# Bioinformatics and Statistical Genetics

## Elective specialization for MDS/MIRI/MAI students

**Marta Castellano**

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya Barcelona,
Spain

marta.castellano@upc.edu

UPC
UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

# Syllabus

Bioinformatics and **Statistical Genetics**

1. Introduction to statistical genetics
2. Hardy-Weinberg equilibrium
3. Linkage disequilibrium and haplotype estimation
4. Population substructure
5. Genetic association analysis
6. Relatedness analysis (allele sharing) 12 December 2023

General review 19 December 2023

EXAM 15 January 2024 (Monday) - bring your calculator!
15:00 - 18:00
A5-102

# Population substructure

## RECAP

Set of methods that allow to detect population substructures, which refers to the existence of groups of individuals in a genetic database that come from **different human populations**.

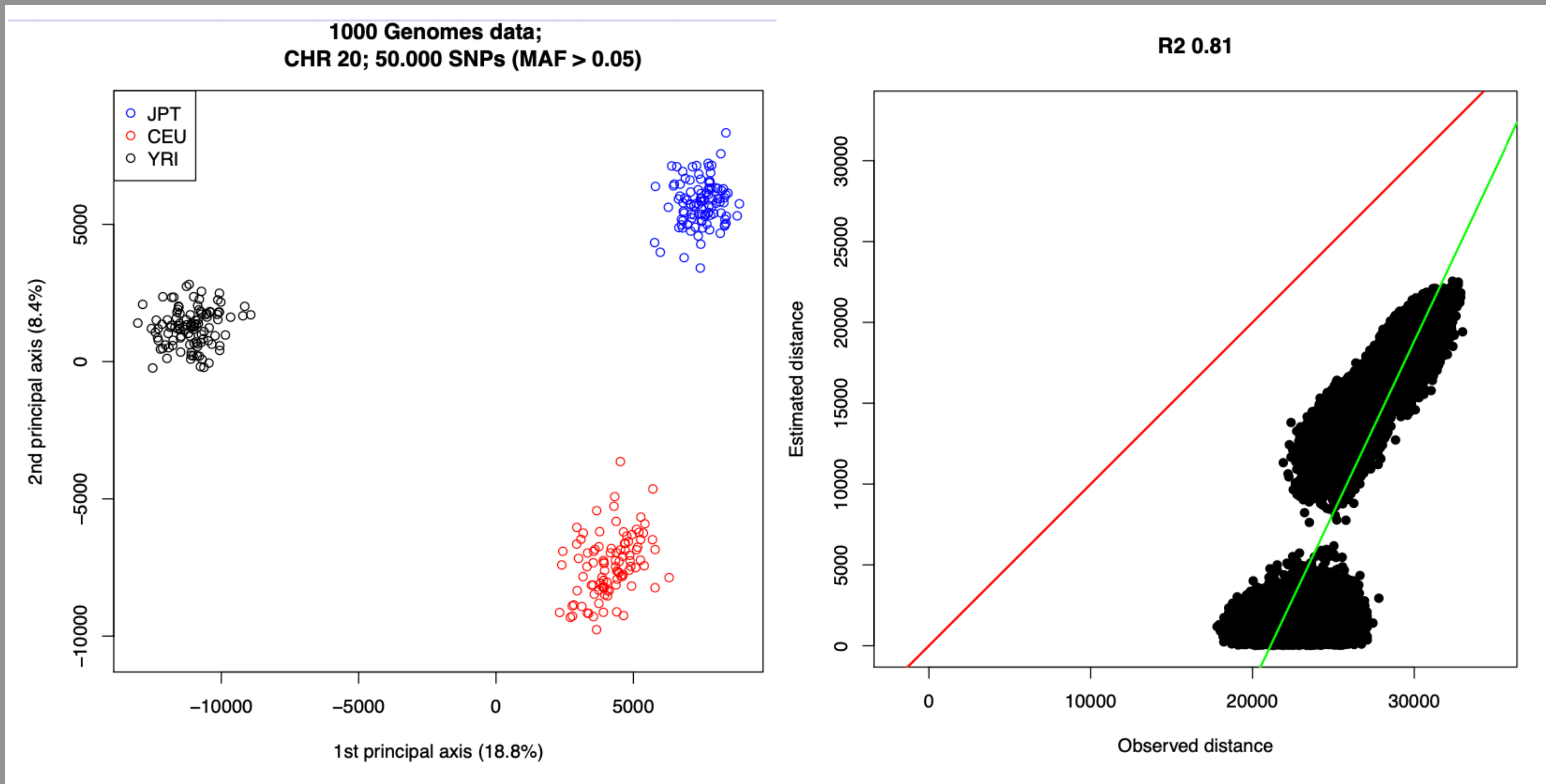
Population structures may influence HWE, LD and marker-trait association studies.

Many methods will assess diverse aspects of population structures. We focused on MDS as a method to estimate populations from genotype data.

- Metric MDS
- Non-metric MDS

Introduced the concepts of ….

- Allele sharing distance
- GOF



# Content

Relatedness analysis

1. Introduction
2. IBS methods
3. IBD methods
4. Computer exercise

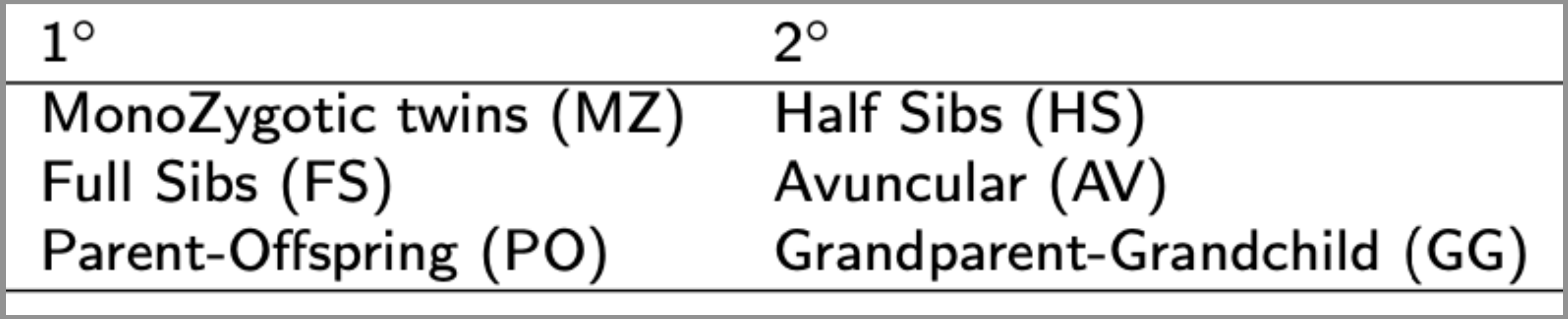
# Relatedness analysis

## Introduction

A distinction is generally drawn between:

- **Close or recent relatedness:** family relationships (MZ, PO, FS, HS, AV, FC, …)
- **Distant or remote relatedness:** population substructure (non-homogeneous genetic data)

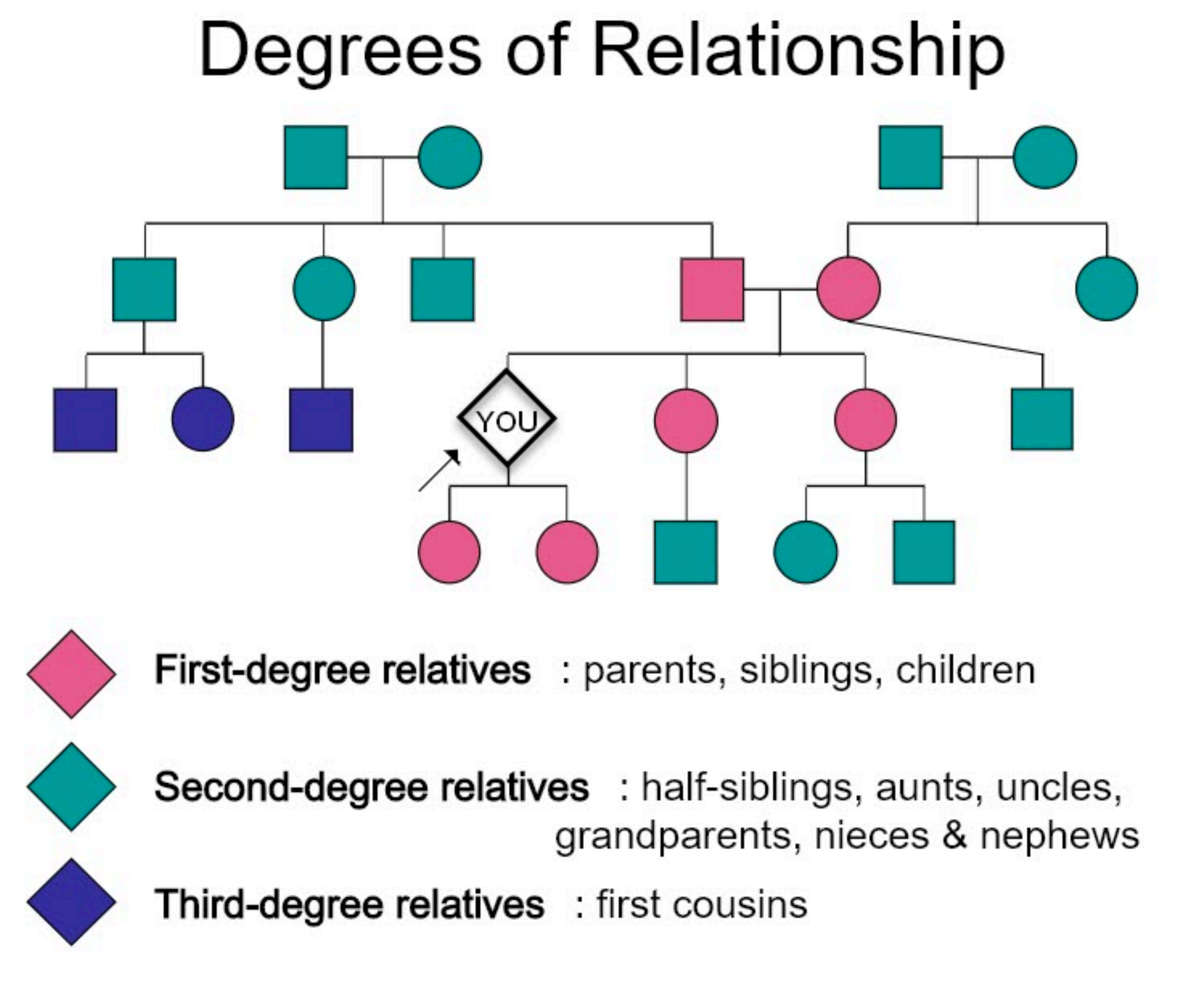
Here we mostly address **recent relatedness.** The focus is on 1°and 2° relationships:



| 1° | 2° |
|---|---|
| MonoZygotic twins (MZ) | Half Sibs (HS) |
| Full Sibs (FS) | Avuncular (AV) |
| Parent-Offspring (PO) | Grandparent-Grandchild (GG) |

