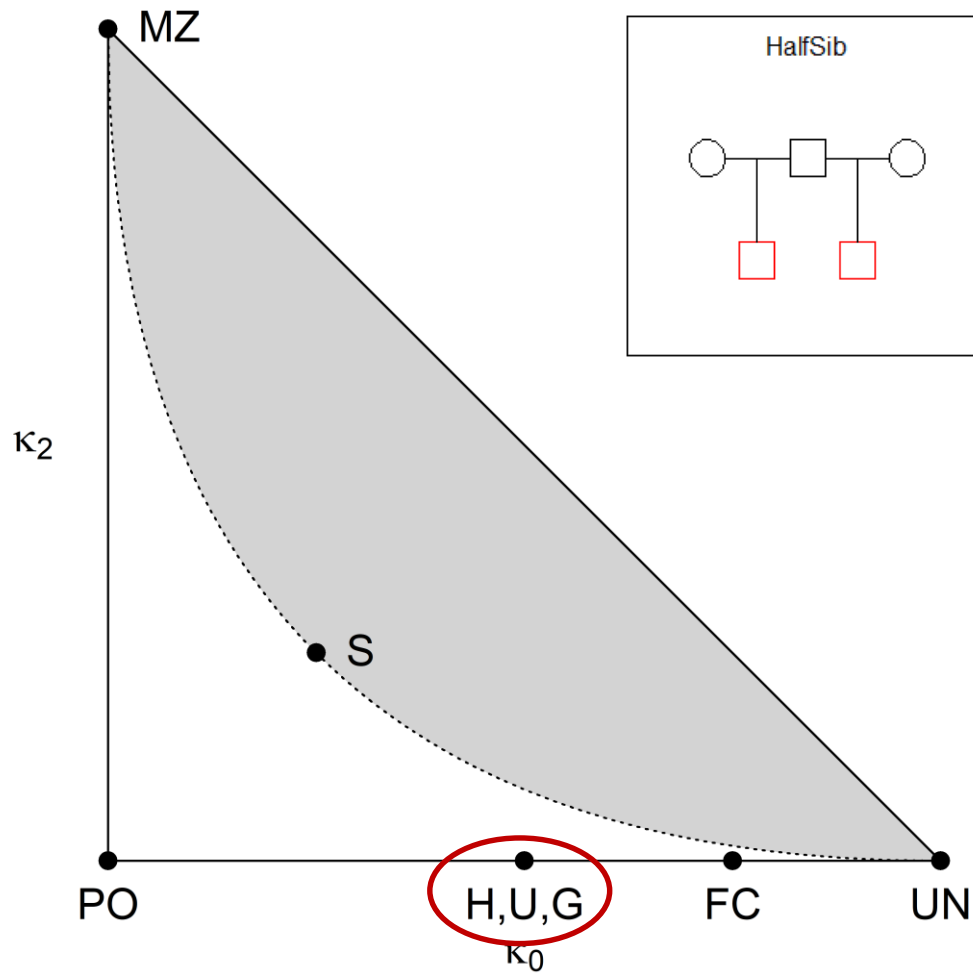
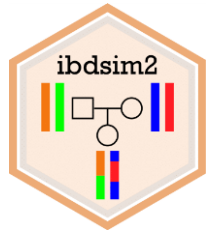
Expected: $f = 0.0625$



