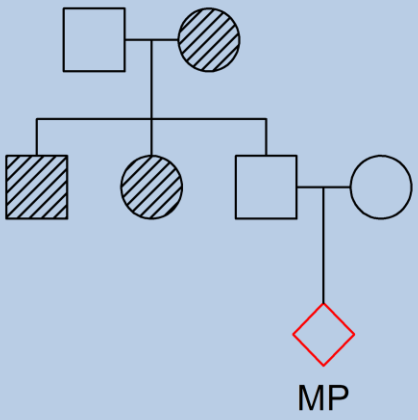


Statistical power in missing person cases

Lecture 1: Do we have enough data?

Magnus Dehli Vigeland

August 18th 2021

PROYECTO DE COOPERACIÓN TRIANGULAR. ARGENTINA - COLOMBIA – UNIÓN EUROPEA.

Overview

1. Introduction to missing person cases
   - Motivating case: *Missing grandchildren of Argentina*
   - Terminology
   - Genetics
   - Likelihood ratio
   - Software

2. Power
   - Inclusion power
   - Exclusion power
   - Power plots
   - Examples from BNDG

How I got involved

Mariana

Franco

Forensic Science International: Genetics
Volume 31, November 2017, Pages 57-66

Research paper

Evaluating the statis
identification, exem
grandchildren of Ar

Daniel Kling [a], Thore Egeland [b], Mariar

Forensic Science International: Genetics
Volume 49, November 2020, 102376

Research paper

Prioritising family r
missing person case
combining the stati
inclusion

Magnus D. Vigeland [a], Franco L. Marsico

Forensic Science International: Genetics
Volume 54, September 2021, 102519

Research paper

Making decisions in missing person identification cases with low statistical power

Franco L. Marsico [a], Magnus D. Vigeland [b], Thore Egeland [c], Mariana Herrera Piñero [a]

Thore

Daniel

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

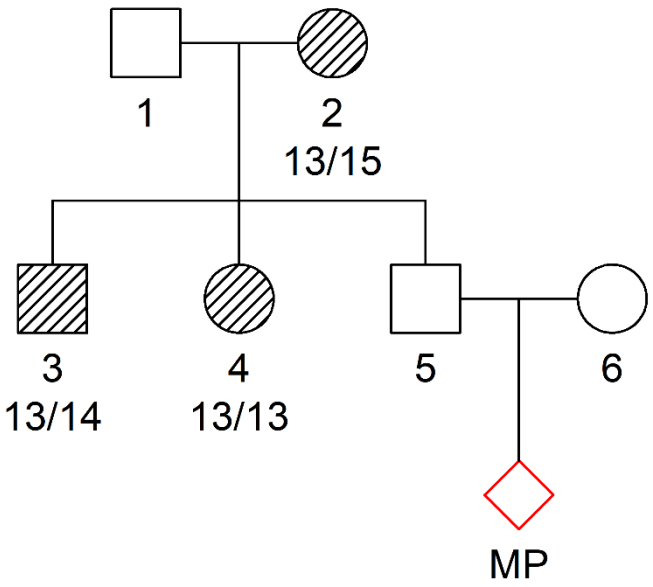
BNDG
BANCO NACIONAL DE DATOS GENÉTICOS

- 2021: 130 reunifications so far


JUICIO Y
CASTIGO
NO
DUDA

Missing person cases: Basics

Reference family



match?

Person of interest (POI)

Currently in BNDG

- ~300 reference fams
- ~10 000 POIs

DNA-based identification

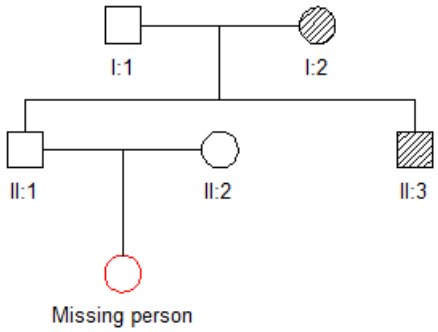
- DNA-based evidence
  - autosomal markers
  - mtDNA
  - Y chromosome

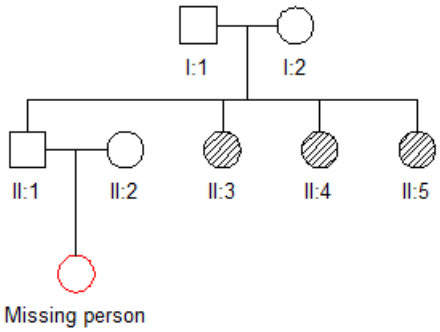
Forensic markers:
- standard kits, 15 - 24 STRs
- up to 50 alleles
- mostly unlinked

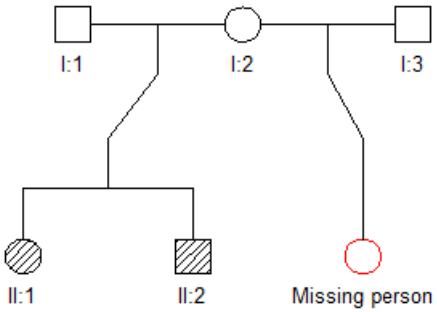
- Simplest when
  - available DNA from the missing person
  - available DNA from parents

Similar to standard paternity cases

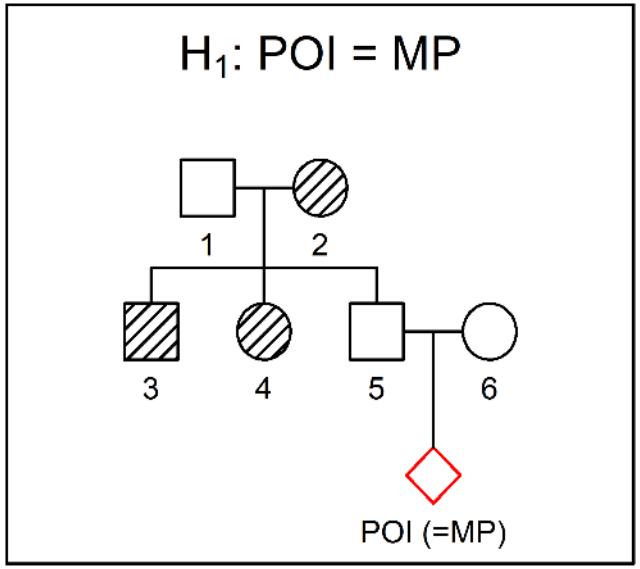
- Argentina: Parents usually unavailable

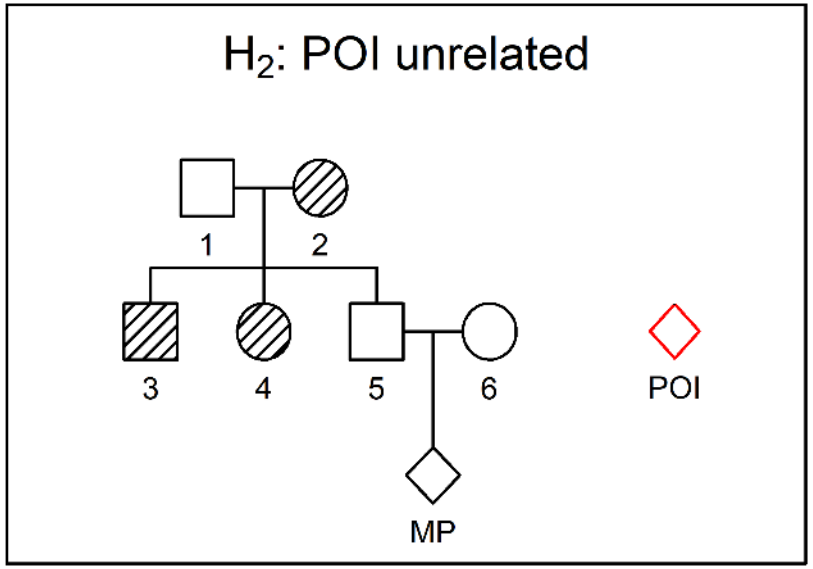






The likelihood ratio (LR)