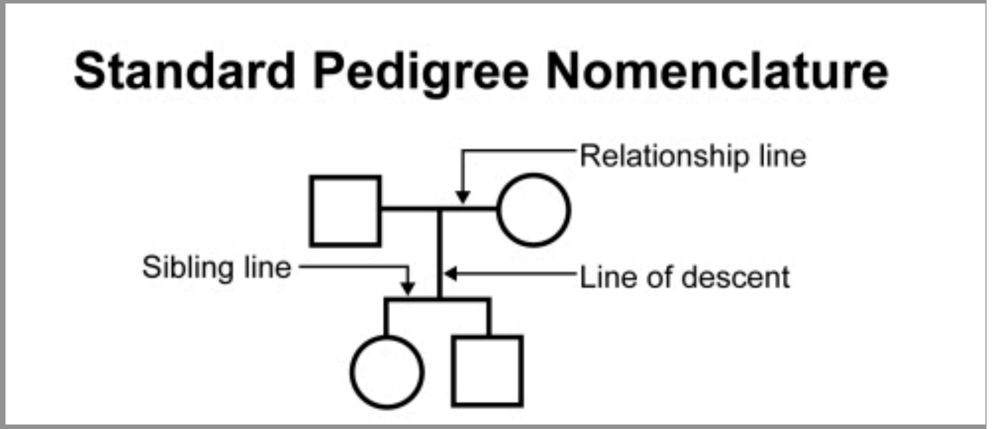
# Relatedness analysis

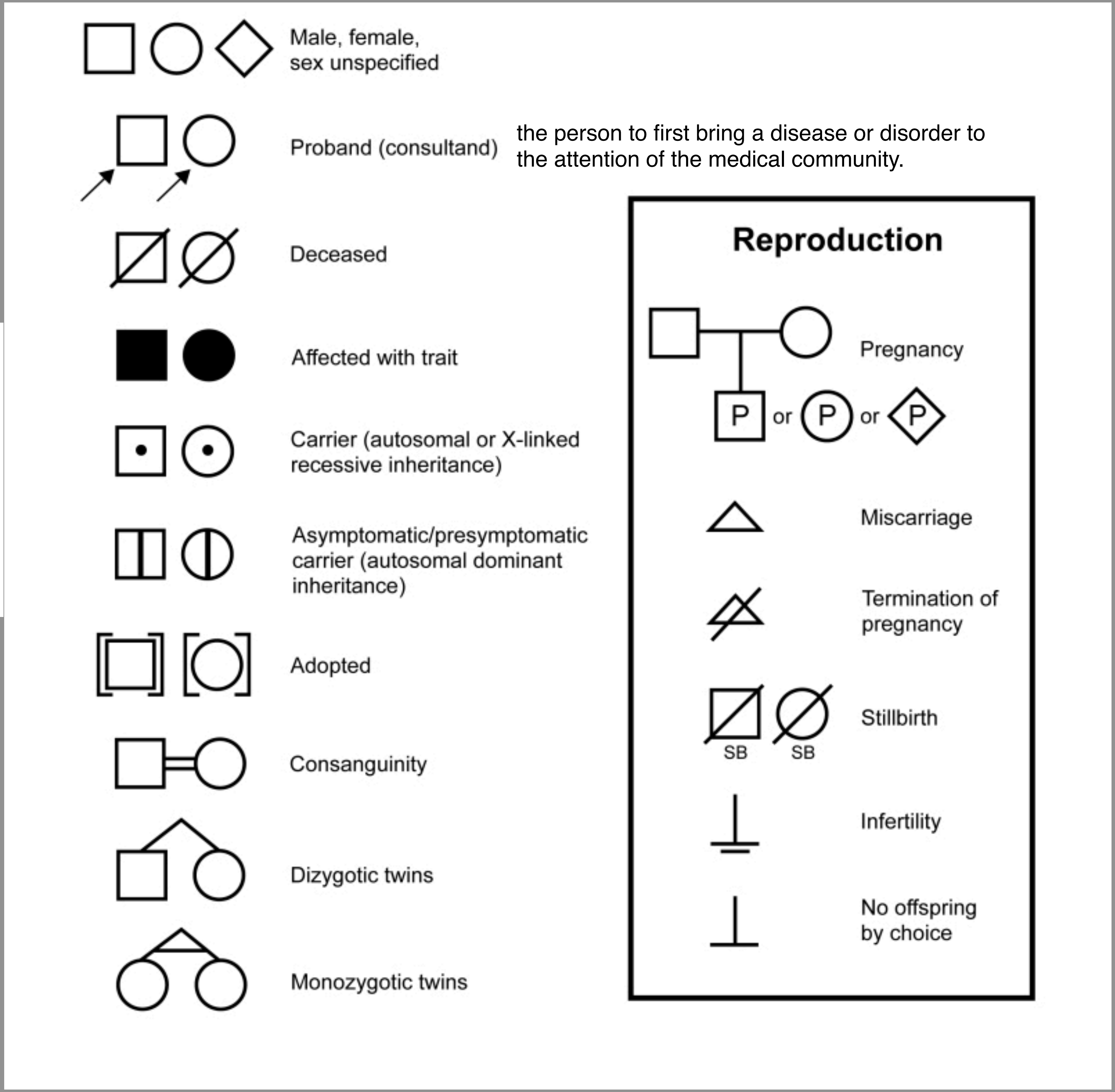
## Introduction - pedigree


Degrees of Relationship
YOU
First-degree relatives : parents, siblings, children
Second-degree relatives : half-siblings, aunts, uncles, grandparents, nieces & nephews
Third-degree relatives : first cousins

# Relatedness analysis

## Introduction - pedigree nomenclature





the person to first bring a disease or disorder to the attention of the medical community.

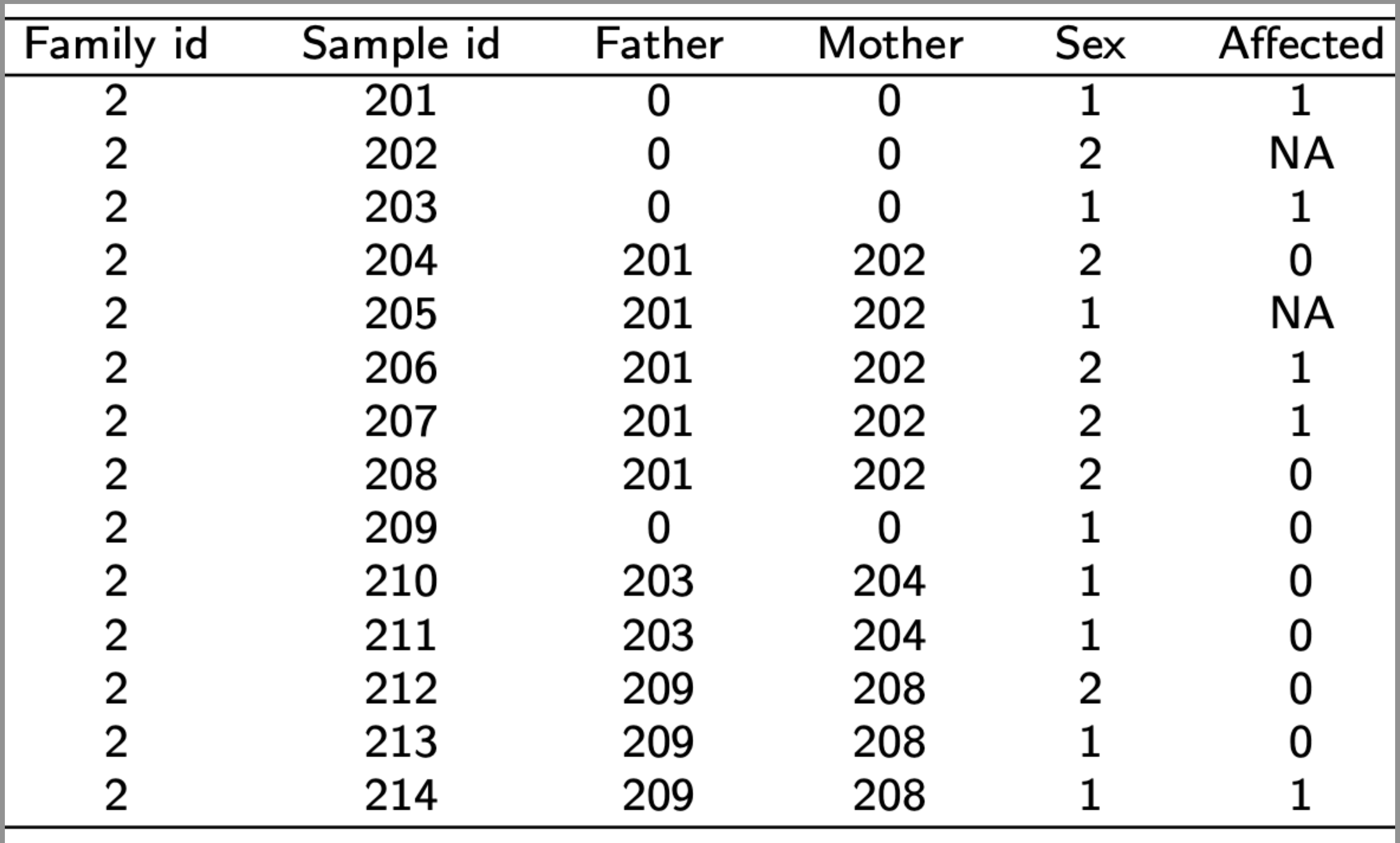
# Relatedness analysis

## Introduction - coding pedigree in the data

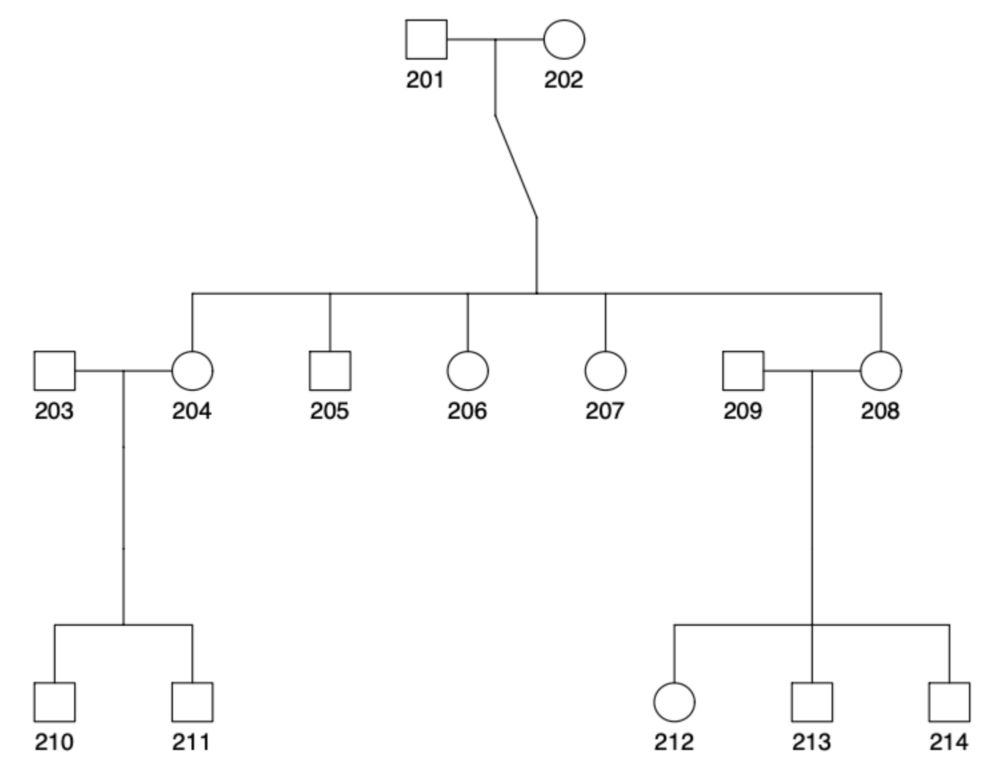
- A database of related individuals is typically coded in .ped file format. Pedigree (.ped) files describe relationships between individuals in your dataset and also store marker genotypes, disease status and quantitative trait values.

Example:

Family ID, Sample ID, Paternal ID, Maternal ID, Sex (1=male; 2=female; other=unknown) and Affection status (1=affected; 0=unaffected) are registered.



| Family id | Sample id | Father | Mother | Sex | Affected |
|---|---|---|---|---|---|
| 2 | 201 | 0 | 0 | 1 | 1 |
| 2 | 202 | 0 | 0 | 2 | NA |
| 2 | 203 | 0 | 0 | 1 | 1 |
| 2 | 204 | 201 | 202 | 2 | 0 |
| 2 | 205 | 201 | 202 | 1 | NA |
| 2 | 206 | 201 | 202 | 2 | 1 |
| 2 | 207 | 201 | 202 | 2 | 1 |
| 2 | 208 | 201 | 202 | 2 | 0 |
| 2 | 209 | 0 | 0 | 1 | 0 |
| 2 | 210 | 203 | 204 | 1 | 0 |
| 2 | 211 | 203 | 204 | 1 | 0 |
| 2 | 212 | 209 | 208 | 2 | 0 |
| 2 | 213 | 209 | 208 | 1 | 0 |
| 2 | 214 | 209 | 208 | 1 | 1 |



# Relatedness analysis

## Introduction - motivation

Goal:
The detection of (closely) related individuals in genetic studies

Motivations:

- In association studies, many methods assume independent individuals. Closely related individuals will not be independent.
- In conservation genetics, breeding programs are set up for preferably unrelated individuals.
- In quality control of the data, samples can be accidentally duplicated, and it is of interest to detect it.
- In paternity testing.
- In forensic genetics, e.g. identification of remains.
- To verify documented family relationships.
- To uncover cryptic relatedness.
- ...

# Relatedness analysis

## Introduction - terminology

Goal:
The detection of (closely) related individuals in genetic studies

- **Genetic relatedness:** historically calculated through pedigree information, an estimation on how similar two individuals are.

- **Kinship coefficient:** probability that an allele at a locus for two individuals was inherited from the same genome. Degree of consanguinity between two individuals.

  - If you were to randomly draw one of the two alleles from a locus of an individual, and randomly draw one of the two alleles from first cousin (same locus), what is the probability that those alleles both come from a common grandparent?

- **Inbreeding coefficient:** probability that the two alleles at a locus for an individual was inherited from the same genome.

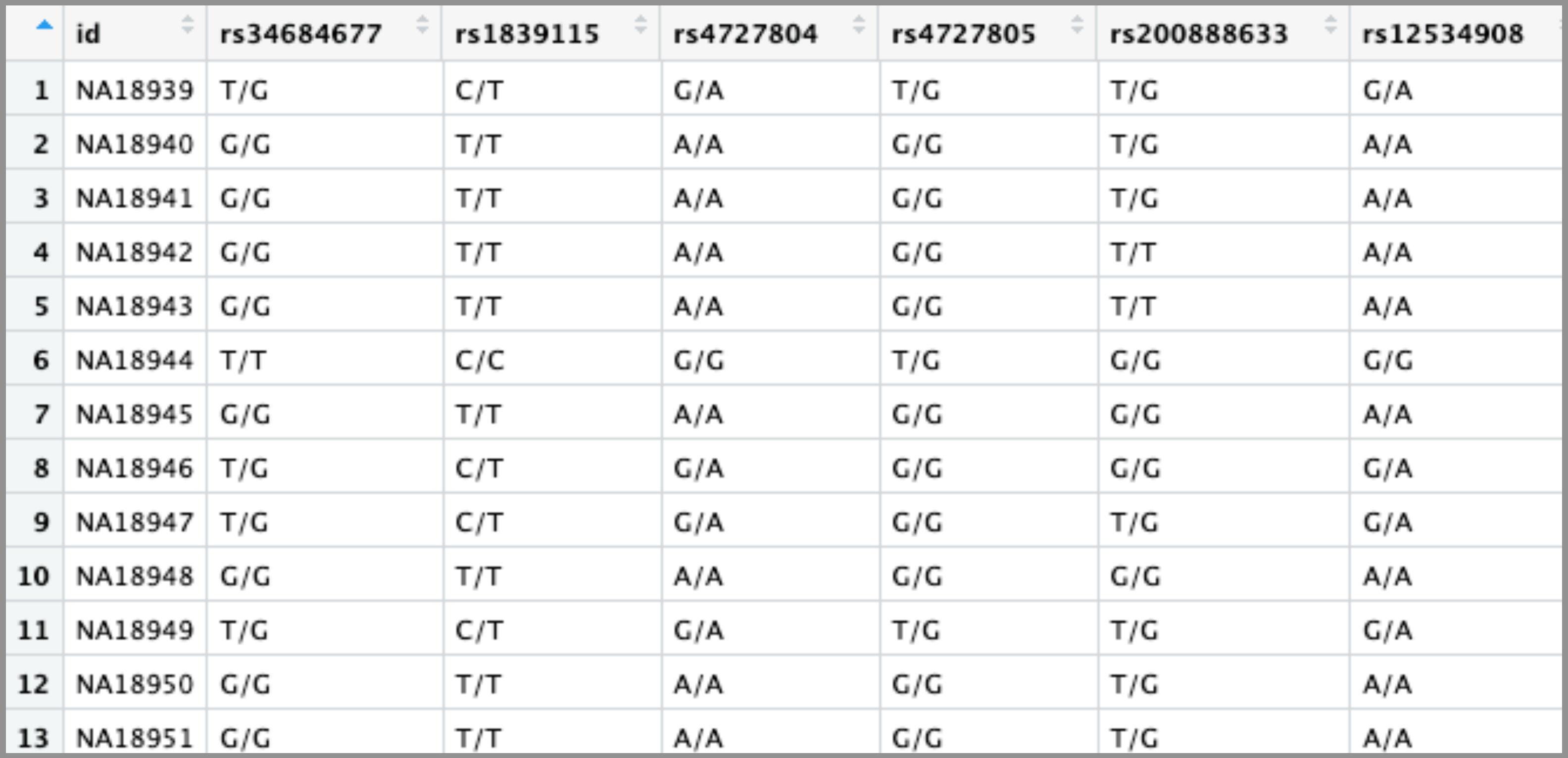
  - If the parents had a recent common ancestor, the inbreeding coefficient would be higher for that individual.