$f = 0.0625$

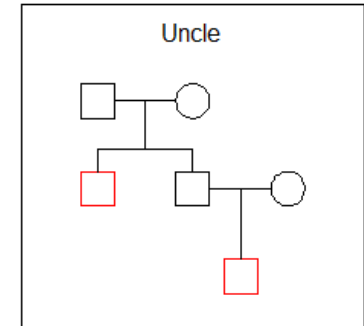
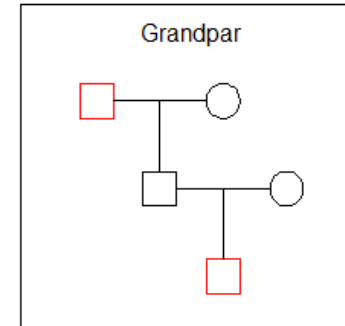
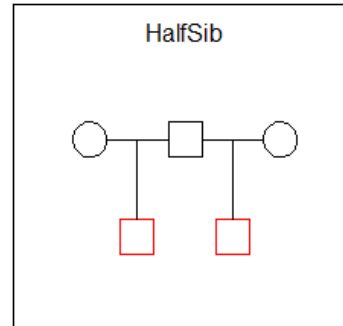
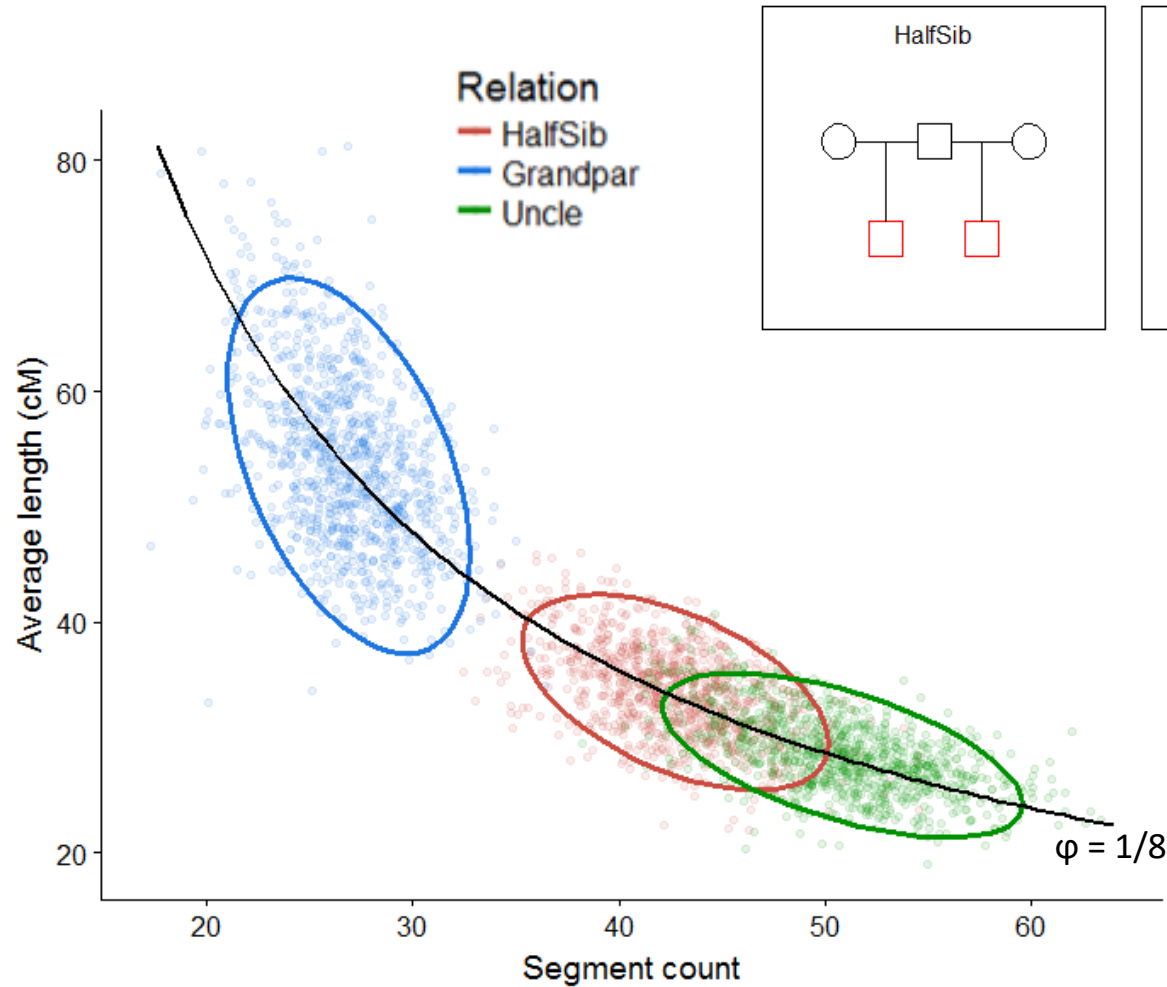


Indistinguishable relationships?



$$\begin{aligned}\kappa_0 &= 0.5 \\ \kappa_1 &= 0.5 \\ \kappa_2 &= 0\end{aligned}$$

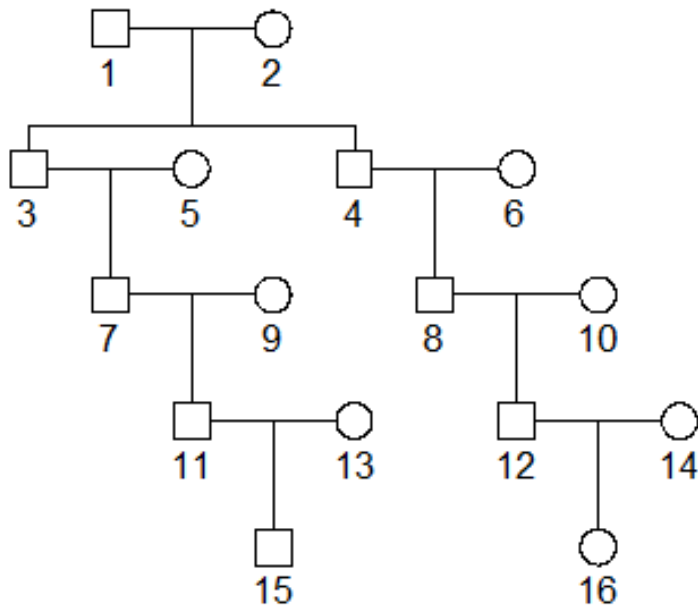
Simulated IBD distributions



Conclusion

In theory often distinguishable!
In practice quite hard.

The probability of zero IBD



N'th cousins	$P(\text{zero IBD})$
first	0.0 %
second	0.0 %
third	1.5 %
fourth	28 %
fifth	67 %

Third cousins: Expected fraction of the genome with shared DNA:

$$k_1 = 1/64$$

Two individuals can have a common ancestor without being genetically related

So...what does it mean to be related?