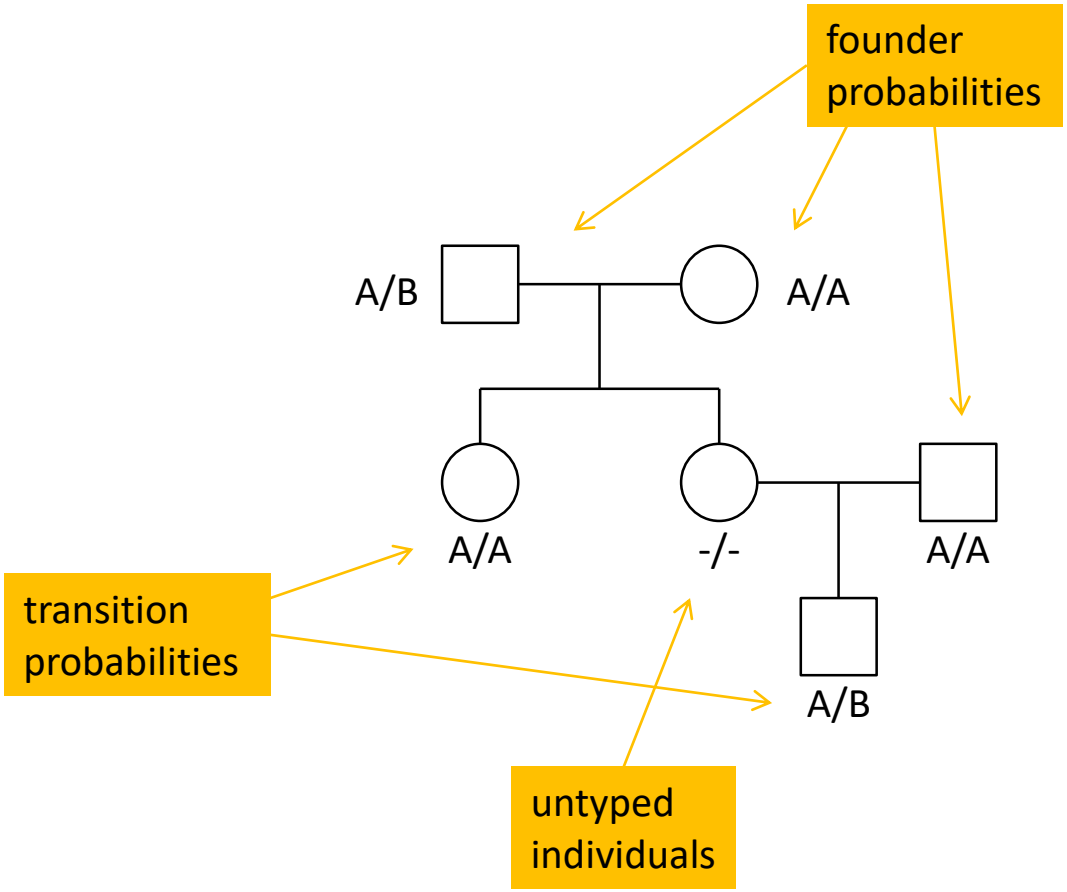


$$LR = \frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_2)}$$

Positive match if LR > 10,000*

*or other suitable threshold

Likelihood computations

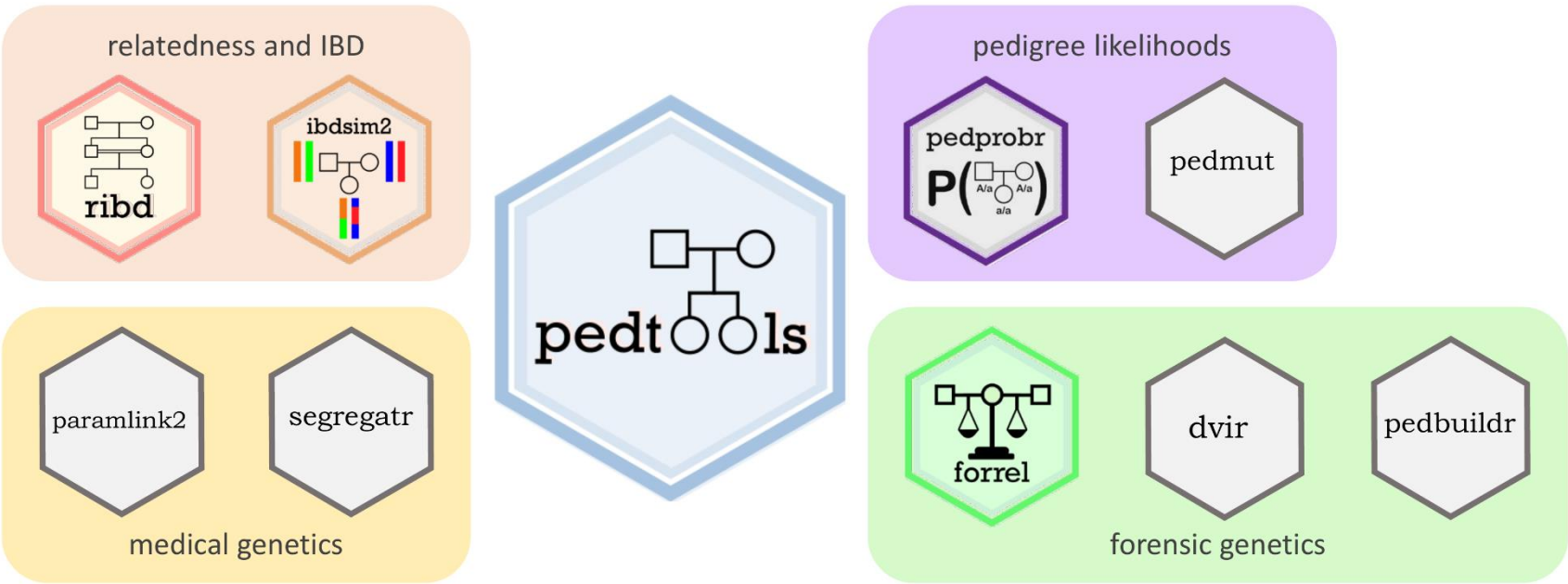


Software

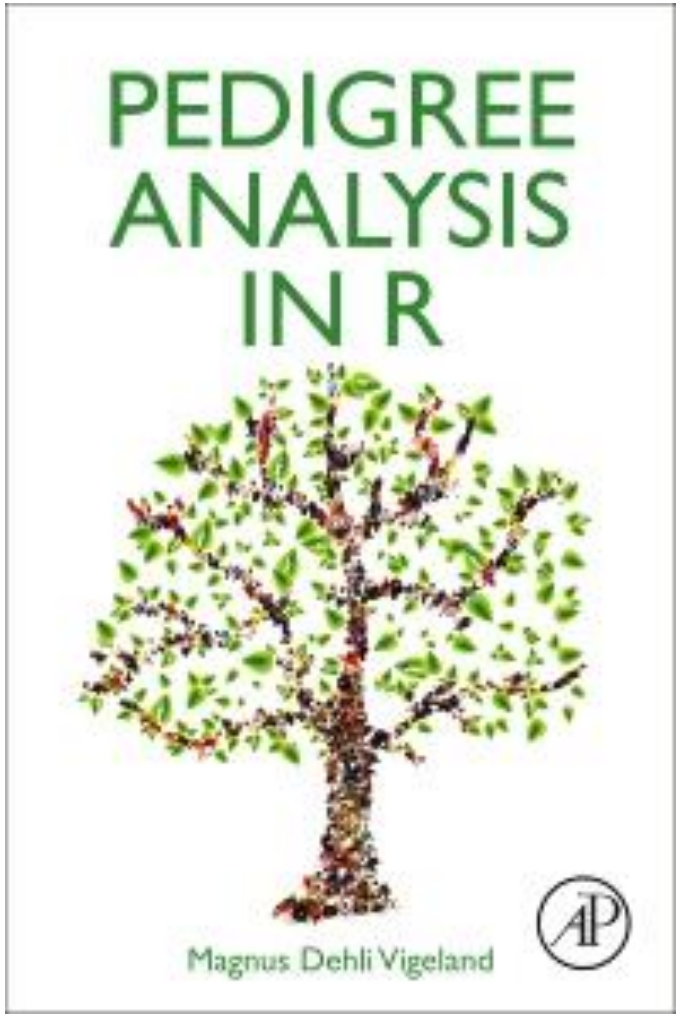
- Familias
  - Original publication: Egeland, Mostad et al, 2000
  - Currently maintained by Daniel Kling
  - Used by BNDG
- R/ped suite
  - Very flexible
  - Great for plotting
  - The latest research!
  - (but requires some programming)
- QuickPed
  - Online tool for creating pedigrees
  - Handy for reports and presentations

The **ped suite**

A collection of packages for pedigree analysis in R



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



Academic Press, 2021

Available in most online book stores, Amazon etc.