# Relatedness analysis

## Introduction - terminology

Genotype data can be used to estimate genetic relatedness:

- **Identity by state (IBS):**
  - Same allele at a locus between two individuals
- **Identity by descent (IBD):**
  - Same allele at a locus between two individuals
  - Inherited from a recent common ancestor



| | id | rs34684677 | rs1839115 | rs4727804 | rs4727805 | rs200888633 | rs12534908 |
|---|---|---|---|---|---|---|---|
| 1 | NA18939 | T/G | C/T | G/A | T/G | T/G | G/A |
| 2 | NA18940 | G/G | T/T | A/A | G/G | T/G | A/A |
| 3 | NA18941 | G/G | T/T | A/A | G/G | T/G | A/A |
| 4 | NA18942 | G/G | T/T | A/A | G/G | T/T | A/A |
| 5 | NA18943 | G/G | T/T | A/A | G/G | T/T | A/A |
| 6 | NA18944 | T/T | C/C | G/G | T/G | G/G | G/G |
| 7 | NA18945 | G/G | T/T | A/A | G/G | G/G | A/A |
| 8 | NA18946 | T/G | C/T | G/A | G/G | G/G | G/A |
| 9 | NA18947 | T/G | C/T | G/A | G/G | T/G | G/A |
| 10 | NA18948 | G/G | T/T | A/A | G/G | G/G | A/A |
| 11 | NA18949 | T/G | C/T | G/A | T/G | T/G | G/A |
| 12 | NA18950 | G/G | T/T | A/A | G/G | T/G | A/A |
| 13 | NA18951 | G/G | T/T | A/A | G/G | T/G | A/A |

CHALLENGE: to move from observed genotype to inference about relatedness.

# Relatedness analysis

## Introduction - terminology

Genotype data can be used to estimate genetic relatedness:

- **Identity by state (IBS):**
  - Same allele at a locus between two individuals

Alleles are IBS if…

- Those alleles come from a common ancestor
- There is an entirely independent mutation that arose in the population

- **Identity by descent (IBD):**
  - Same allele at a locus between two individuals
  - Inherited from a recent common ancestor

RECALL human genetic variability:

- There is about 3-4 million SNPs (human genome is about 3 Gbp long and there is about one SNP per every 1000 bp). Allele frequencies differ by 15% on average across populations.

- High MAF variants are more informative for discriminating relationship categories.
  - At the population level, we have greater genetic variation at that locus, increased chance of heterozygosity…they are more informative than low MAF variants.
  - If a variant is rare, it may help us identify a recent common ancestor, but it's less likely shared across a broader range of relatives.

- Witherspoon, D.J., Wooding, S., Rogers, A.R., Marchani, E.E., Watkins, W.S., Batzer, M.A. and Jorde, L.B., 2007. Genetic similarities within and between human populations. Genetics, 176(1), pp.351-359.
- Miller, R.D., Phillips, M.S., Jo, I., Donaldson, M.A., Studebaker, J.F., Addleman, N., Alfisi, S.V., Ankener, W.M., Bhatti, H.A., Callahan, C.E. and Carey, B.J., 2005. High-density single-nucleotide polymorphism maps of the human genome. Genomics, 86(2), pp.117-126.

# Relatedness analysis

## Introduction - terminology

Genotype data can be used to estimate genetic relatedness:

- **Identity by state (IBS):**
  - Same allele at a locus between two individuals
- **Identity by descent (IBD):**
  - Same allele at a locus between two individuals
  - Inherited from a recent common ancestor

Alleles are IBS if…

- Those alleles come from a common ancestor
- There is an entirely independent mutation that arose in the population

RECALL human genetic variability:

- There is about 3-4 million SNPs (human genome is about 3 Gbp long and there is about one SNP per every 1000 bp). Allele frequencies differ by 15% on average across populations.

- High MAF variants are more informative for discriminating relationship categories.
- Variants with multiple alleles (e.g. microsatellites) are more informative for discriminating relationship categories than bi-allelic variants (SNP data).
- Consider enriching your dataset for those loci that are already likely to vary between individuals (HighD sites).

- Witherspoon, D.J., Wooding, S., Rogers, A.R., Marchani, E.E., Watkins, W.S., Batzer, M.A. and Jorde, L.B., 2007. Genetic similarities within and between human populations. Genetics, 176(1), pp.351-359.
- Miller, R.D., Phillips, M.S., Jo, I., Donaldson, M.A., Studebaker, J.F., Addleman, N., Alfisi, S.V., Ankener, W.M., Bhatti, H.A., Callahan, C.E. and Carey, B.J., 2005. High-density single-nucleotide polymorphism maps of the human genome. Genomics, 86(2), pp.117-126.

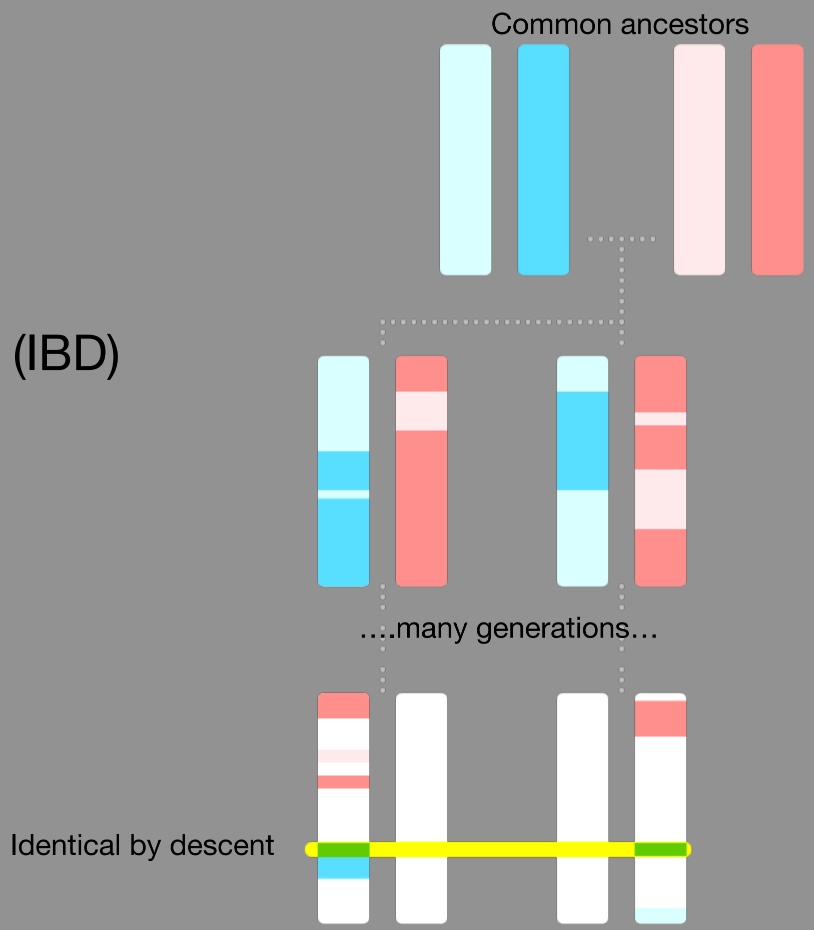
# Relatedness analysis

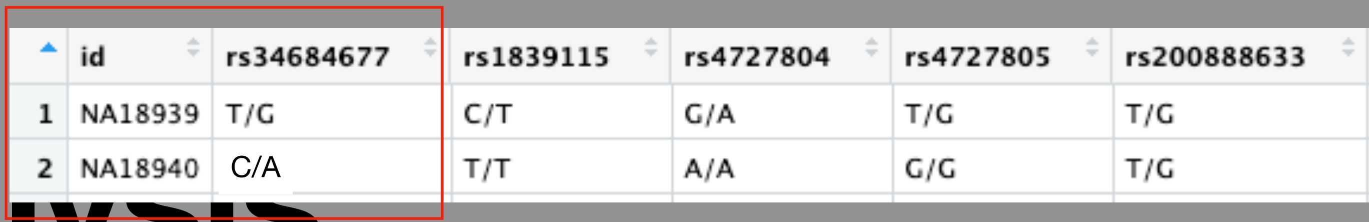
## Introduction - terminology

Genotype data can be used to estimate genetic relatedness:

- **Identity by state (IBS):**
    - Same allele at a locus between two individuals
- **Identity by descent (IBD):**
    - Same allele at a locus between two individuals
    - Inherited from a recent common ancestor

- A pair of alleles can be identical by state (IBS) or identical by descent (IBD)
    - IBS alleles simply match irrespective of their provenance
    - IBD alleles match because of a common ancestor.
    - IBD implies IBS but not the reverse.





| | id | rs34684677 | rs1839115 | rs4727804 | rs4727805 | rs200888633 |
|---|---|---|---|---|---|---|
| 1 | NA18939 | T/G | C/T | G/A | T/G | T/G |
| 2 | NA18940 | C/A | T/T | A/A | G/G | T/G |

# Relatedness analysis

## Introduction - terminology

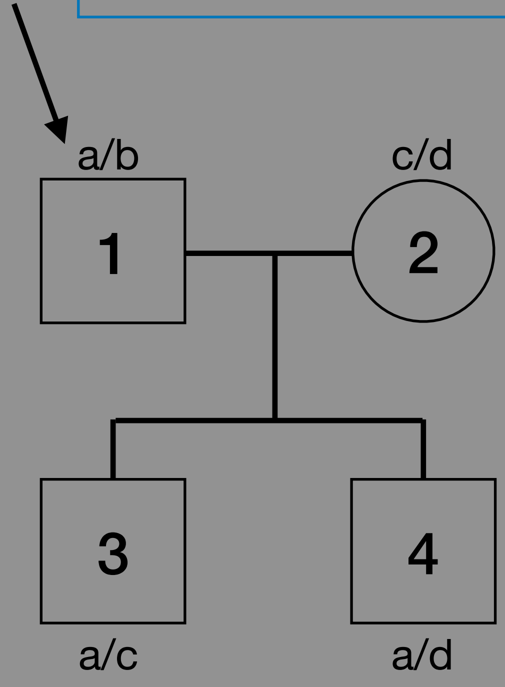
Genotype data can be used to estimate genetic relatedness:

- **Identity by state (IBS):**
  - Same allele at a locus between two individuals
- **Identity by descent (IBD):**
  - Same allele at a locus between two individuals
  - Inherited from a recent common ancestor

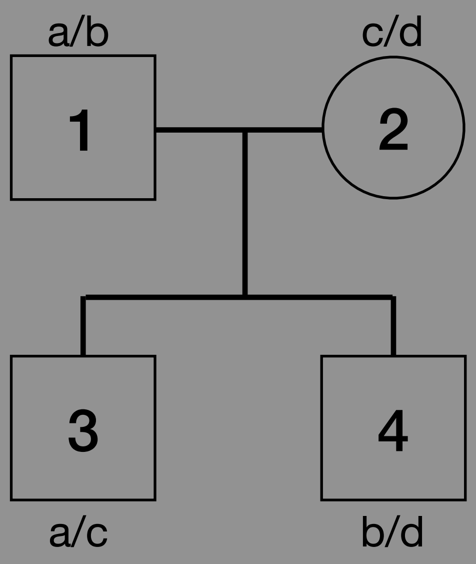
Punnett square to reflect the 4 possible genotypes in the next generation:

| | **a** | **b** |
|---|---|---|
| **c** | ac | bc |
| **d** | ad | bd |

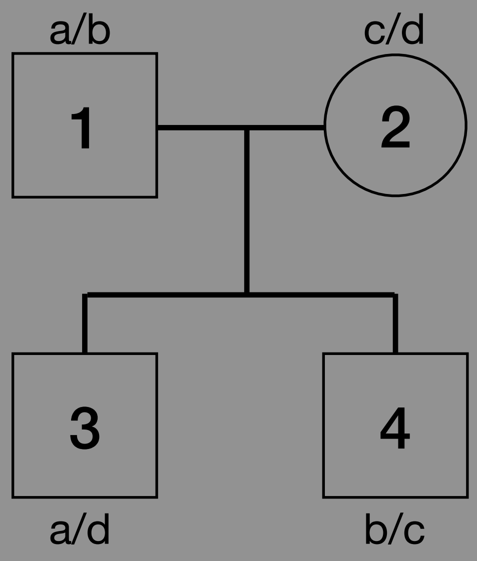
One locus with different alleles



- Allele a is IBD between individuals 3 and 4.



- No IBD alleles between individuals 3 and 4



- No IBD alleles between individuals 3 and 4

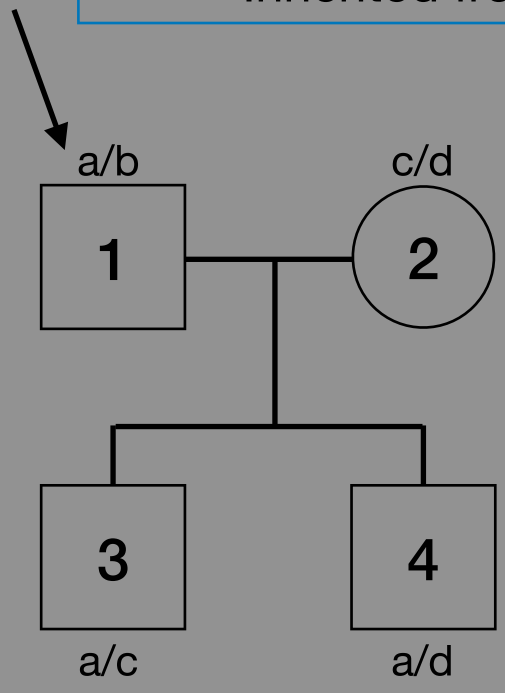
# Relatedness analysis

## Introduction - terminology

Genotype data can be used to estimate genetic relatedness:

- **Identity by state (IBS):**
  - Same allele at a locus between two individuals
- **Identity by descent (IBD):**
  - Same allele at a locus between two individuals
  - Inherited from a recent common ancestor

One locus with different alleles



- Allele a is IBD between individuals 3 and 4.