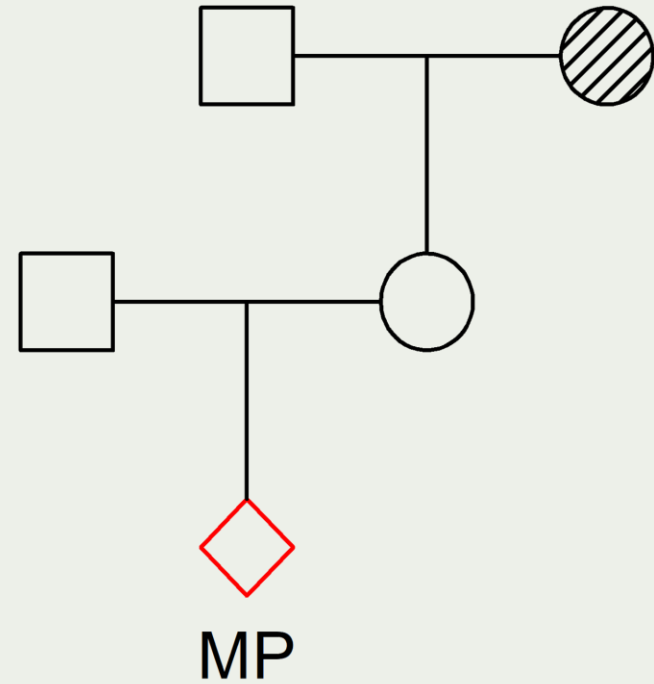
- Pedigree based definition: $\varphi > 0$
potentially having alleles IBD
- Genomic definition (**realised** relatedness):
actually having alleles IBD



Questions



Coffee break



Forensic application 1: Missing person identification

Argentina 1976 - 1983

- Military dictatorship
- *Dirty war* against left-wing guerrillas
- Opponents killed or disappeared
 - counts: 20,000 - 30,000



- 500 children abducted
 - kidnapped with their parents, or born in captivity
 - parents killed
 - raised by police or military families.



The missing grandchildren

- *Grandmothers of Plaza de Mayo*
 - formed in 1977
 - weekly marches ever since
- 1984: First grandchild recovered
 - HLA typing + blood groups
- 1989: National genetic data bank



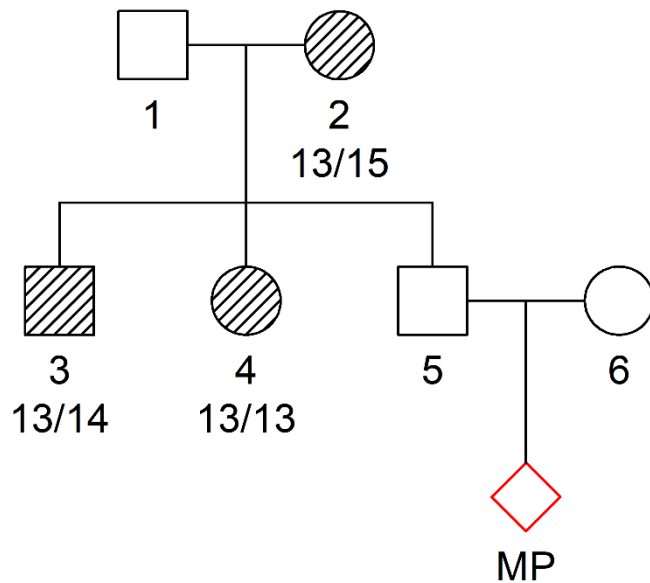
- 2022: 130 reunifications so far





Missing person cases: Basics

Reference family



Person of interest (POI)



match?



DNA-based identification

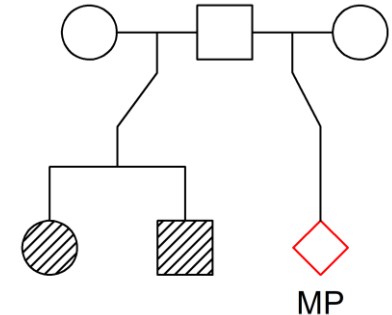
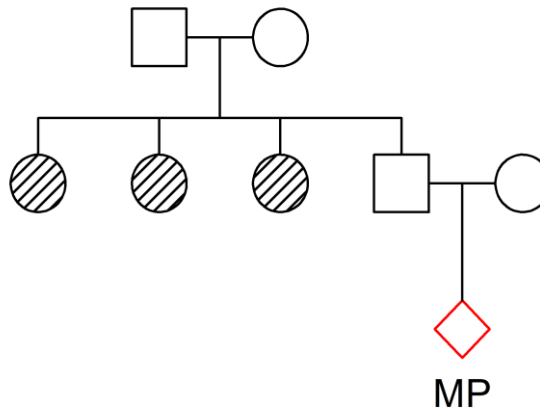
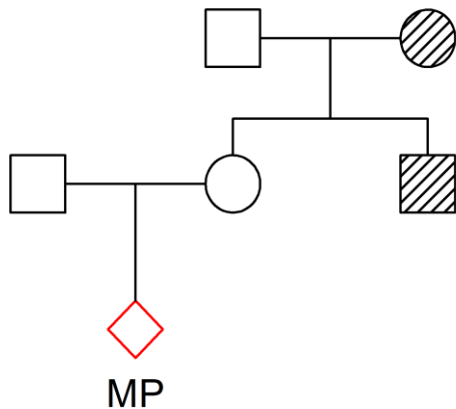
- Forensic markers
 - autosomal, X, Y, mtDNA
- Simplest with DNA from
 - the missing person
 - parents of the missing