Chapter 6

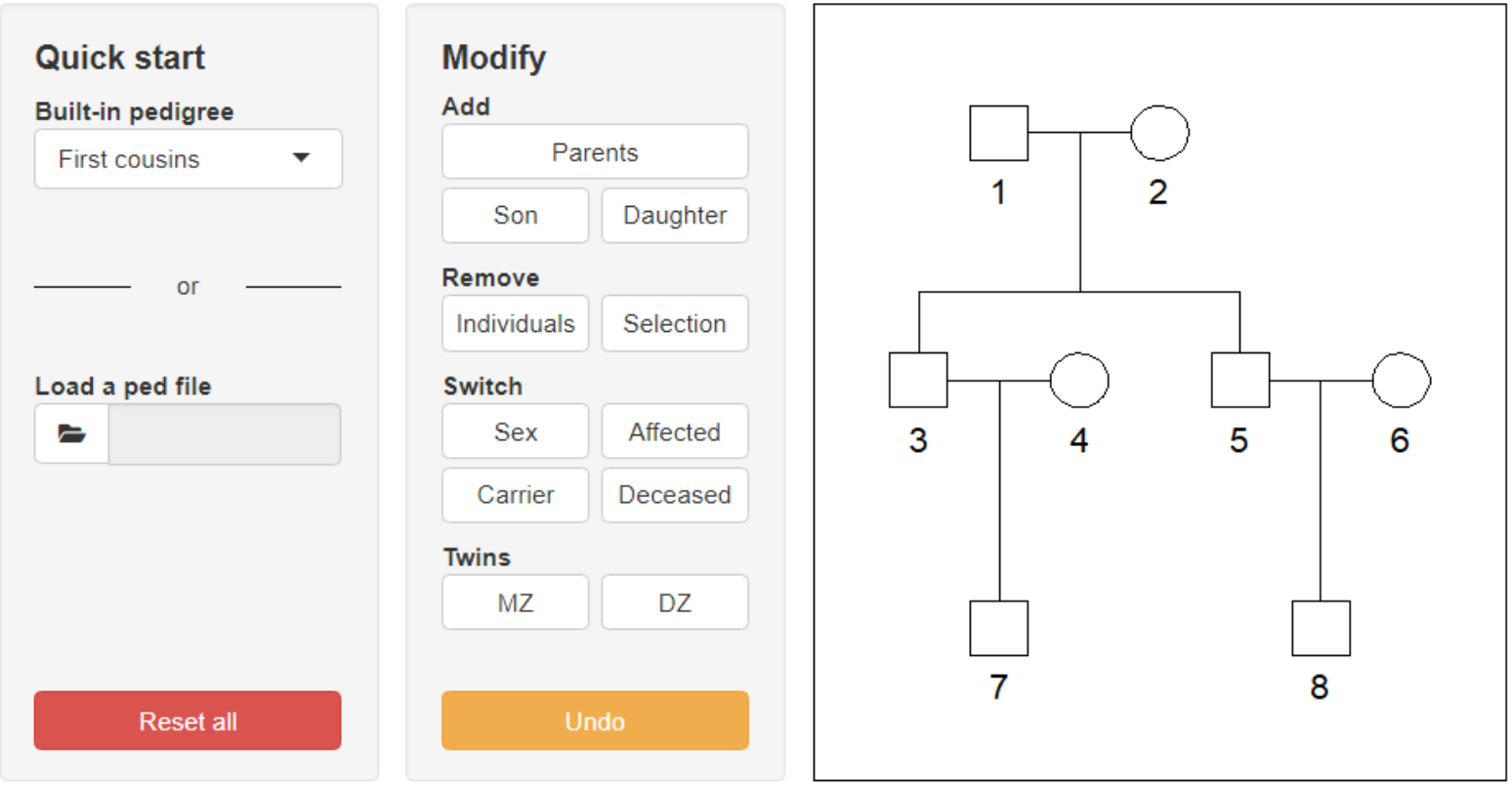
- Kinship testing
- Missing person cases

QuickPed

https://magnusdv.shinyapps.io/quickped

Quick demo!

- A free online tool for building pedigrees



Questions

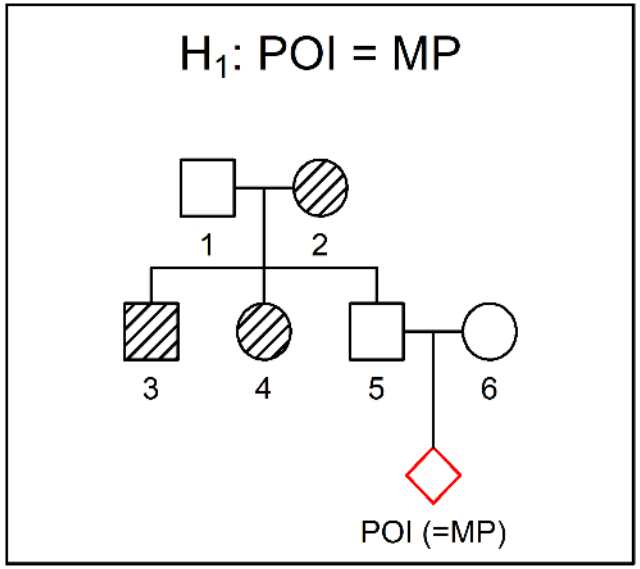
Coffee break

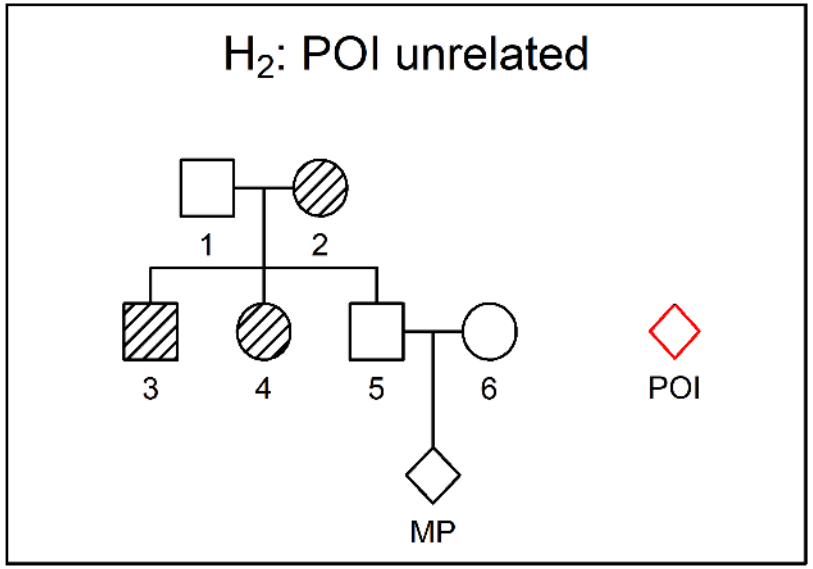
Overview

1. Introduction to missing person cases
   - Motivating case: *Missing grandchildren of Argentina*
   - Terminology
   - Genetics
   - Likelihood ratio
   - Software

2. Power
   - Inclusion power
   - Exclusion power
   - Power plots
   - Examples from BNDG

Where we left off...





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

Positive match if LR > 10,000*

*or other suitable threshold

Statistical power: Brief recap

- Classical hypothesis testing

$$H_0\text{:}\ \ P(\text{head}) = 0.5$$
$$H_A\text{:}\ \ P(\text{head}) = 0.7$$

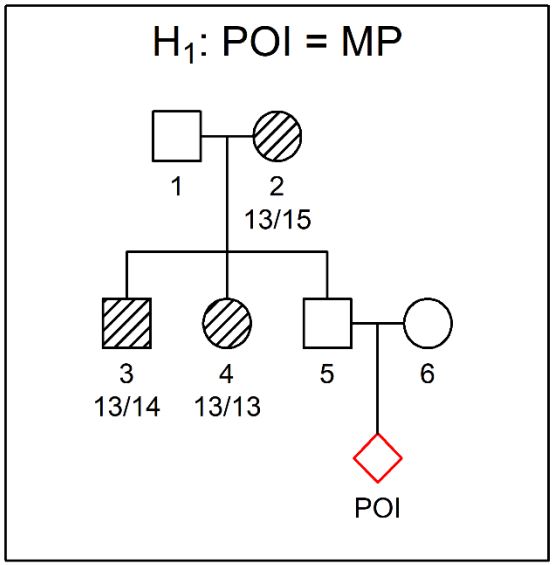
- Procedure:
  - Flip 30 times
  - Reject H_0 if $\#\text{heads} \geq 20$
  - (Gives significance level $\alpha \approx 0.05$)