Punnett square to reflect the 4 possible genotypes in the next generation:

| | **a** | **b** |
|---|---|---|
| **c** | ac | bc |
| **d** | ad | bd |

Table to reflect the probability of IBD for full siblings. Fill in number of same alleles:

| | **ac** | **cb** | **ad** | **db** |
|---|---|---|---|---|
| **ac** | 2 | 1 | 1 | 0 |
| **cb** | 1 | 2 | 0 | 1 |
| **ad** | 1 | 0 | 2 | 1 |
| **db** | 0 | 1 | 1 | 2 |

1/4 genotypes/individuals have 2 alleles in IBD
1/2 genotypes/individuals share one allele in IBD
1/4 genotypes/individuals are not IBD

# Relatedness analysis

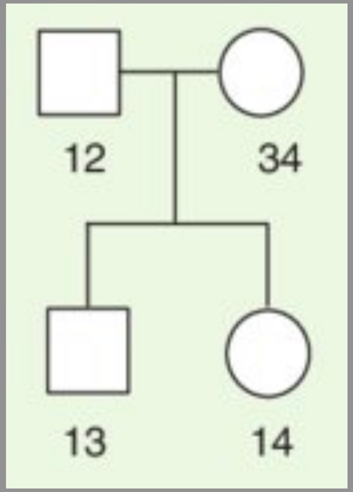
## Introduction - terminology

Genotype data can be used to estimate genetic relatedness:

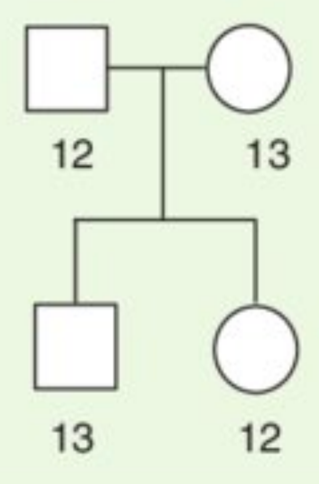
- **Identity by state (IBS):**
  - Same allele at a locus between two individuals
- **Identity by descent (IBD):**
  - Same allele at a locus between two individuals
  - Inherited from a recent common ancestor

**Example of IBD with uncertain info:**

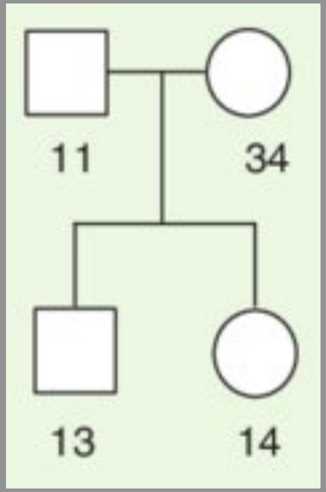
Parent-offspring share exactly 1 allele IBD



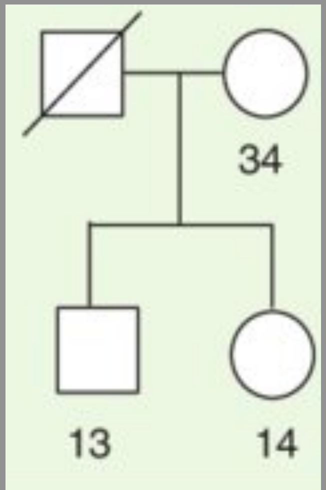
allele 1 IBS
(assuming parents unrelated)



Not possible to tell whether allele 1 in the two sibs is IBD or just IBS because the father is homozygous for allele 1 (and thus not informative)



The probability of allele 1 being IBD can be calculated based on its population frequency, and thus the probability of the father carrying 1 or 2 copies of the same allele.



Forabosco, P., Falchi, M. and Devoto, M., 2005. Statistical tools for linkage analysis and genetic association studies. Expert Review of Molecular Diagnostics, 5(5), pp.781-796.

# Relatedness analysis

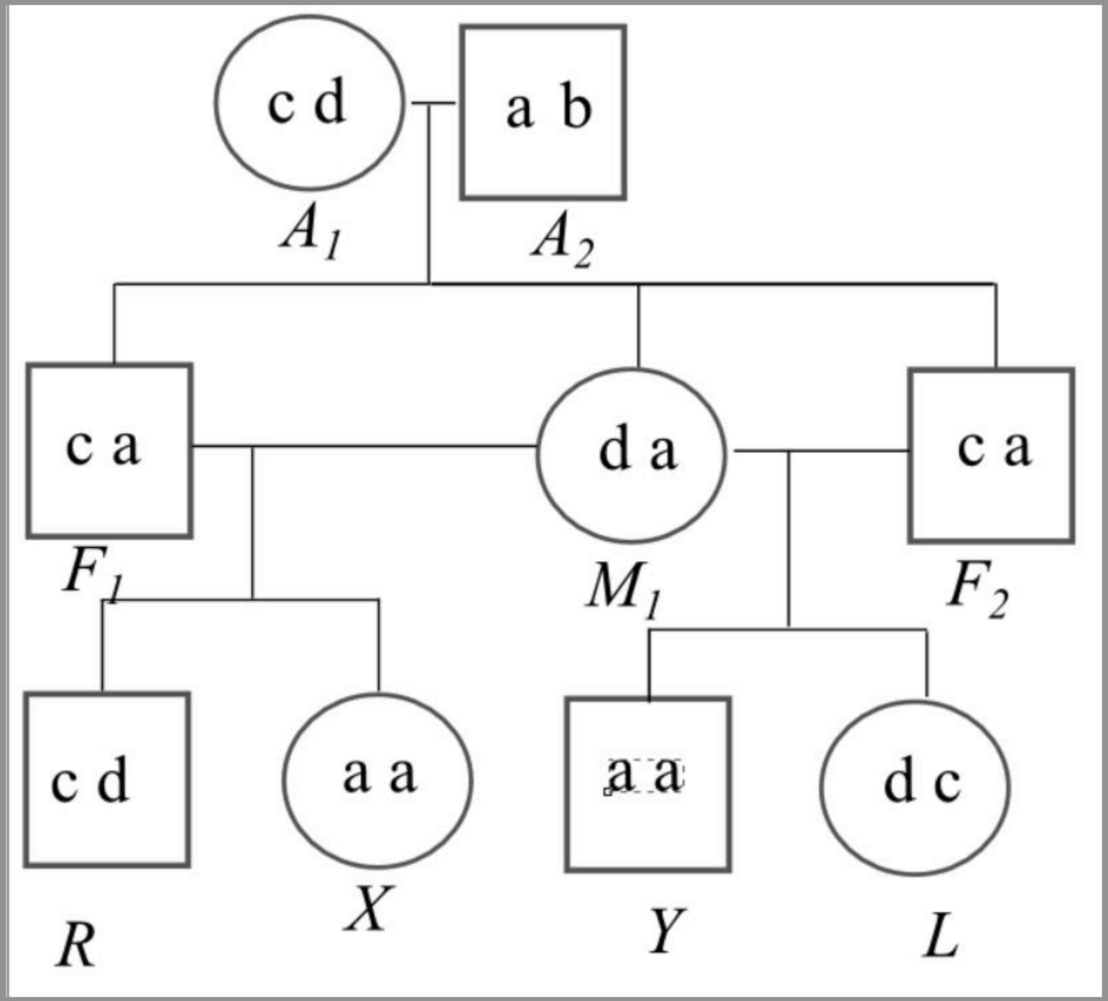
## Introduction - terminology

Genotype data can be used to estimate genetic relatedness:

- **Identity by state (IBS):**
  - Same allele at a locus between two individuals
- **Identity by descent (IBD):**
  - Same allele at a locus between two individuals
  - Inherited from a recent common ancestor

**Example of IBD with inbreeding:**
How many alleles are IBD for R, X and Y individuals?

- R,X individual has 0 allele IBD
- R,Y individual has 0 allele IBD
- X,Y individual has 2 alleles IBD



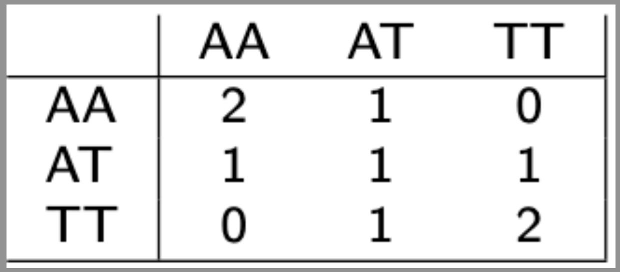
# Relatedness analysis

## Introduction - allele sharing

Much of relatedness research is based on the principle of **allele sharing**

- For diploid individuals, a pair of individuals can share 0, 1 or 2 alleles for a certain locus.
- The degree to which individuals share alleles indicates the extent to which they are related.

- For any locus, we can record for a pair of individuals how many alleles are IBS (how many alleles ”match”) and this can be 0, 1 or 2.
- Example: for an A/T single nucleotide polymorphism (SNP):



| | AA | AT | TT |
|---|---|---|---|
| AA | 2 | 1 | 0 |
| AT | 1 | 1 | 1 |
| TT | 0 | 1 | 2 |

- The **number of IBS alleles** can be recorded for many loci, and averaged over loci.

- This principle can be used to uncover closely related individuals, or to detect sample heterogeneity (individuals from different populations).

# Relatedness analysis

## Introduction - allele sharing

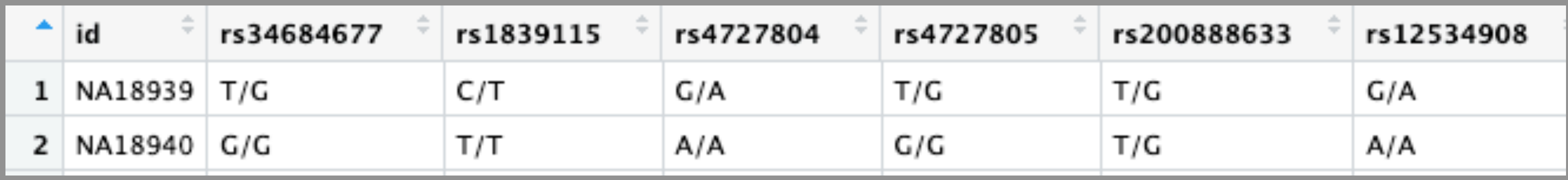
Much of relatedness research is based on the principle of **allele sharing**

- For diploid individuals, a pair of individuals can share 0, 1 or 2 alleles for a certain locus.
- The degree to which individuals share alleles indicates the extent to which they are related.

RECALL:

- The **allele sharing distance** is an often used measure to estimate genetic distance between individuals. Used as input for MDS.
- Let $x_{ijk}$ be the number of shared alleles of individual i and j for variant k
- Define $d_{ijk} = 2 - x_{ijk}$
- Typically averaged over K genetic variants:

$$d_{ij} = \frac{1}{K}\sum_{k=1}^{K} d_{ijk}$$



|  | id | rs34684677 | rs1839115 | rs4727804 | rs4727805 | rs200888633 | rs12534908 |
|---|---|---|---|---|---|---|---|
| 1 | NA18939 | T/G | C/T | G/A | T/G | T/G | G/A |
| 2 | NA18940 | G/G | T/T | A/A | G/G | T/G | A/A |

# Content

Relatedness analysis (allele sharing)

1. Introduction
2. IBS methods
3. IBD methods
4. Computer exercise

# Relatedness analysis

## Introduction - IBS methods

Much of relatedness research is based on the principle of **allele sharing**

- For diploid individuals, a pair of individuals can share 0, 1 or 2 alleles for a certain locus.
- The degree to which individuals share alleles indicates the extent to which they are related.

Allele sharing statistics are often graphed in one of the following ways:

- By plotting means ($m$) and standard deviation ($s$) of IBS statistics: $(m, s)$ plot
- By plotting percentages of markers with 0, 1 or 2 IBS alleles: $(p_0, p_2)$ plot

# Relatedness analysis

## Introduction - IBS methods - $(m, s)$ plot

GOAL: plot the means ($m$) and standard deviation ($s$) of IBS statistics by means of the $(m, s)$ plot

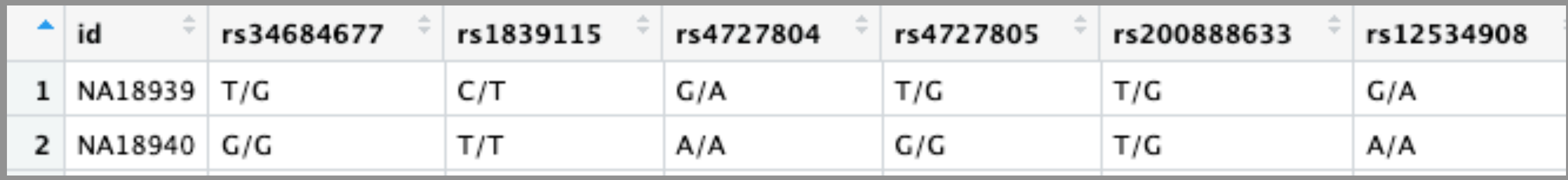
- Let $x_{ijk}$ be the number of shared alleles of individual i and j for variant k

- Compute $m_{ij}$ and $s_{ij}$ over K genetic variants:

$$m_{ij} = \frac{1}{K}\sum_{k=1}^{K} x_{ijk} \text{ and}$$

$$s_{ij} = \frac{1}{K-1}\sum_{k=1}^{K} (x_{ijk} - m_{ij})^2$$

- Plot $m_{ij}$ against $s_{ij}$
- Precise position of the different clusters depends on the distribution of the allele frequencies.
- This plot reveals characteristic clusters that correspond to the different family relationships.



| | id | rs34684677 | rs1839115 | rs4727804 | rs4727805 | rs200888633 | rs12534908 |
|---|---|---|---|---|---|---|---|
| 1 | NA18939 | T/G | C/T | G/A | T/G | T/G | G/A |
| 2 | NA18940 | G/G | T/T | A/A | G/G | T/G | A/A |