Standard forensic kits

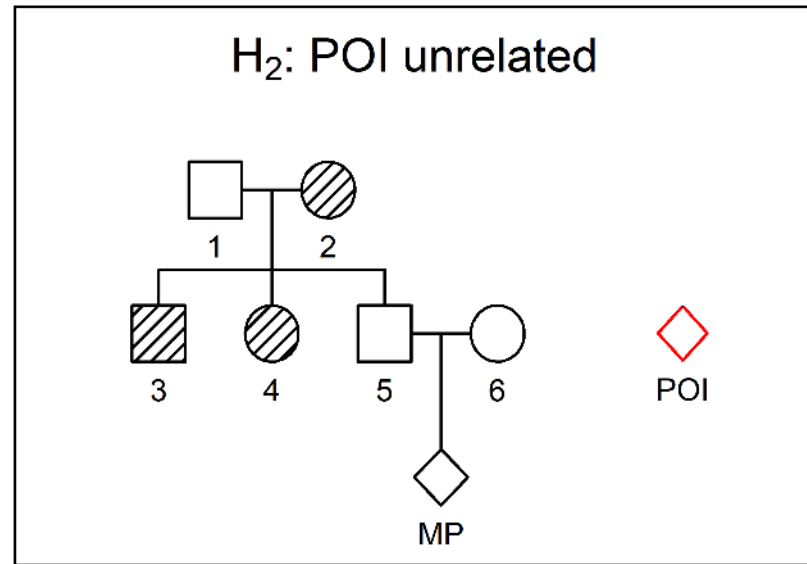
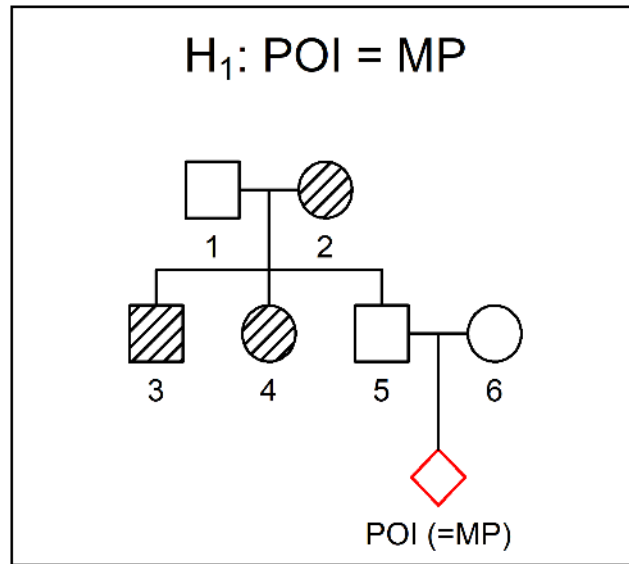
- 15 - 24 autosomal STRs
- Typically 10 - 50 alleles
- Mostly unlinked

← \approx paternity case

- Argentina: Parents usually unavailable



The likelihood ratio (LR)