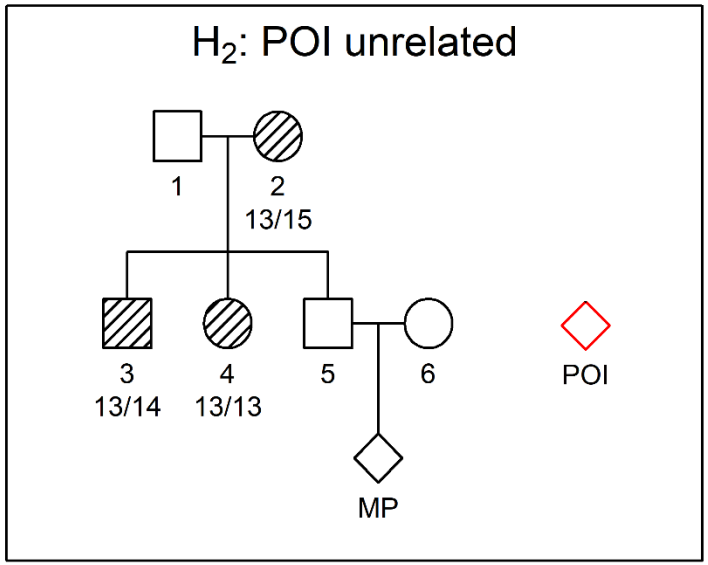
- If the coin is fake, what is the probability that we detect it?

$$power = P(\#\text{heads} \geq 20 \mid H_A)$$

- In this case: power = <u>0.73</u>

Power in missing person cases

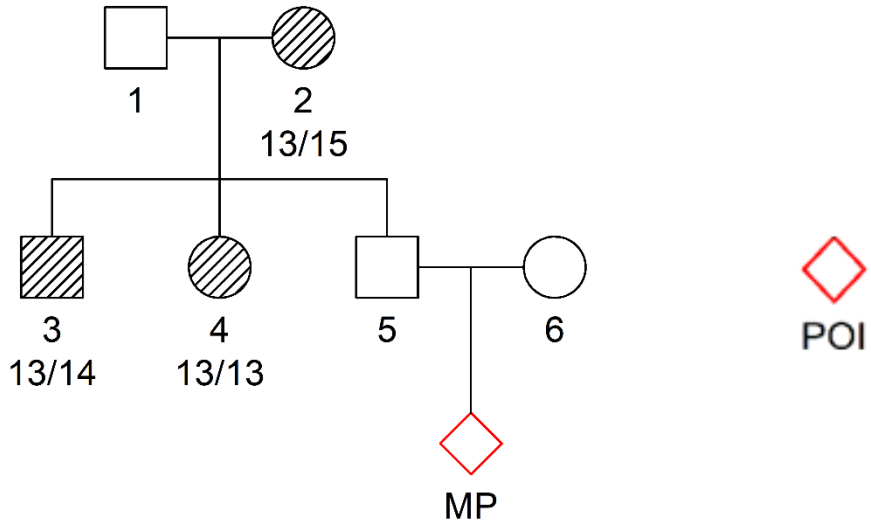




- Two complementary measures of power
  - Inclusion: The probability of recognizing the true MP
  - Exclusion: The probability of excluding an unrelated POI

- Note: Computed <u>before</u> POI is genotyped!

Inclusion power (IP)



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

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

also known as the exceedence probability E_{10000}

- Computed by simulation
  - Unconditional → average for pedigrees of this type
  - Conditional → probability for this particular case

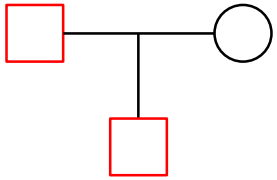
Depends on

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

Unconditional simulation

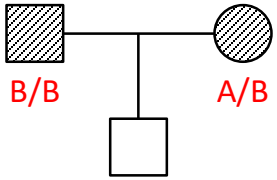
SNP: $p_A = p_B = 0.5$

A/A	A/B	B/B
0.25	0.5	0.25

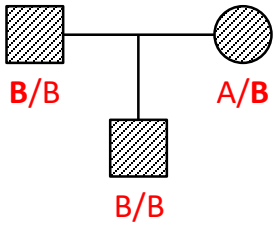


A/A	A/B	B/B
0.25	0.5	0.25

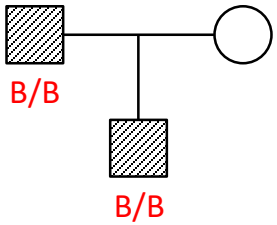
1. Founders (HW)



2. Gene dropping

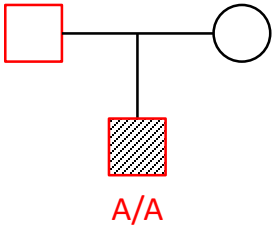


3. Remove untyped



Conditional simulation

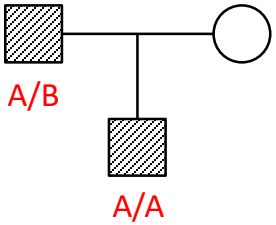
SNP: $p_A = p_B = 0.5$



1. Compute <u>conditional</u> distribution in the father

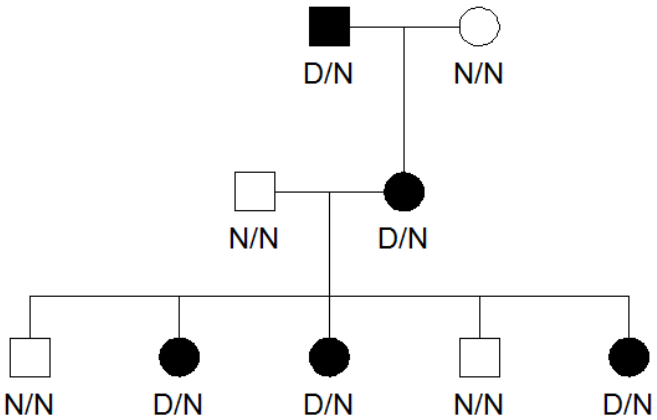
A/A	A/B	B/B
0.5	0.5	0

2. Sample from this



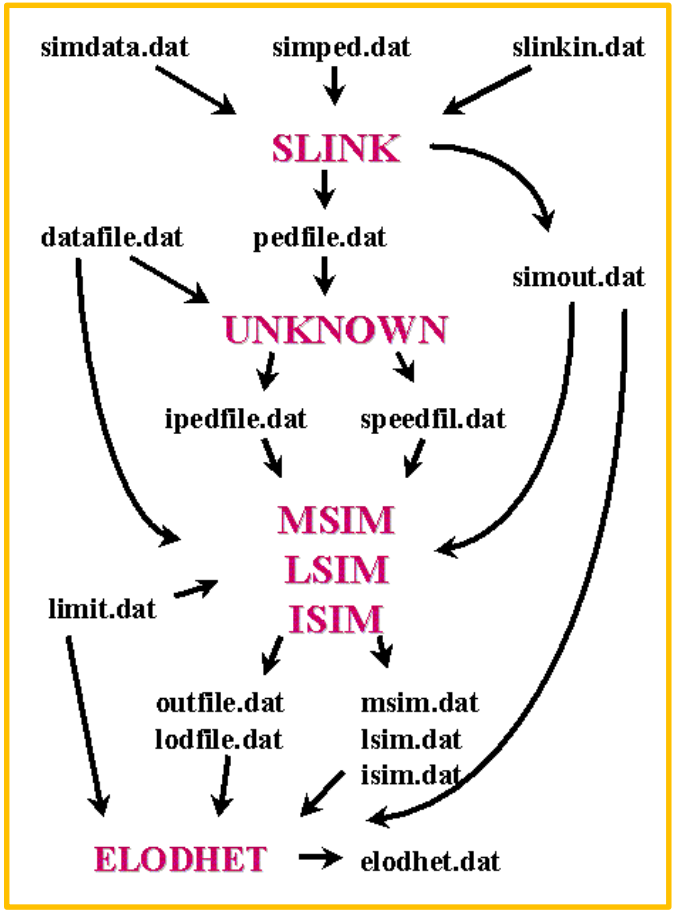
More computer intensive than unconditional

Conditional sims: first done in linkage analysis



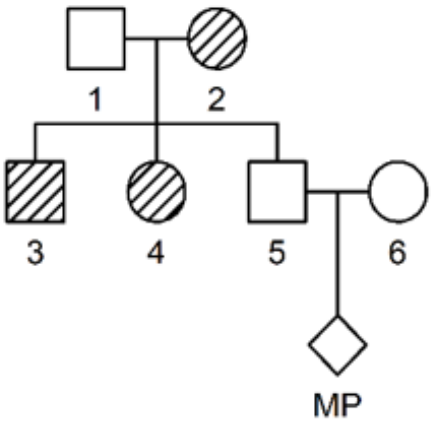
- Power analysis for linkage requires simulation
  - conditional on disease genotype
  - conditional on the distance to disease locus
- Weeks, Ott, Lathrop (1990)
  - SLINK: a general simulation program for linkage analysis

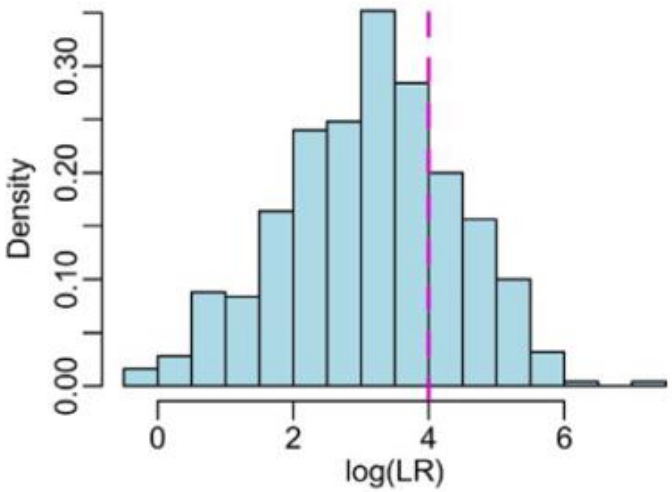
Not for the faint of heart...