# Relatedness analysis

## Introduction - IBS methods - $(m, s)$ plot

GOAL: plot the means ($m$) and standard deviation ($s$) of IBS statistics by means of the $(m, s)$ plot

- Let $x_{ijk}$ be the number of shared alleles of individual i and j for variant k

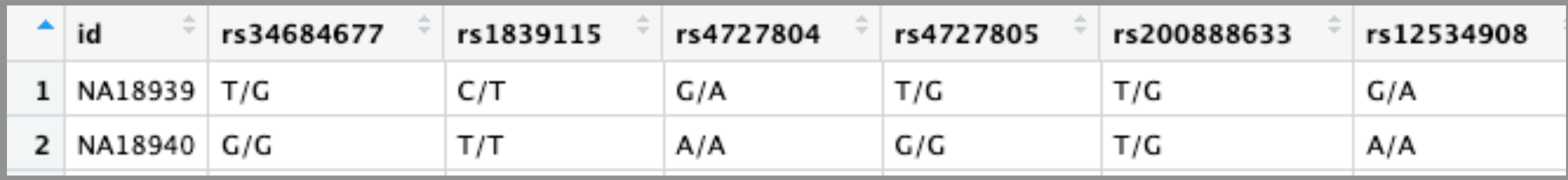
- Compute $m_{ij}$ and $s_{ij}$ over K genetic variants:

$$m_{ij} = \frac{1}{K}\sum_{k=1}^{K} x_{ijk} \text{ and}$$

$$s_{ij} = \frac{1}{K-1}\sum_{k=1}^{K} (x_{ijk} - m_{ij})^2$$

- Plot $m_{ij}$ against $s_{ij}$
- Precise position of the different clusters depends on the distribution of the allele frequencies.
- This plot reveals characteristic clusters that correspond to the different family relationships.

$$m_{12} = \frac{1}{6}\sum_{k=1}^{6} x_{ijk} = \frac{7}{6} = 1.16$$



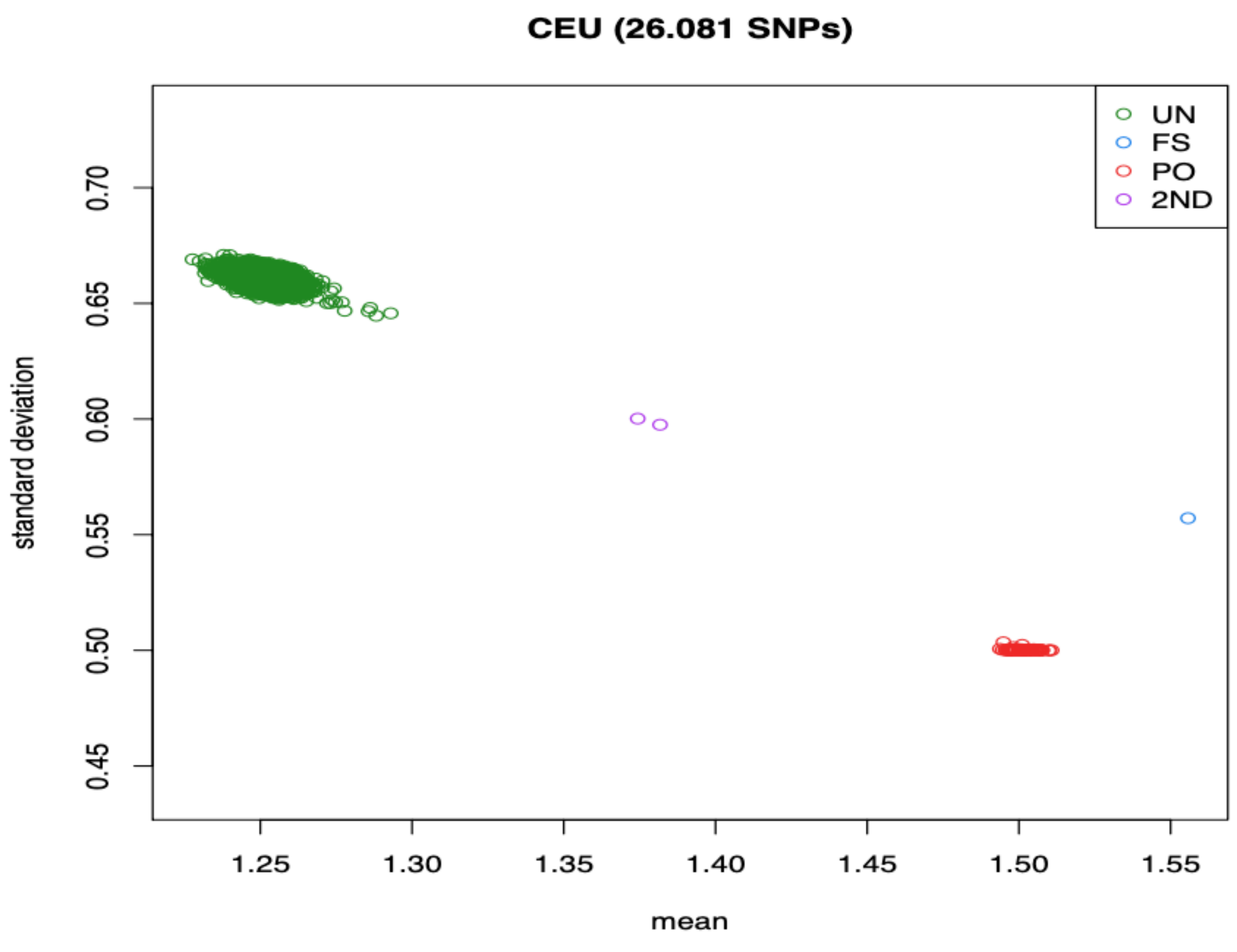
| | id | rs34684677 | rs1839115 | rs4727804 | rs4727805 | rs200888633 | rs12534908 |
|---|---|---|---|---|---|---|---|
| 1 | NA18939 | T/G | C/T | G/A | T/G | T/G | G/A |
| 2 | NA18940 | G/G | T/T | A/A | G/G | T/G | A/A |

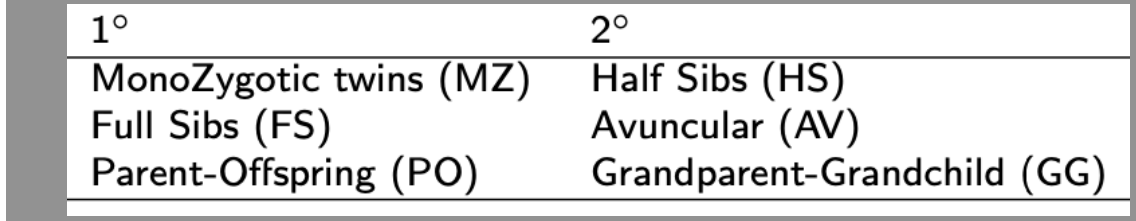
# Relatedness analysis

## Introduction - IBS methods - $(m, s)$ plot

Example: CEU sample from the 1000 Genome project. (n = 165, p = 26.081 pruned highly variable SNPs)

RECALL: LD pruning to thin the markers so that they are approximately independent in the population





| 1° | 2° |
|---|---|
| MonoZygotic twins (MZ) | Half Sibs (HS) |
| Full Sibs (FS) | Avuncular (AV) |
| Parent-Offspring (PO) | Grandparent-Grandchild (GG) |

An average of 2 would mean that the two individuals are identical (monozygotic twins) or that a sample has been accidentally duplicated.

# Relatedness analysis

## Introduction - IBS methods - $(p_0, p_2)$ plot

GOAL: plotting percentages of markers with 0, 1 or 2 IBS alleles by means of the $(p_0, p_2)$ plot

- Let $x_{ijk}$ be the number of shared alleles of individual i and j for variant k
- Compute over K genetic variants:

$$p_0 = \frac{1}{K}\sum_{k=1}^{K} I(x_{ijk} = 0)$$

$$p_1 = \frac{1}{K}\sum_{k=1}^{K} I(x_{ijk} = 1)$$

$$p_2 = \frac{1}{K}\sum_{k=1}^{K} I(x_{ijk} = 2)$$

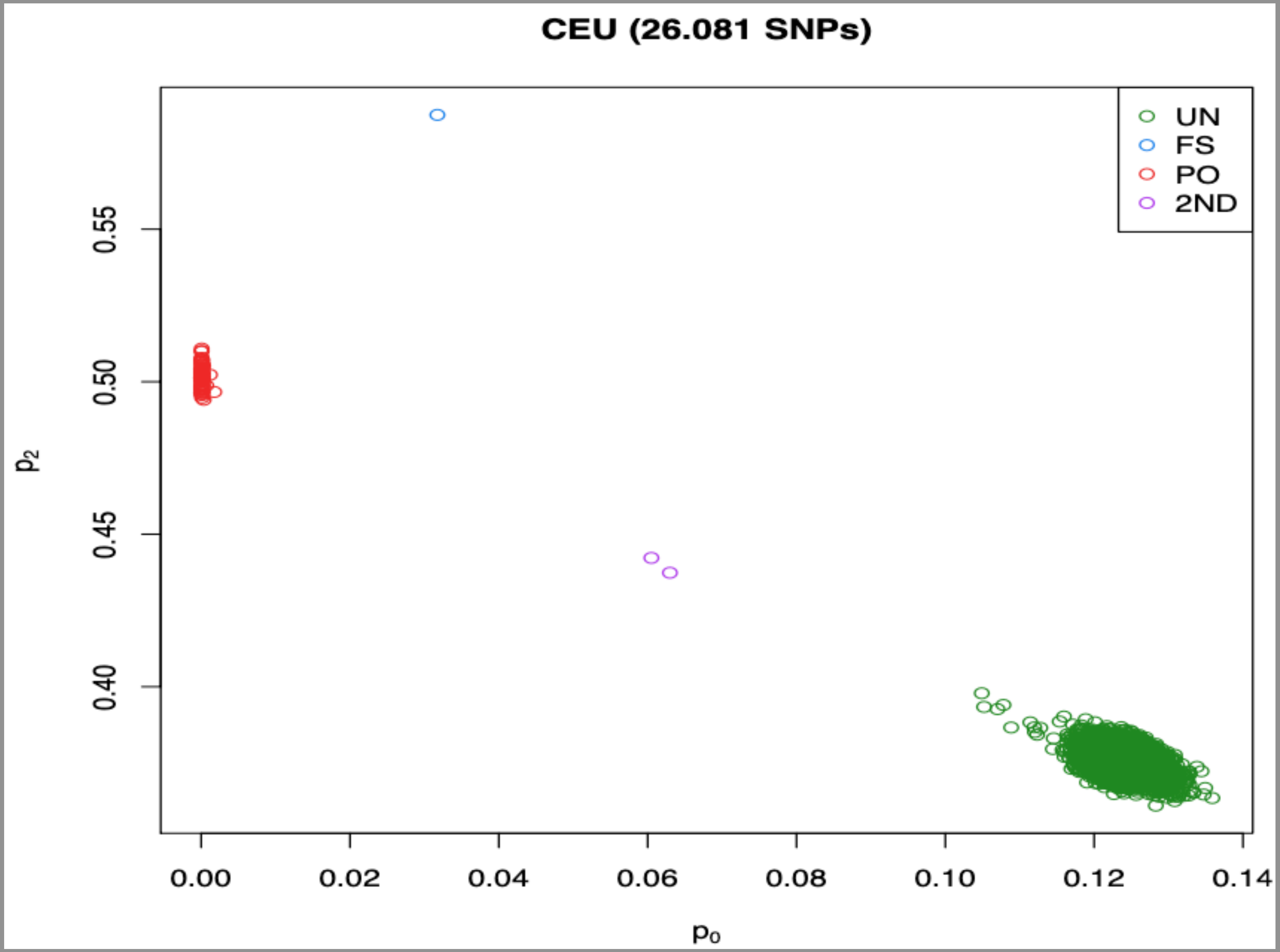
- Where $I(x_{ijk})$ reflects whether individual i and j share 0, 1 or 2 alleles for variant k. Plot $p_0$ against $p_2$…or other combinations.
- The $(p_0, p_2)$ plot leaves out one of the three proportions. The three proportions can be explicitly visualized simultaneously in a ternary diagram
- Precise position of the different clusters depends on the distribution of the allele frequencies. This plot reveals characteristic clusters that correspond to the different family relationships.
- $(p_0, p_2)$ and $(m, s)$ are mathematically related:

$$m = 1 - p_0 + p_2 \text{ and } s = \sqrt{p_0(1 - p_0) + p_2(1 - p_2) + 2p_0p_2}$$

# Relatedness analysis

## Introduction - IBS methods - $(m, s)$ plot

Example: CEU sample from the 1000 Genome project. (n = 165, p = 26.081 pruned highly variable SNPs)



# Content

Relatedness analysis (allele sharing)