$$LR = \frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_2)}$$

Positive match if $LR > 10,000^*$

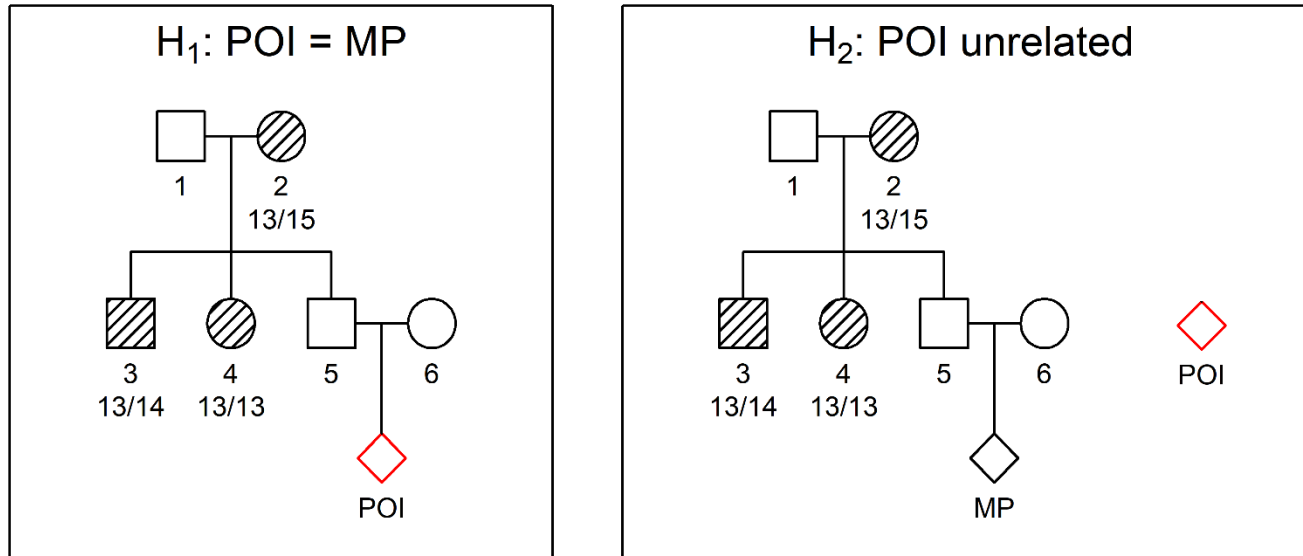
*or other suitable threshold

Pedigree likelihood - software

- Familias
 - Egeland, Mostad et al, 2000
 - Currently maintained by Daniel Kling
- R/ped suite
 - Very flexible
 - Great for plotting
 - Advanced analysis
 - Example: **Power analysis**



Power in missing person cases

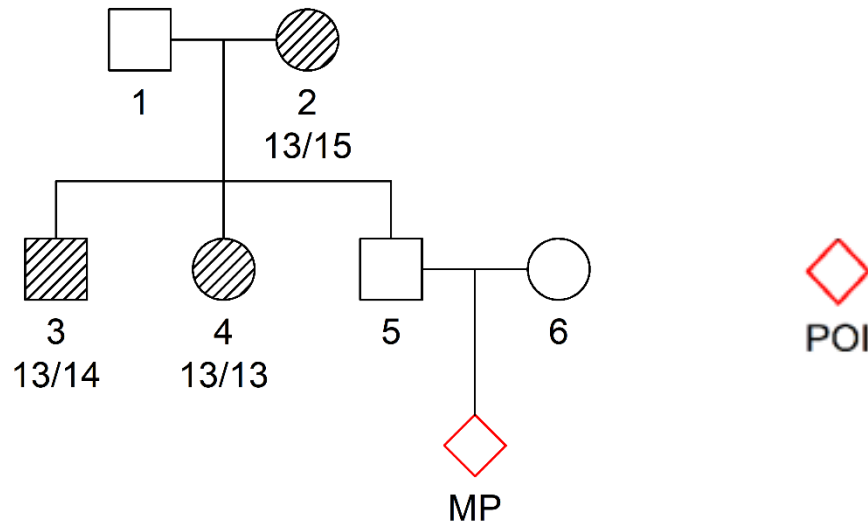


- Two complementary measures of power
 - **Inclusion**: The probability of recognizing the true MP
 - **Exclusion**: The probability of excluding an unrelated POI
- Note: Computed before POI is genotyped!

Depend on

- Reference individuals
- Reference genotypes
- Number of markers
- Allele frequencies

Inclusion power (IP)

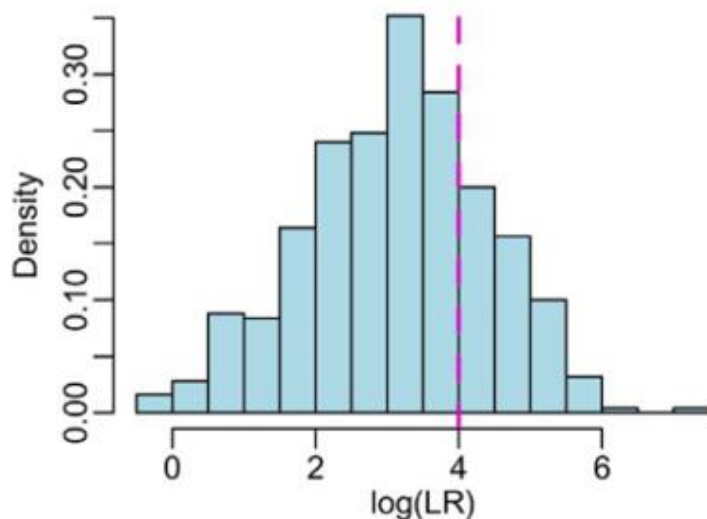
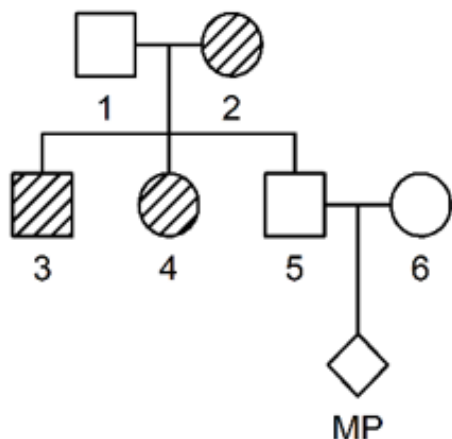


- If POI = MP: Do we have enough data to detect it?

$$IP_{10000} = P(LR > 10,000 \mid POI = MP)$$

- Computed by simulations of MP - conditional on the reference

Inclusion power in R



R code

```
> library(forrel)

> ref = readFam(...)
> missingPersonIP(ref, nsim = 500,
                  threshold = 10000)
```

Using all 20 attached markers
Simulating 500 profiles...done
Computing likelihood ratios...done
Total time used: 9.87 secs