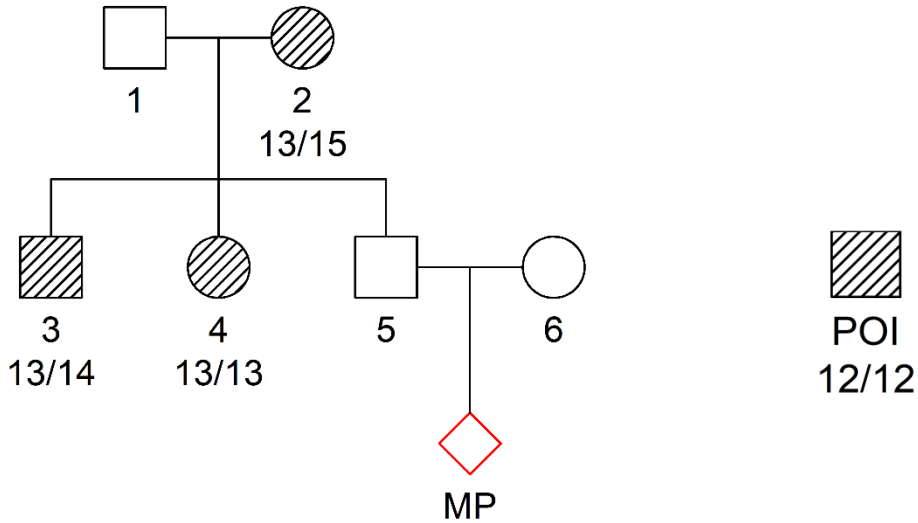
Inclusion power in R

```
> library(forrel)

> ref = readFam(...)
> missingPersonIP(ref, missing = "MP",
                  nsim = 500, seed = 42,
                  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(LR >= 10000) = 0.248
```





Exclusion power (EP)



MP cannot be 12/12!
→ POI excluded

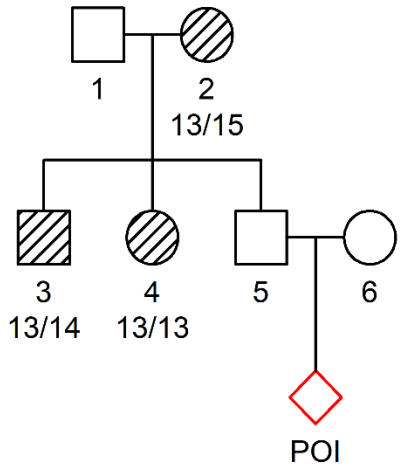
- If POI ≠ MP: What is the probability of mismatch in at least 1 marker?

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

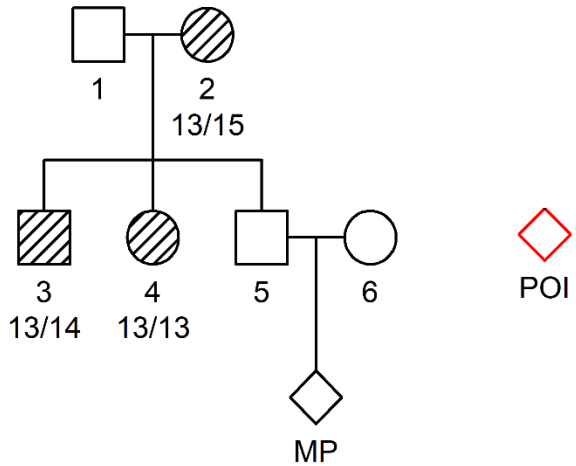
- Can be computed exactly!

The exclusion power formula

Claim: H_1



True: H_2



Single marker: $EP_i = P(\text{mismatch in } H_1 \text{ for marker } i \mid H_2)$

$$= \sum_g I(g \mid H_1) \cdot P(g \mid H_2)$$

genotype of POI

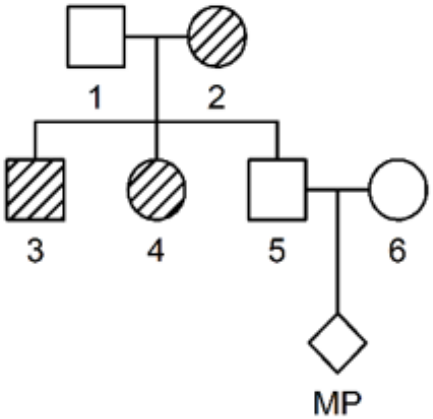
$$= \begin{cases} 1, \text{ if } g \text{ incompat with } H_1 \\ 0, \text{ otherwise} \end{cases}$$

Total power: $EP = 1 - \prod(1 - EP_i)$

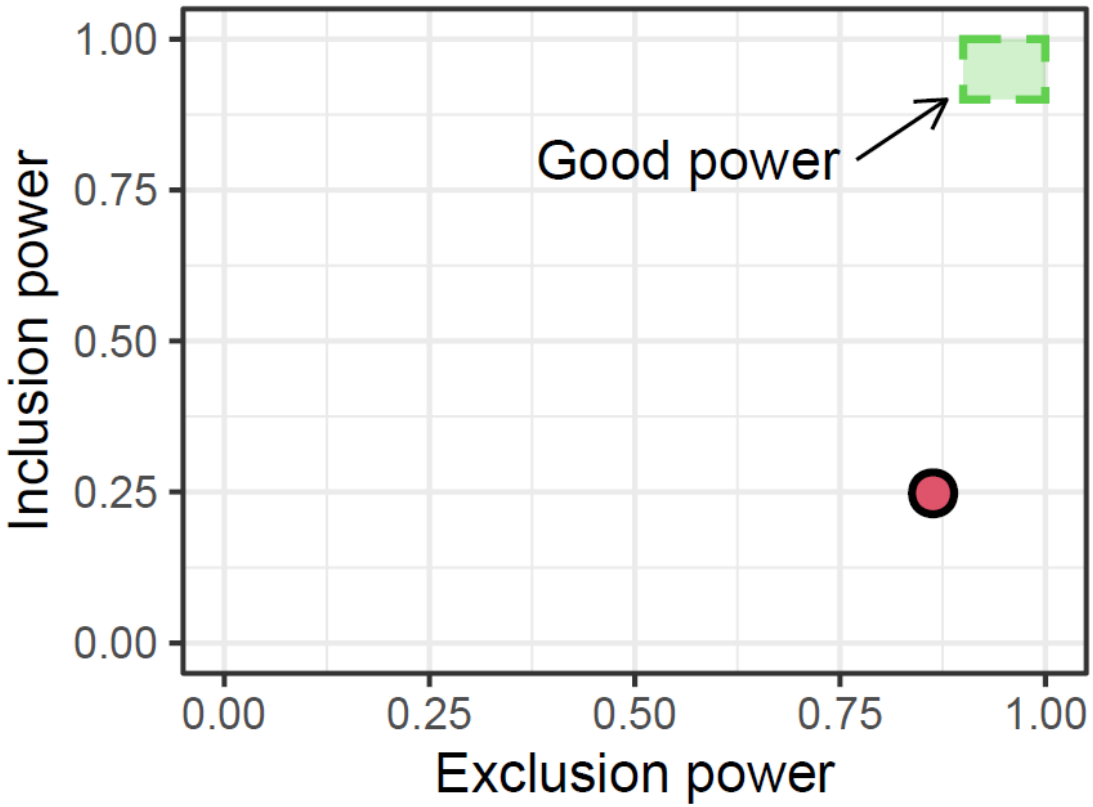
Egeland, Pinto, Vigeland. *A general approach to power calculation for relationship testing*. FSI Genetics (2014)