1. Introduction
2. IBS methods
3. IBD methods
4. Computer exercise

# Relatedness analysis

## Introduction - IBD methods

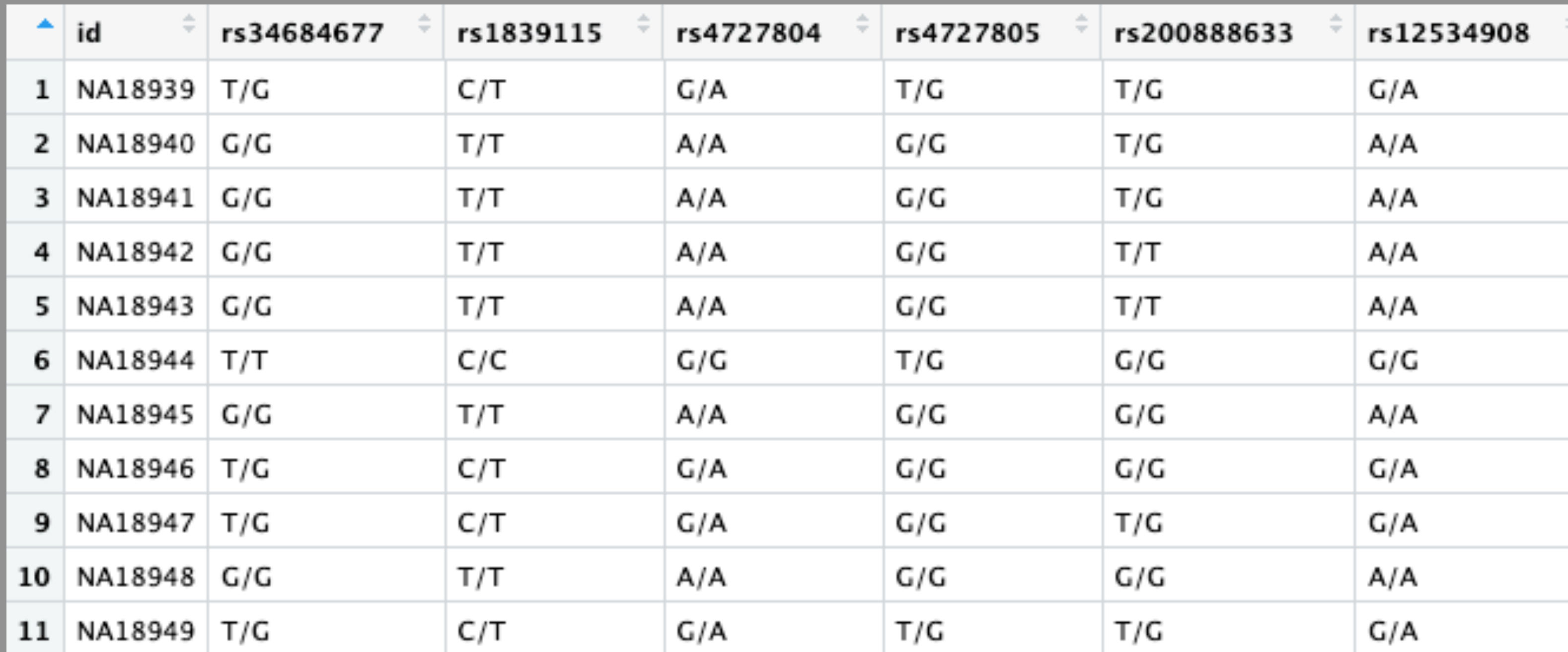
- Identity by descent (IBD) refers to the number of shared alleles at a locus between individuals, that have been inherited from a common ancestor.

- IBD probabilities for a given genotype data can be estimated (Thompson, 1975)

- IBD methods rely on the principle that individuals who share a recent common ancestor are more likely to have similar genetic material.

- If the estimated probabilities are ”close” to one of the standard relationships, then we infer that particular relationship.

- The inferred relationship may (or not) differ from the putative relationship.



| | id | rs34684677 | rs1839115 | rs4727804 | rs4727805 | rs200888633 | rs12534908 |
|---|---|---|---|---|---|---|---|
| 1 | NA18939 | T/G | C/T | G/A | T/G | T/G | G/A |
| 2 | NA18940 | G/G | T/T | A/A | G/G | T/G | A/A |
| 3 | NA18941 | G/G | T/T | A/A | G/G | T/G | A/A |
| 4 | NA18942 | G/G | T/T | A/A | G/G | T/T | A/A |
| 5 | NA18943 | G/G | T/T | A/A | G/G | T/T | A/A |
| 6 | NA18944 | T/T | C/C | G/G | T/G | G/G | G/G |
| 7 | NA18945 | G/G | T/T | A/A | G/G | G/G | A/A |
| 8 | NA18946 | T/G | C/T | G/A | G/G | G/G | G/A |
| 9 | NA18947 | T/G | C/T | G/A | G/G | T/G | G/A |
| 10 | NA18948 | G/G | T/T | A/A | G/G | G/G | A/A |
| 11 | NA18949 | T/G | C/T | G/A | T/G | T/G | G/A |

# Relatedness analysis

## Introduction - IBD methods

- IBD probabilities for a given family relationship can be estimated.
- Consider estimating the probability of obtaining different number of alleles in IBD for full siblings (FS):

$$k_0 = p(\#IBD = 0 \,|\, FS) = 0.25$$
$$k_1 = p(\#IBD = 1 \,|\, FS) = 0.50$$
$$k_2 = p(\#IBD = 2 \,|\, FS) = 0.25$$

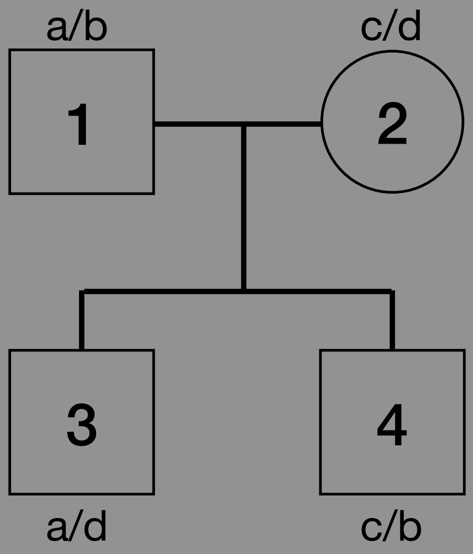
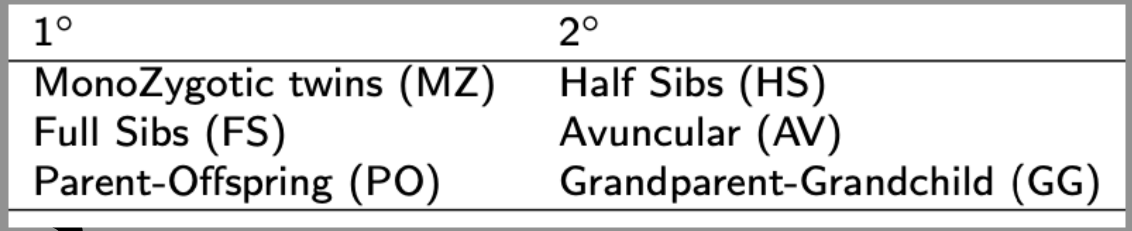


Table to reflect the probability of IBD for full siblings. Fill in number of same alleles:

| | **ac** | **cb** | **ad** | **db** |
|---|---|---|---|---|
| **ac** | 2 | 1 | 1 | 0 |
| **cb** | 1 | 2 | 0 | 1 |
| **ad** | 1 | 0 | 2 | 1 |
| **db** | 0 | 1 | 1 | 2 |

1/4 genotypes/individuals have 2 alleles in IBD
1/2 genotypes/individuals share one allele in IBD
1/4 genotypes/individuals are not IBD

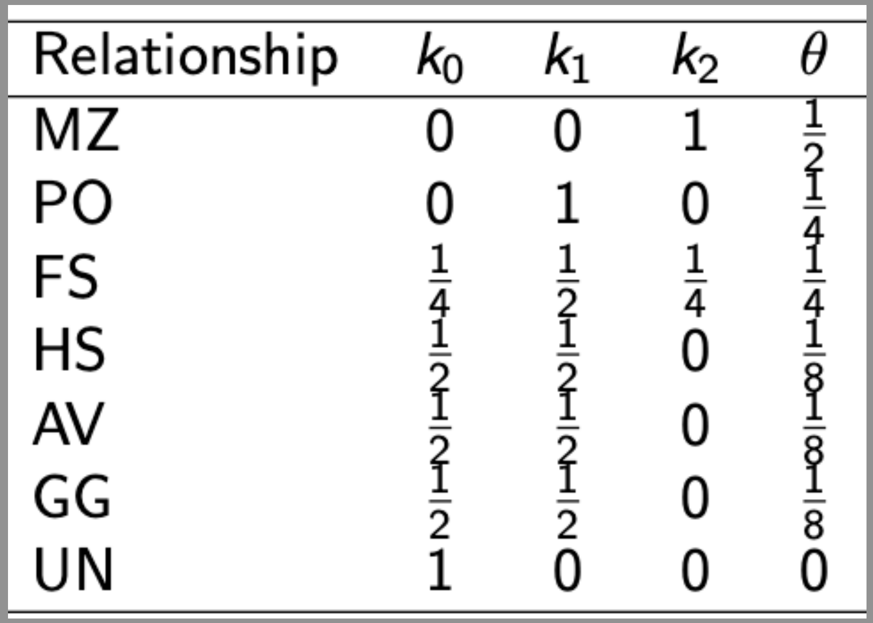
# Relatedness analysis



| 1° | 2° |
|---|---|
| MonoZygotic twins (MZ) | Half Sibs (HS) |
| Full Sibs (FS) | Avuncular (AV) |
| Parent-Offspring (PO) | Grandparent-Grandchild (GG) |

## Introduction - IBD methods

- IBD probabilities for a given family relationship can be estimated:



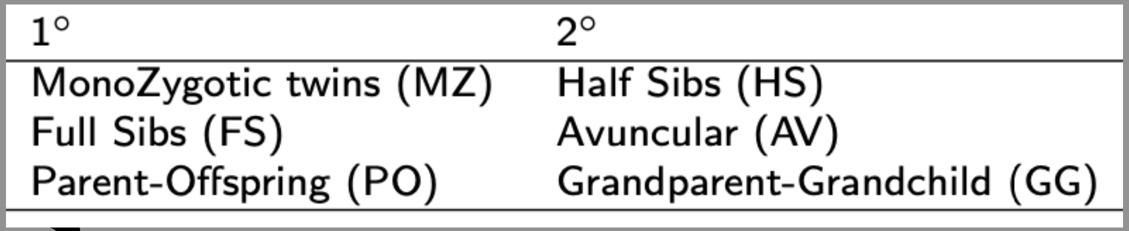
| Relationship | $k_0$ | $k_1$ | $k_2$ | $\theta$ |
|---|---|---|---|---|
| MZ | 0 | 0 | 1 | $\frac{1}{2}$ |
| PO | 0 | 1 | 0 | $\frac{1}{4}$ |
| FS | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| HS | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| AV | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| GG | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| UN | 1 | 0 | 0 | 0 |

$$k_0 = p(\#IBD = 0 \,|\, relationship)$$
$$k_1 = p(\#IBD = 1 \,|\, relationship)$$
$$k_2 = p(\#IBD = 2 \,|\, relationship)$$

- Kinship or coancestry coefficient $\theta = 1/4 \cdot k_1 + 1/2 \cdot k_2$
- **Kinship coefficient:** probability that an allele at a locus for two individuals was inherited from the same genome. Degree of consanguinity between two individuals.
- Updated definition: Probability that, for a given genetic locus, a randomly selected allele from individual i and a randomly selected allele from individual j are identical-by-descent.

# Relatedness analysis

## Introduction - IBD estimation

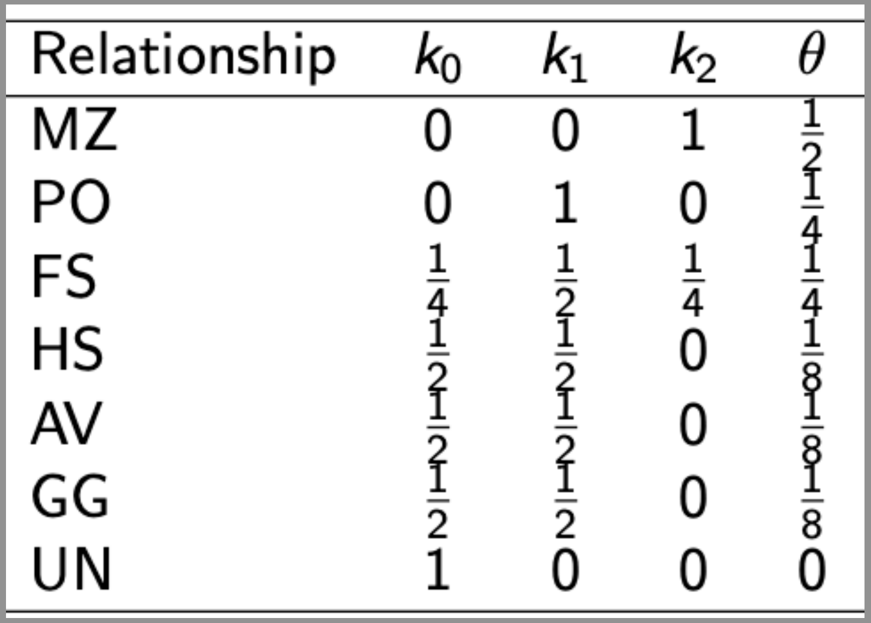


| 1° | 2° |
|---|---|
| MonoZygotic twins (MZ) | Half Sibs (HS) |
| Full Sibs (FS) | Avuncular (AV) |
| Parent-Offspring (PO) | Grandparent-Grandchild (GG) |

- Identity by descent (IBD) refers to the number of shared alleles at a locus between individuals, that have been inherited from a common ancestor.