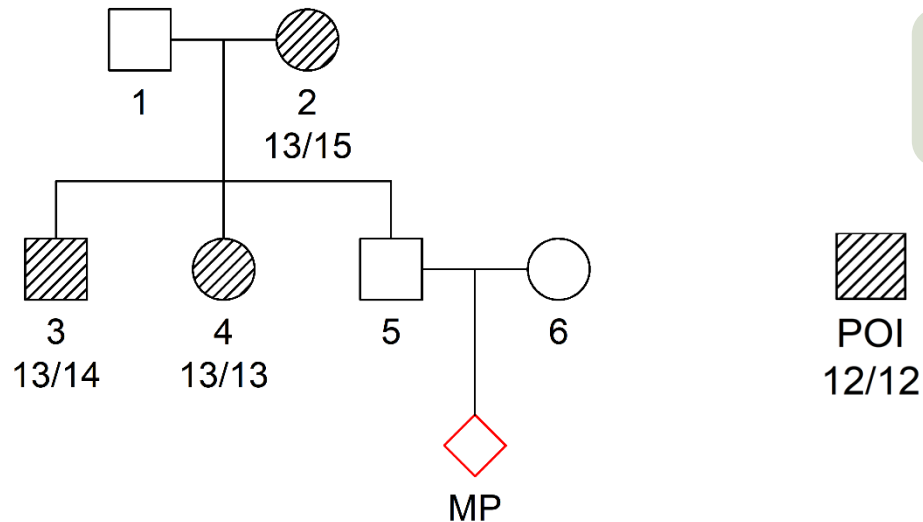
Mean LR: 65947.04

Mean log₁₀(LR): 3.121

Estimated power:

$P(\text{LR} \geq 10000) = 0.248$

Exclusion power (EP)



MP cannot be 12/12!
→ POI excluded

- **If $POI \neq MP$:** Probability of mismatch in at least 1 marker?

$$EP = P(\text{exclusion} \mid POI \text{ unrelated})$$

- Can be computed exactly!

Egeland, Pinto, Vigeland (2014). *A general approach to power calculation for relationship testing*

Exclusion power in R



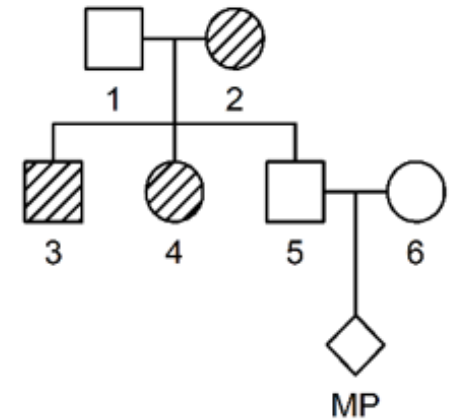
R code

```
> missingPersonEP(ref, missing = "MP")
```

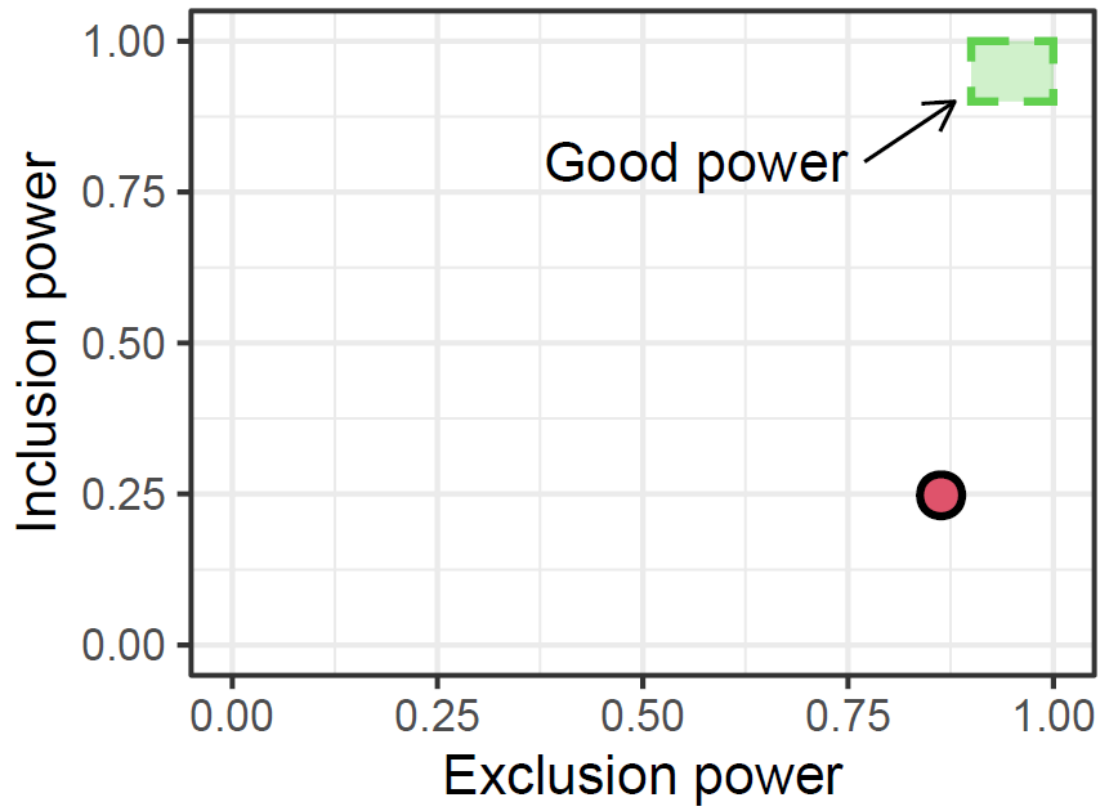
```
Potential mismatches: 8 (D3S1358, D7S820, CSF1PO,  
PENTA_D, VWA, TPOX, D19S433, D2S1338)
```

```
Expected mismatches: 1.679
```