Exclusion power in R

```
> 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

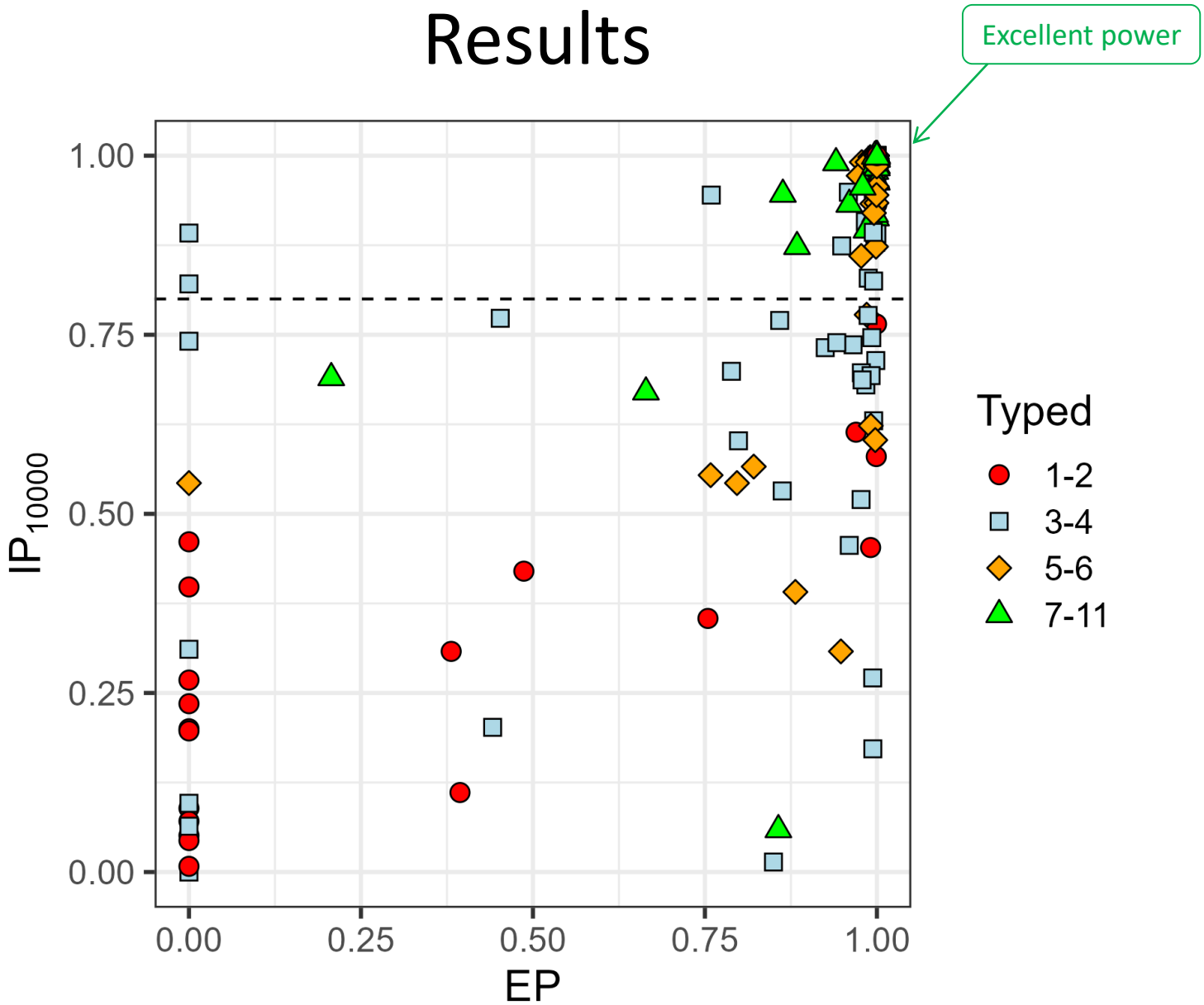

Good power
Inclusion power
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Exclusion power

Back to Argentina ...

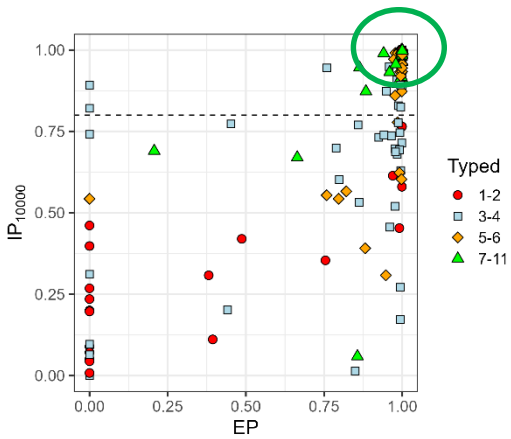
Power assessment of the BNDG database

- Selection of ~200 reference families
- For each family: Compute IP_{10000} and EP

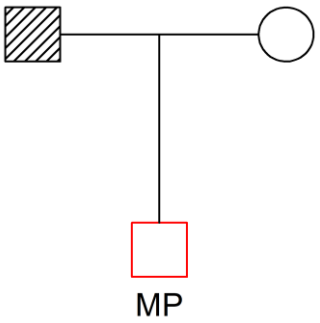
Results

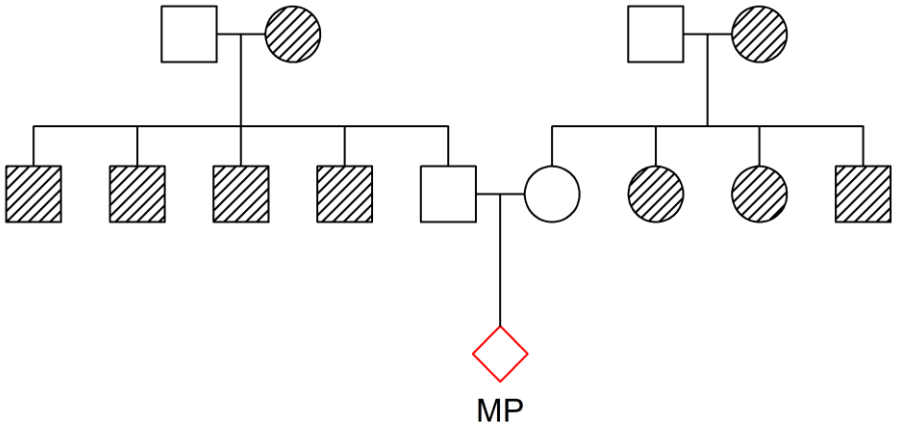

Excellent power
1.00
0.75
0.50
0.25
0.00
IP_{10000}
0.00
0.25
0.50
0.75
1.00
EP
Typed
1-2
3-4
5-6
7-11

Excellent power