- IBD probabilities for a given genotype data can be estimated (Thompson, 1975)

- If the estimated probabilities are ”close” to one of the standard relationships, then we infer that particular relationship.
- The inferred relationship may (or not) differ from the putative relationship.



| Relationship | $k_0$ | $k_1$ | $k_2$ | $\theta$ |
|---|---|---|---|---|
| MZ | 0 | 0 | 1 | $\frac{1}{2}$ |
| PO | 0 | 1 | 0 | $\frac{1}{4}$ |
| FS | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| HS | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| AV | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| GG | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| UN | 1 | 0 | 0 | 0 |

$$k_0 = p(\#IBD = 0 \,|\, relationship)$$
$$k_1 = p(\#IBD = 1 \,|\, relationship)$$
$$k_2 = p(\#IBD = 2 \,|\, relationship)$$

$$\theta = 1/4 \cdot k_1 + 1/2 \cdot k_2$$

# Relatedness analysis

## Introduction - IBD estimation

- Let $G_1$ and $G_2$ be the pair of genotypes observed at a locus for two individuals.
- Let $m$ (0, 1 or 2) represent the number of IBD alleles.
- And

$$k_0 = p(\#IBD = 0 \,|\, relationship)$$
$$k_1 = p(\#IBD = 1 \,|\, relationship)$$
$$k_2 = p(\#IBD = 2 \,|\, relationship)$$

- By the law of total probability, the joint genotypic probabilities:

$$P(G_1 \cap G_2 \,|\, k_0, k_1, k_2) = P(G_1 \cap G_2 \,|\, m = 0) \cdot k_0 + P(G_1 \cap G_2 \,|\, m = 1) \cdot k_1 + P(G_1 \cap G_2 \,|\, m = 2) \cdot k_2$$

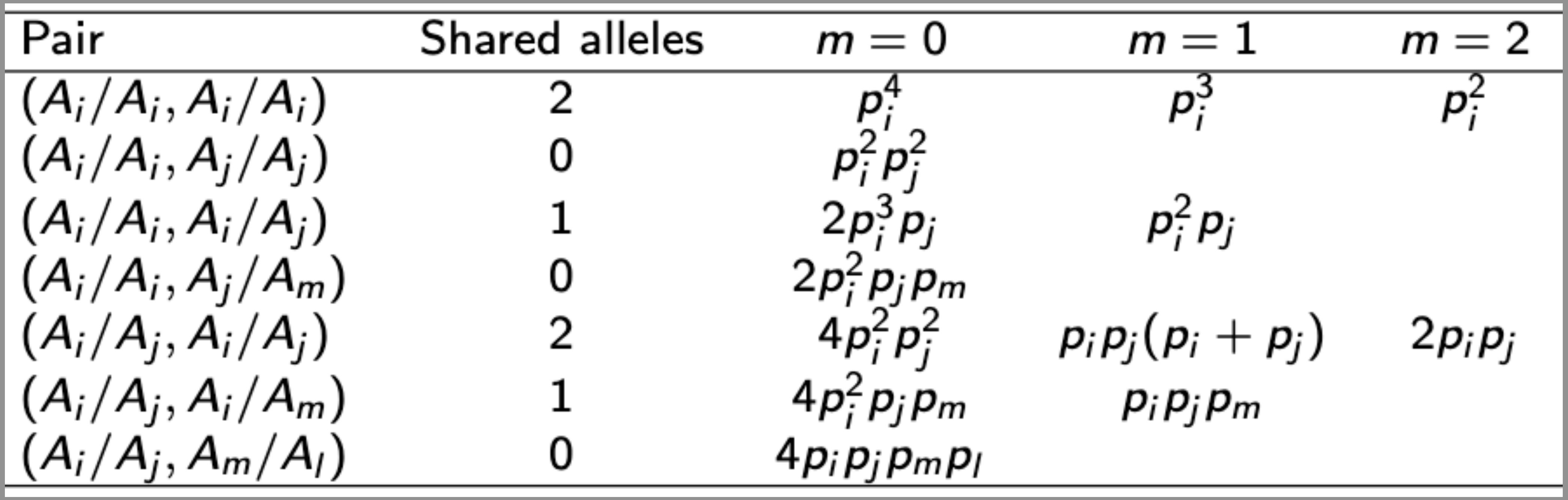
- The probabilities $P(G_1 \cap G_2 \,|\, m)$ depend on the genotypes of the individuals and are calculated from the allele frequencies in the population.
- Let $p_i$ be the $i$th allele frequency. Consider two homozygous individuals:

$$P(G_1 = A_i/A_i \cap G_2 = A_i/A_i \,|\, m = 0) = P(G_1 = A_i/A_i)P(G_2 = A_i/A_i) = p_i^2 p_i^2 = p_i^4$$
$$P(G_1 = A_i/A_i \cap G_2 = A_i/A_i \,|\, m = 1) = P(G_1 = A_i/A_i)P(G_2 = A_i/A_i \,|\, G_1 = A_i/A_i \,|\, m = 1) = p_i^2 p_i = p_i^3$$
$$P(G_1 = A_i/A_i \cap G_2 = A_i/A_i \,|\, m = 2) = P(G_1 = A_i/A_i) = P(G_2 = A_i/A_i) = p_i^2$$

# Relatedness analysis

## Introduction - IBD estimation

- For all possible genotype pairs:



| Pair | Shared alleles | $m = 0$ | $m = 1$ | $m = 2$ |
|---|---|---|---|---|
| $(A_i/A_i, A_i/A_i)$ | 2 | $p_i^4$ | $p_i^3$ | $p_i^2$ |
| $(A_i/A_i, A_j/A_j)$ | 0 | $p_i^2 p_j^2$ | | |
| $(A_i/A_i, A_i/A_j)$ | 1 | $2p_i^3 p_j$ | $p_i^2 p_j$ | |
| $(A_i/A_i, A_j/A_m)$ | 0 | $2p_i^2 p_j p_m$ | | |
| $(A_i/A_j, A_i/A_j)$ | 2 | $4p_i^2 p_j^2$ | $p_i p_j (p_i + p_j)$ | $2p_i p_j$ |
| $(A_i/A_j, A_i/A_m)$ | 1 | $4p_i^2 p_j p_m$ | $p_i p_j p_m$ | |
| $(A_i/A_j, A_m/A_l)$ | 0 | $4p_i p_j p_m p_l$ | | |

$$P(G_1 \cap G_2 | k_0, k_1, k_2) = P(G_1 \cap G_2 | m = 0) \cdot k_0 + P(G_1 \cap G_2 | m = 1) \cdot k_1 + P(G_1 \cap G_2 | m = 2) \cdot k_2$$

$$P(G_1 = A_i/A_i \cap G_2 = A_i/A_i | k_0, k_1, k_2) = p_i^4 \cdot k_0 + p_i^3 \cdot k_1 + p_i^2 \cdot k_2$$

- Let $p_i$ be the $i$th allele frequency. Consider two homozygous individuals:

$$P(G_1 = A_i/A_i \cap G_2 = A_i/A_i | m = 0) = P(G_1 = A_i/A_i)P(G_2 = A_i/A_i) = p_i^2 p_i^2 = p_i^4$$

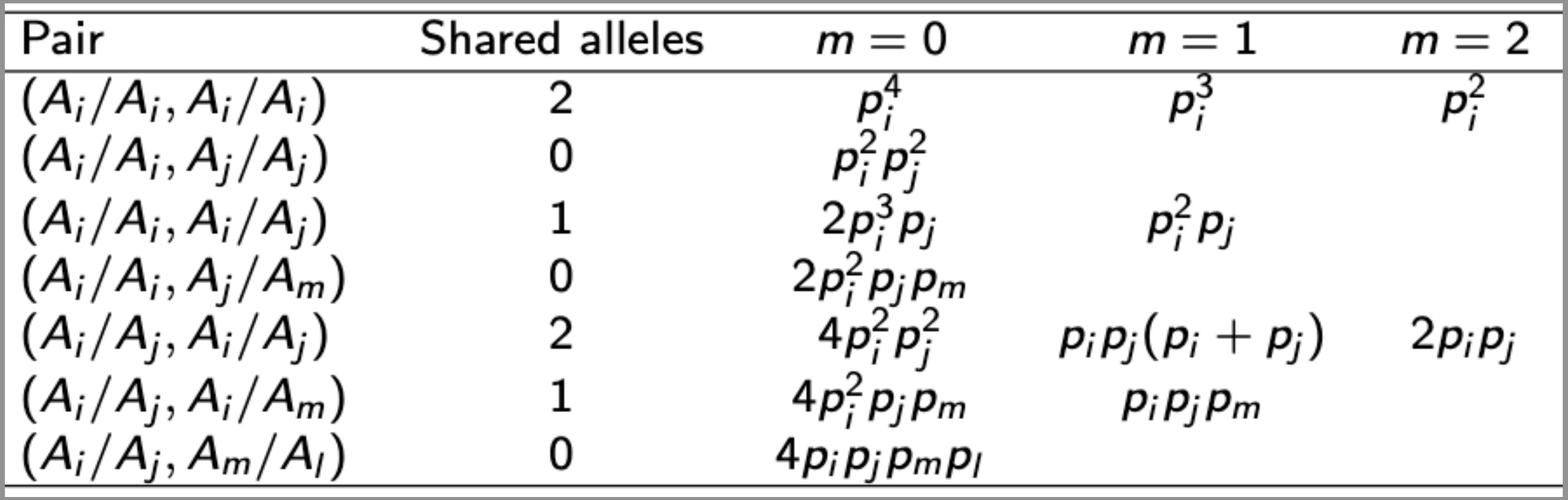
$$P(G_1 = A_i/A_i \cap G_2 = A_i/A_i | m = 1) = P(G_1 = A_i/A_i)P(G_2 = A_i/A_i | G_1 = A_i/A_i | m = 1) = p_i^2 p_i = p_i^3$$

$$P(G_1 = A_i/A_i \cap G_2 = A_i/A_i | m = 2) = P(G_1 = A_i/A_i) = P(G_2 = A_i/A_i) = p_i^2$$

# Relatedness analysis

## Introduction - IBD methods

- For all possible genotype pairs:



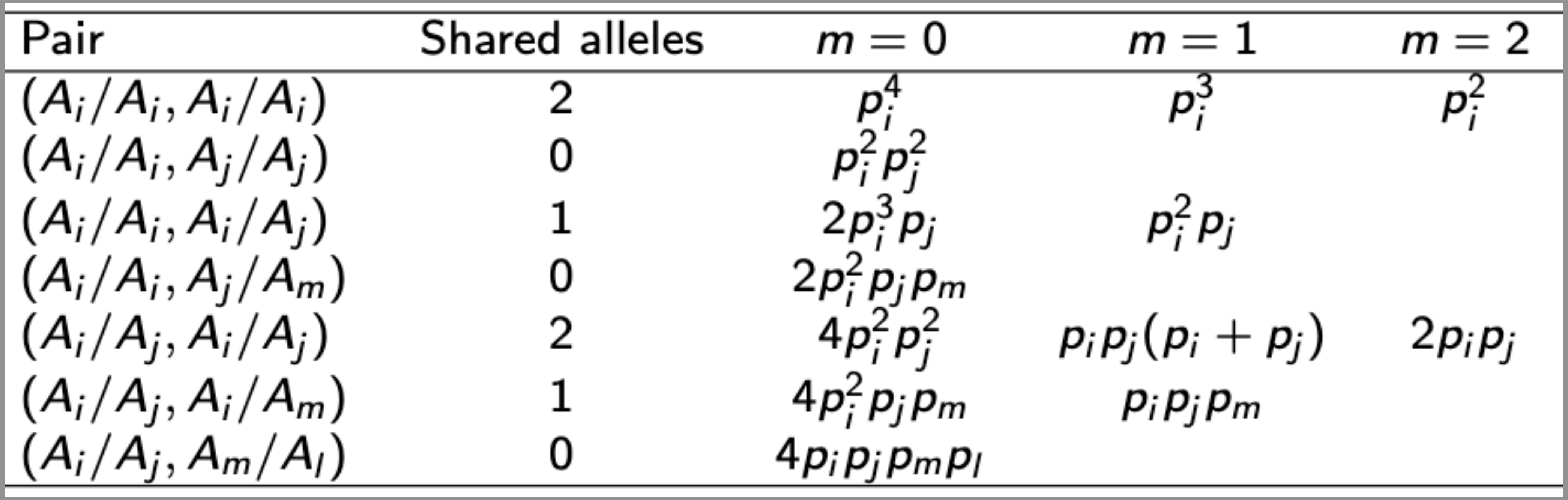
| Pair | Shared alleles | $m = 0$ | $m = 1$ | $m = 2$ |
|---|---|---|---|---|
| $(A_i/A_i, A_i/A_i)$ | 2 | $p_i^4$ | $p_i^3$ | $p_i^2$ |
| $(A_i/A_i, A_j/A_j)$ | 0 | $p_i^2 p_j^2$ | | |
| $(A_i/A_i, A_i/A_j)$ | 1 | $2p_i^3 p_j$ | $p_i^2 p_j$ | |
| $(A_i/A_i, A_j/A_m)$ | 0 | $2p_i^2 p_j p_m$ | | |
| $(A_i/A_j, A_i/A_j)$ | 2 | $4p_i^2 p_j^2$ | $p_i p_j (p_i + p_j)$ | $2p_i p_j$ |
| $(A_i/A_j, A_i/A_m)$ | 1 | $4p_i^2 p_j p_m$ | $p_i p_j p_m$ | |
| $(A_i/A_j, A_m/A_l)$ | 0 | $4p_i p_j p_m p_l$ | | |

- Assumptions:
  - Hardy-Weinberg equilibrium
  - Known population allele frequency
  - Independent variants

# Relatedness analysis

## Introduction - IBD methods

- For all possible genotype pairs:



| Pair | Shared alleles | $m = 0$ | $m = 1$ | $m = 2$ |
|---|---|---|---|---|
| $(A_i/A_i, A_i/A_i)$ | 2 | $p_i^4$ | $p_i^3$ | $p_i^2$ |
| $(A_i/A_i, A_j/A_j)$ | 0 | $p_i^2 p_j^2$ | | |
| $(A_i/A_i, A_i/A_j)$ | 1 | $2p_i^3 p_j$ | $p_i^2 p_j$ | |
| $(A_i/A_i, A_j/A_m)$ | 0 | $2p_i^2 p_j p_m$ | | |
| $(A_i/A_j, A_i/A_j)$ | 2 | $4p_i^2 p_j^2$ | $p_i p_j (p_i + p_j)$ | $2p_i p_j$ |
| $(A_i/A_j, A_i/A_m)$ | 1 | $4p_i^2 p_j p_m$ | $p_i p_j p_m$ | |
| $(A_i/A_j, A_m/A_l)$ | 0 | $4p_i p_j p_m p_l$ | | |

$$P(G_1 \cap G_2 | k_0, k_1, k_2) = P(G_1 \cap G_2 | m = 0) \cdot k_0 + P(G_1 \cap G_2 | m = 1) \cdot k_1 + P(G_1 \cap G_2 | m = 2) \cdot k_2$$
$$P(G_1 \cap G_2 | k_0, k_1, k_2) = d_0 \cdot k_0 + d_1 \cdot k_1 + d_2 \cdot k_2$$

IBD probabilities can be estimated via maximum likelihood. For $n$ individuals:

$$L(k_0, k_1, k_2 | G) = \prod_{i=1}^{n} d_{0i} \cdot k_0 + d_{1i} \cdot k_1 + d_{2i} \cdot k_2$$

# Relatedness analysis

## Introduction - IBD methods

**Software available**

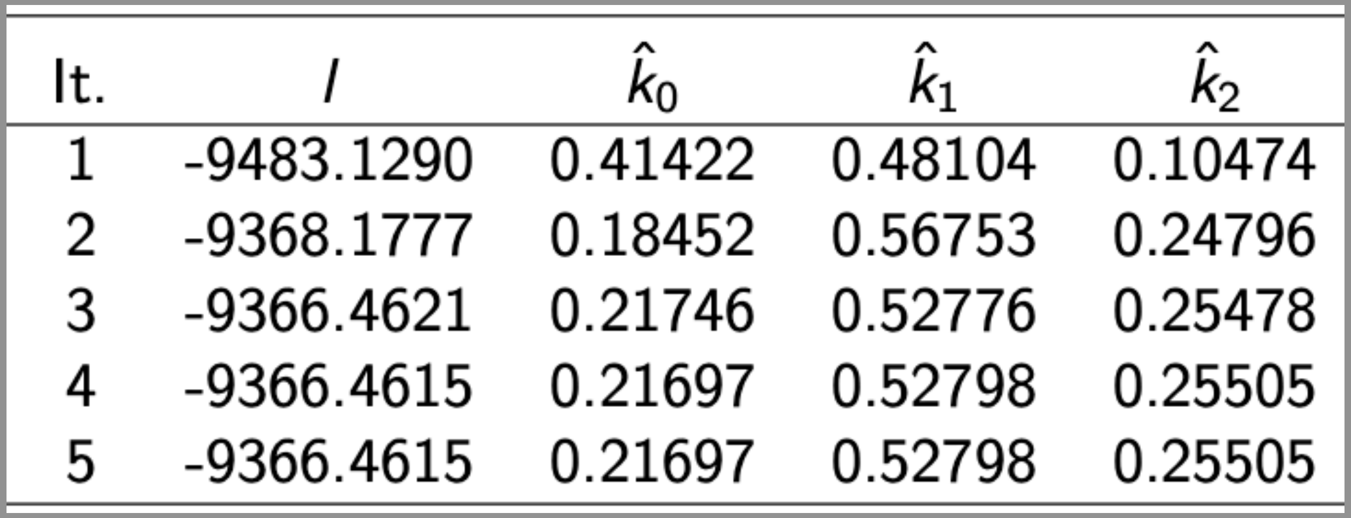
- R-package SNPRelate
- R-package GWASTools
- GRR
- Relpair
- PLINK
- ...