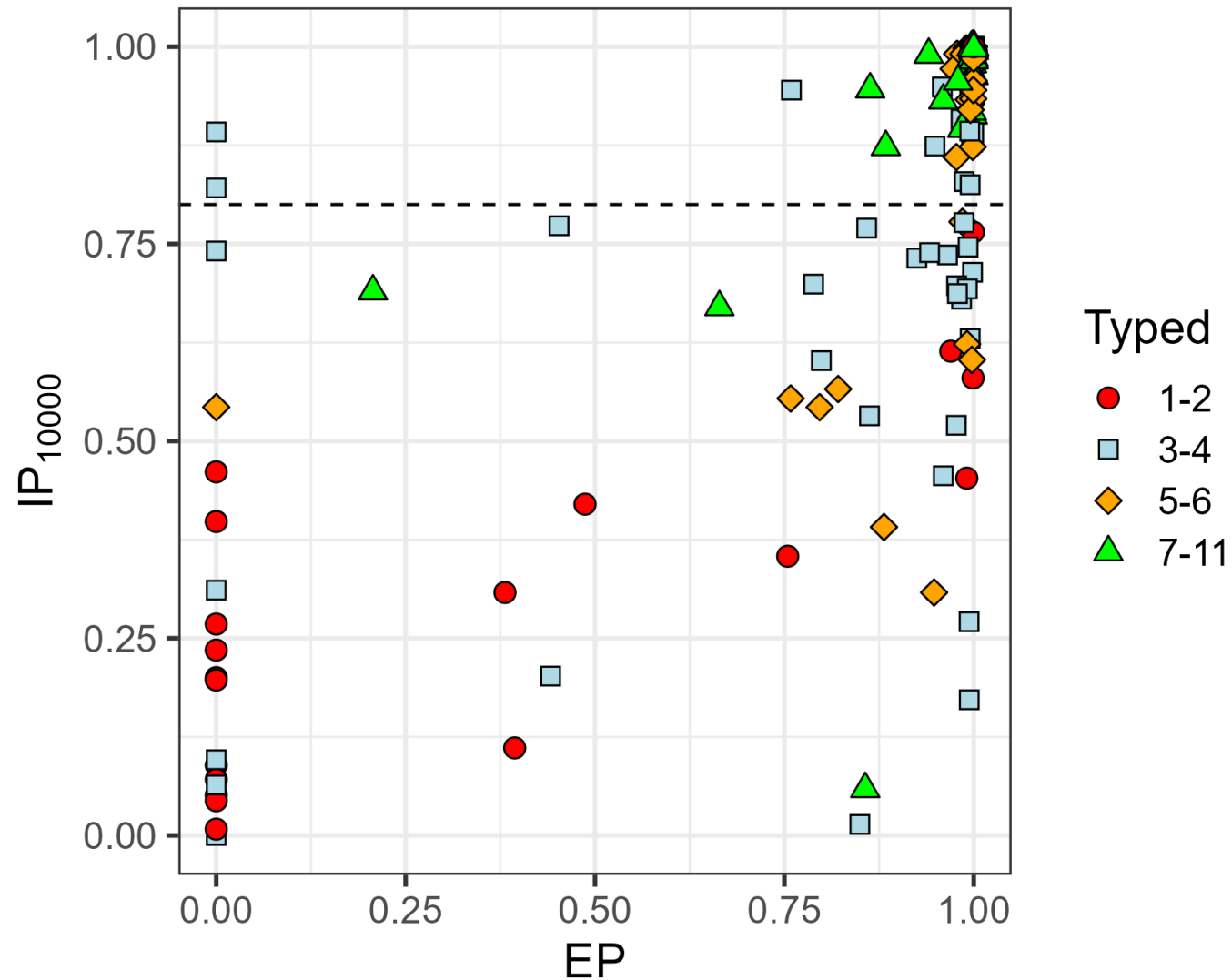
```
P(at least 1 mismatch): 0.863
```



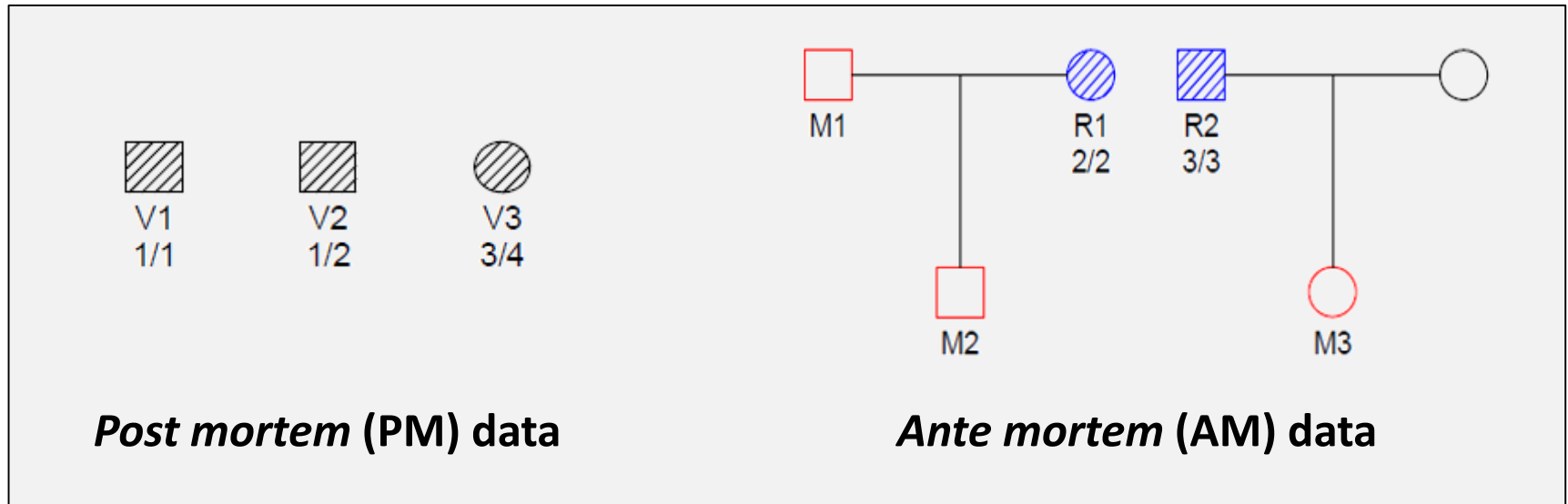
Power plot



Argentina: Power analysis of 200 unsolved cases



Disaster victim identification (DVI)



High-profile DVI cases

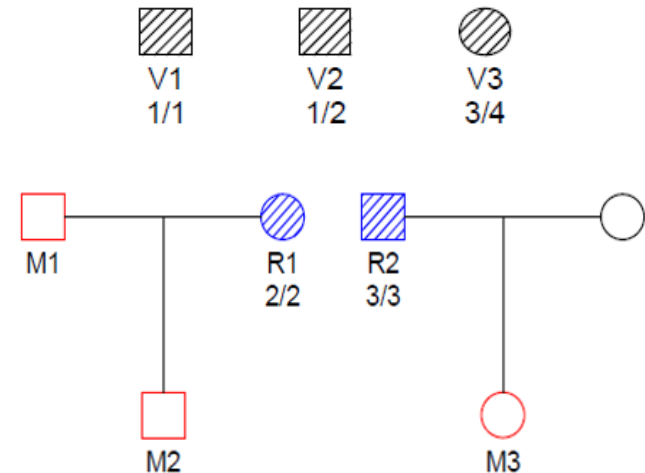
- World Trade Center attack
- Drowned immigrants
- Thailand tsunami



Disaster Victim Identification in R



- Traditional approach/software
 - One victim or family at a time
 - Manual sequential analysis
- Further possibilities in R (**dvir**)
 - Joint analysis!
- Key functions
 - `pairwiseLR()`
 - `jointDVI()`



	V ₁	V ₂	V ₃	loglik	LR	posterior
1	M ₁	M ₂	M ₃	-16.12	250.00	0.72
2	M ₁	M ₂	*	-17.73	50.00	0.14
3	*	M ₂	M ₃	-18.42	25.00	0.07
4	M ₁	*	M ₃	-20.03	5.00	0.01
5	*	M ₁	M ₃	-20.03	5.00	0.01
6	*	M ₂	*	-20.03	5.00	0.01
7	*	*	M ₃	-20.03	5.00	0.01
8	M ₁	*	*	-21.64	1.00	0.00
9	*	M ₁	*	-21.64	1.00	0.00
10	*	*	*	-21.64	1.00	0.00

R code

```
> library(dvir)

> pm = example2$pm
> am = example2$am
> missing = example2$missing

> jointDVI(pm, am, missing)
```

Biased selection of references



Thore

Article | [Open Access](#) | Published: 01 July 2021

Joint DNA-based disaster victim identification

Magnus D. Vigeland & Thore Egeland

Scientific Reports **11**, Article number: 13661 (2021) | [Cite this article](#)

Contents lists available at [ScienceDirect](#)

Forensic Science International: Genetics

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

Research paper

Prioritising family members for genotyping in missing person cases: A general approach combining the statistical power of exclusion and inclusion

Magnus D. Vigeland ^{a,*}, Franco L. Marsico ^b, Mariana Herrera Piñero ^b, Thore Egeland ^c



Mariana

Forensic Science International: Genetics

Volume 31, November 2017, Pages 57–66

Research paper

Evaluating the statistical power of DNA-based identification, exemplified by ‘The missing grandchildren of Argentina’

Daniel Kling ^a , Thore Egeland ^b , Mariana Herrera Piñero ^c , Magnus Dehli Vigeland ^d



Daniel



Lourdes