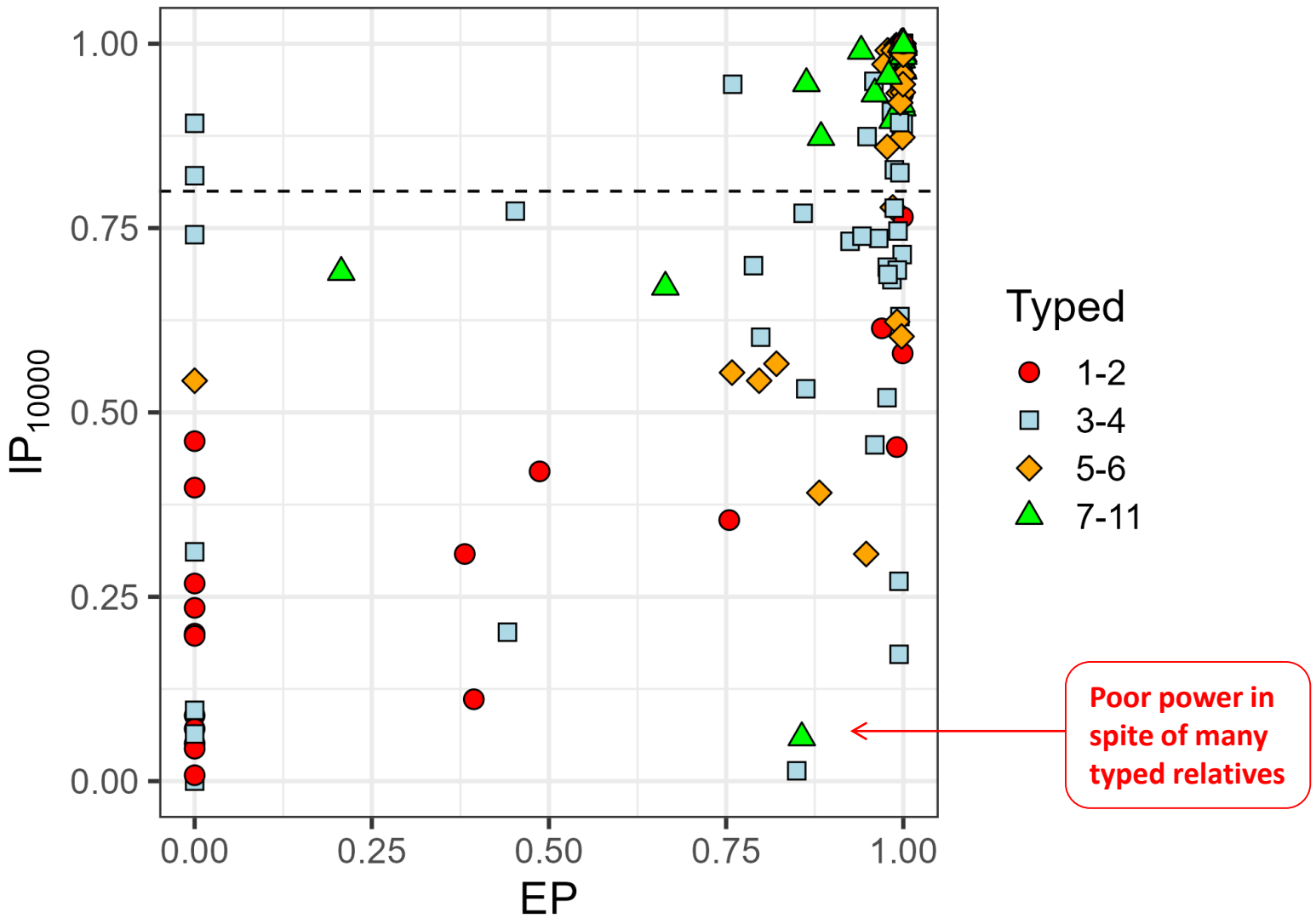
- 68 families with both EP > 99% and IP_{10000} > 99%
- Included ~30 cases with parental data



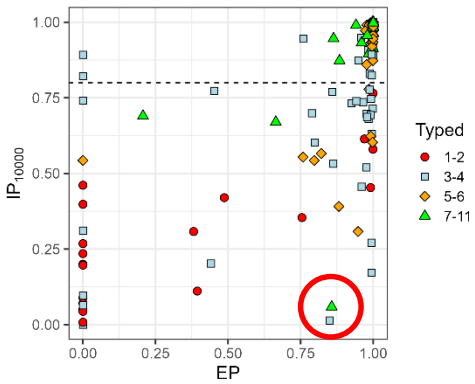


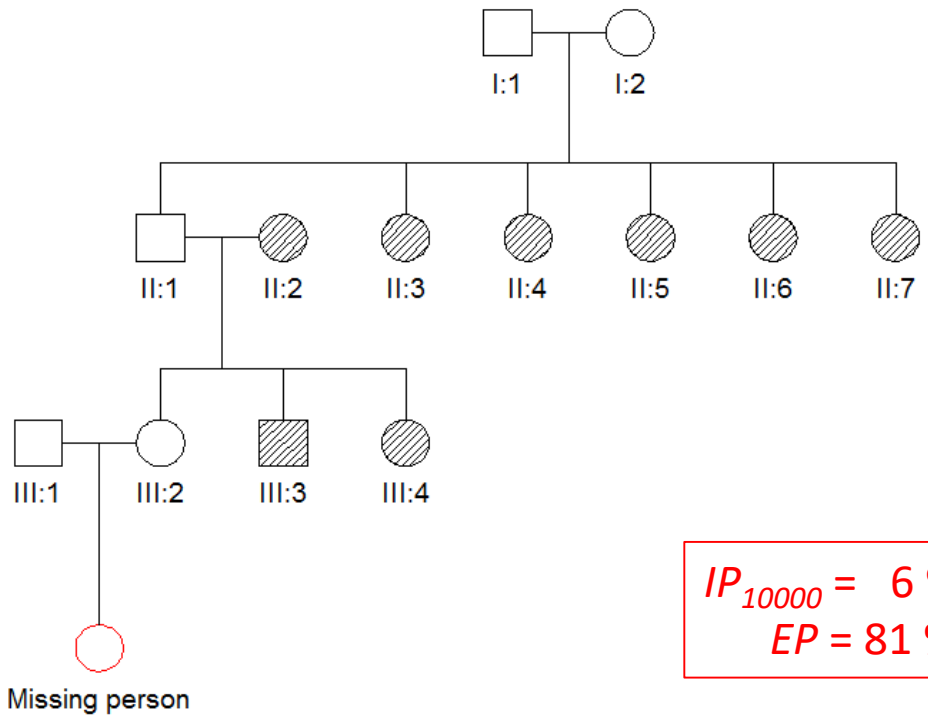
IP_{10000} = 100 %
EP = 100 %

Results


1.00
0.75
0.50
0.25
0.00
IP_{10000}
0.00
0.25
0.50
0.75
1.00
EP
Typed
1-2
3-4
5-6
7-11
Poor power in spite of many typed relatives

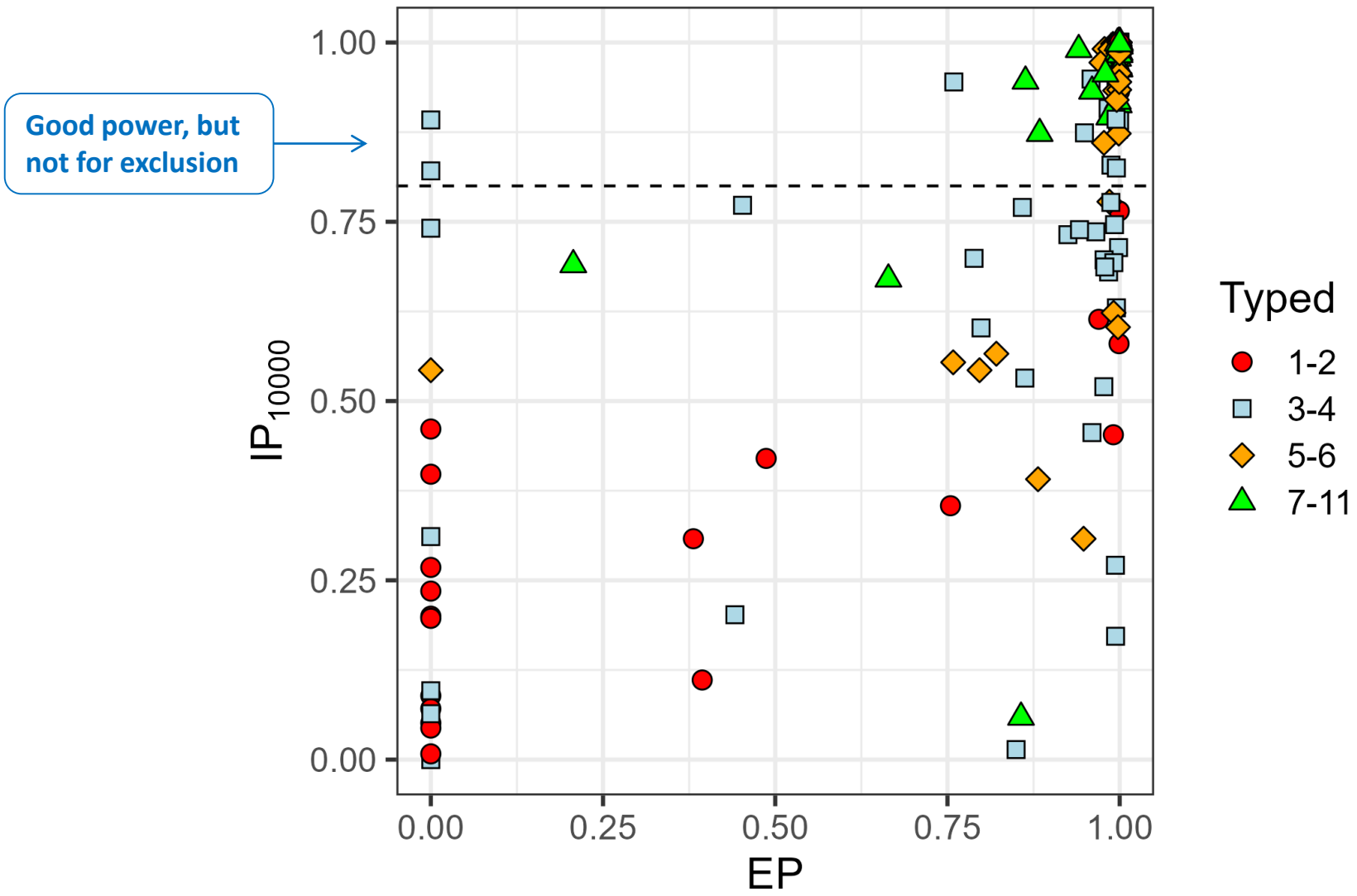
Low power despite many typed



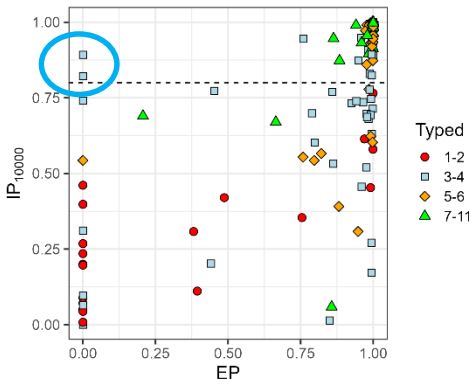


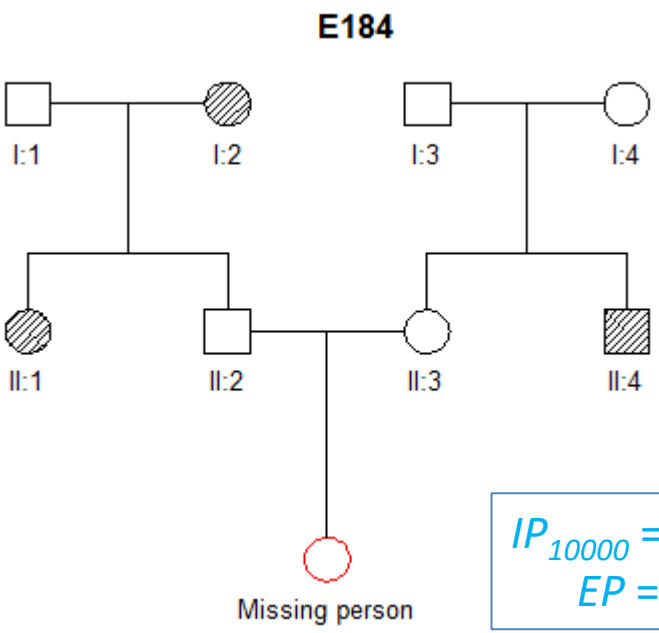
IP_{10000} = 6 %
EP = 81 %

Results

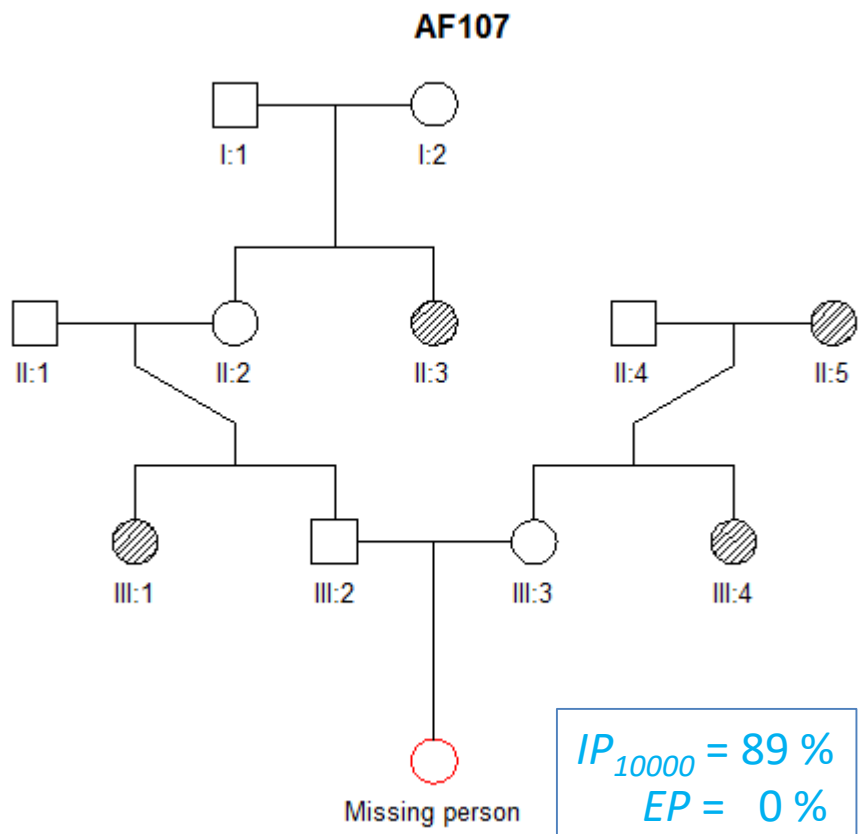

Good power, but not for exclusion
1.00
0.75
0.50
0.25
0.00
IP_{10000}
0.00
0.25
0.50
0.75
1.00
EP
Typed
1-2
3-4
5-6
7-11

Good power, but exclusion impossible



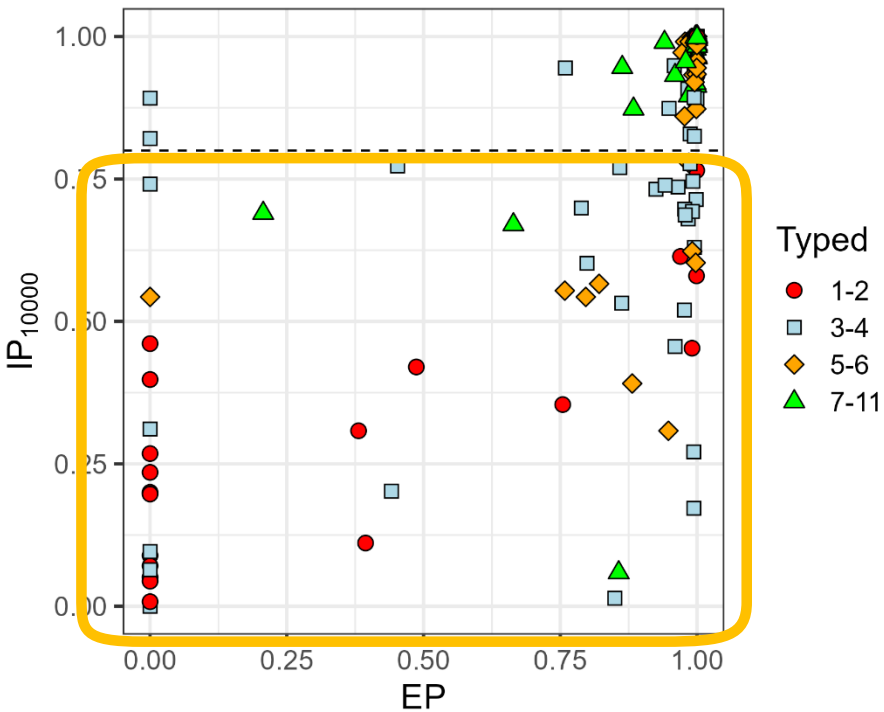


IP_{10000} = 82 %
EP = 0 %



IP_{10000} = 89 %
EP = 0 %

Overall

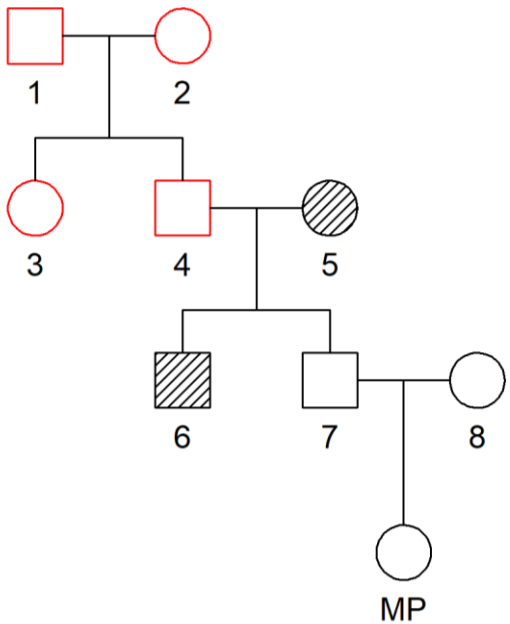


- 34% of the unsolved families had poor power
- Reasons:
  - few markers
  - few typed relatives

Ongoing actions:
- retyping 1000 individuals
- exhumation of informative relatives

Next week

Lecture 2: Who should be exhumed?



Prioritization problems in missing person cases