# Relatedness analysis

## Introduction - IBD methods

**Example: HapMap Phase III, Mexican population (n = 86)**

- ML estimation of IBD probabilities of a FS pair, using 5.000 SNPs, with initial point (0.575,0.400,0.025). Iteration history for the maximization of the log-likelihood (l)



| It. | $l$ | $\hat{k}_0$ | $\hat{k}_1$ | $\hat{k}_2$ |
|---|---|---|---|---|
| 1 | -9483.1290 | 0.41422 | 0.48104 | 0.10474 |
| 2 | -9368.1777 | 0.18452 | 0.56753 | 0.24796 |
| 3 | -9366.4621 | 0.21746 | 0.52776 | 0.25478 |
| 4 | -9366.4615 | 0.21697 | 0.52798 | 0.25505 |
| 5 | -9366.4615 | 0.21697 | 0.52798 | 0.25505 |

- IBD statistics can be visualized by plotting estimates of IBD probabilities with 0, 1 or 2: $(k_0, k_1)$ plot. The $(k_0, k_1)$ plot leave out one of the three proportions. The three proportions can be explicitly visualized simultaneously in a ternary diagram.

# Relatedness analysis

## Introduction - IBD methods

### Example: IBD probabilities with SNPRelate

- ML estimation of IBD probabilities of a 279 individuals with pruned 910 SNPs. Hapmap data.

```
> library(gdsfmt)
> library(SNPRelate)

> data(hapmap_geno)

> ibd <- snpgdsIBDMLE(genofile,maf=0.05, missing.rate=0.05,snp.id=snpset.id, num.thread=2)


# Create a gds file
> snpgdsCreateGeno("test.gds", genmat = hapmap_geno$genotype,
               sample.id = hapmap_geno$sample.id, snp.id = hapmap_geno$snp.id,
               snp.chromosome = hapmap_geno$snp.chromosome,
               snp.position = hapmap_geno$snp.position,
               snp.allele = hapmap_geno$snp.allele, snpfirstdim=TRUE)

# Open the GDS file
> genofile <- snpgdsOpen(“test.gds")

# LD pruning

> snpset <- snpgdsLDpruning(genofile, ld.threshold=0.2)
> snpset.id <- unlist(unname(snpset))

# Estimate IBD coefficients
> ibd <- snpgdsIBDMLE(genofile,maf=0.05, missing.rate=0.05,snp.id=snpset.id, num.thread=2)
```

# Relatedness analysis

## Introduction - IBD methods

**Example: IBD probabilities with SNPRelate**

- ML estimation of IBD probabilities of a 279 individuals with pruned 910 SNPs. Hapmap data.

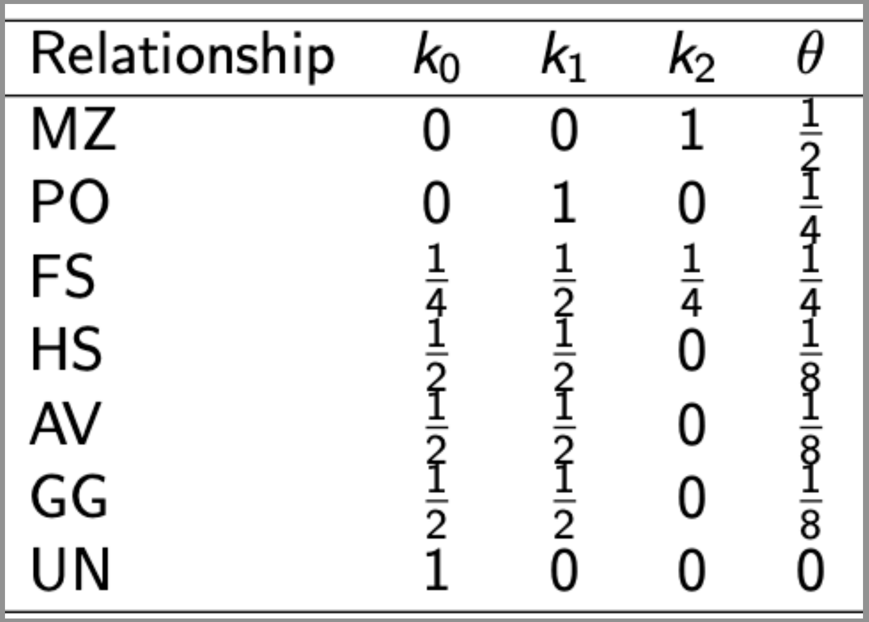
```
# Estimate IBD coefficients
> ibd <- snpgdsIBDMLE(genofile,maf=0.05, missing.rate=0.05,snp.id=snpset.id, num.thread=2)

Identity-By-Descent analysis (MLE) on genotypes:
Excluding 90 SNPs (non-autosomes or non-selection)
Excluding 129 SNPs (monomorphic: TRUE, MAF: 0.05, missing rate: 0.05)
    # of samples: 279
    # of SNPs: 781
    using 2 threads
MLE IBD:    the sum of all selected genotypes (0,1,2) = 214309
MLE IBD:   Mon Dec 11 13:51:22 2023      0%
MLE IBD:   Mon Dec 11 13:51:47 2023      100%


# Make a data.frame
> ibd.coeff <- snpgdsIBDSelection(ibd)

# Plot
> plot(ibd.coeff$k0, ibd.coeff$k1, xlim=c(0,1), ylim=c(0,1),
     xlab="k0", ylab="k1", main="YRI samples (MLE)")
> lines(c(0,1), c(1,0), col="red", lty=2)
```

# Relatedness analysis

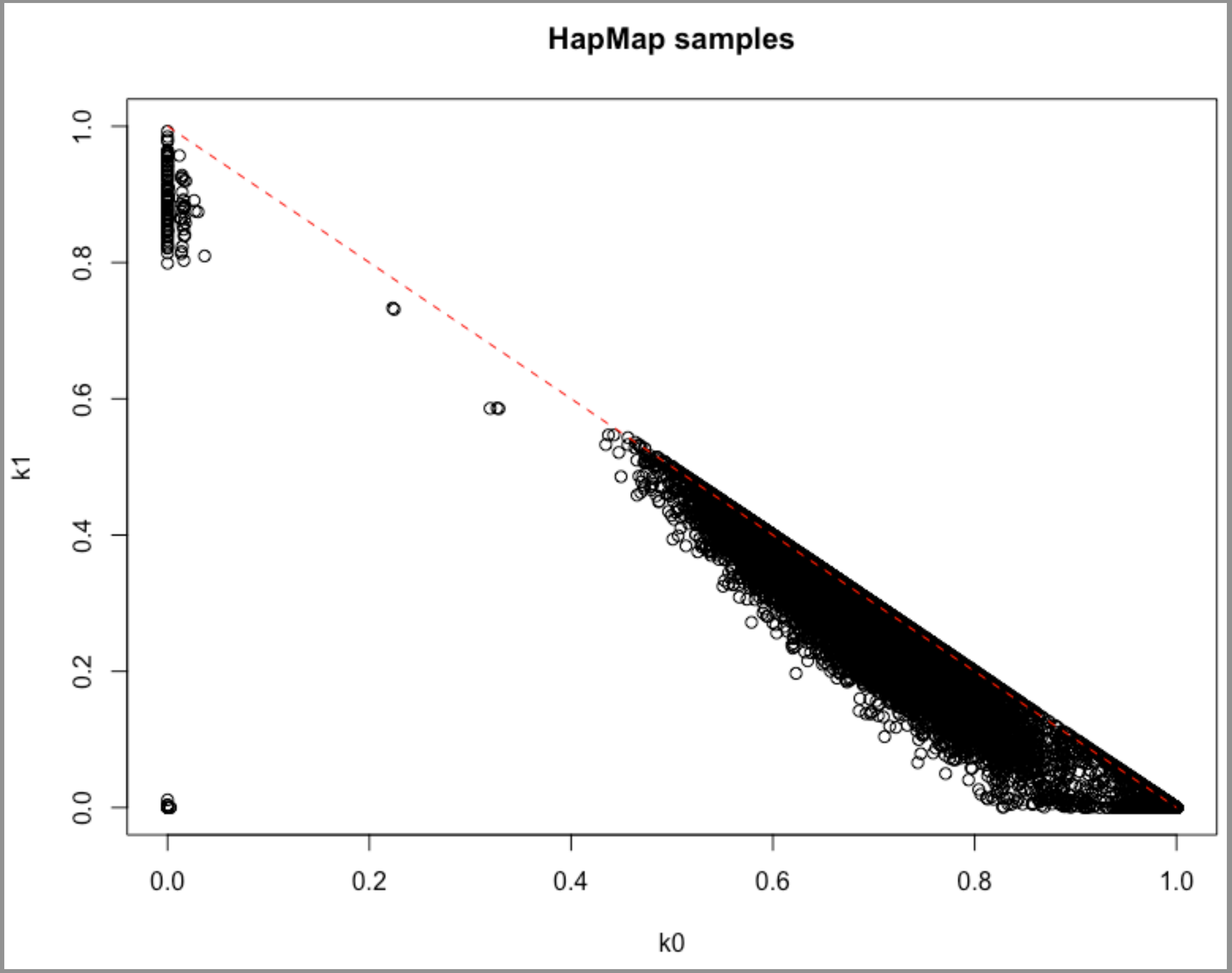
## Introduction - IBD methods

**Example: IBD probabilities with SNPRelate**

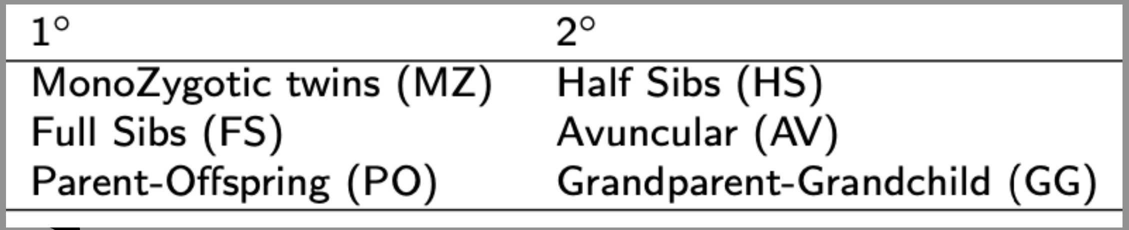
- ML estimation of IBD probabilities of a 279 individuals with pruned 910 SNPs. Hapmap data.



| Relationship | $k_0$ | $k_1$ | $k_2$ | $\theta$ |
|---|---|---|---|---|
| MZ | 0 | 0 | 1 | $\frac{1}{2}$ |
| PO | 0 | 1 | 0 | $\frac{1}{4}$ |
| FS | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| HS | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| AV | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| GG | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| UN | 1 | 0 | 0 | 0 |



$$k_0 = p(\#IBD = 0\,|\,relationship)$$
$$k_1 = p(\#IBD = 1\,|\,relationship)$$
$$k_2 = p(\#IBD = 2\,|\,relationship)$$



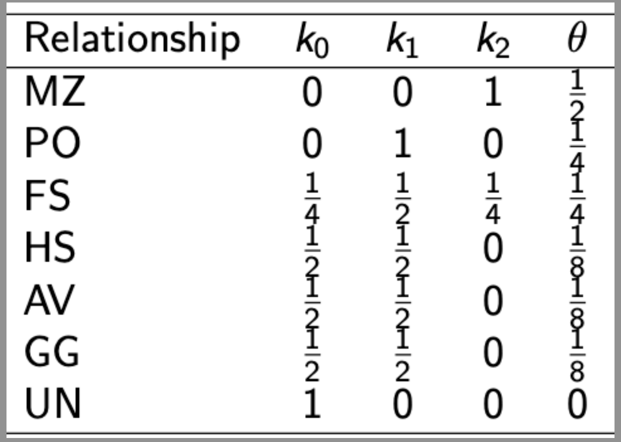
| 1° | 2° |
|---|---|
| MonoZygotic twins (MZ) | Half Sibs (HS) |
| Full Sibs (FS) | Avuncular (AV) |
| Parent-Offspring (PO) | Grandparent-Grandchild (GG) |

# Relatedness analysis

## Introduction - IBD methods

**Example: IBD probabilities with SNPRelate**

- ML estimation of IBD probabilities of a 279 individuals with pruned 910 SNPs. Hapmap data.



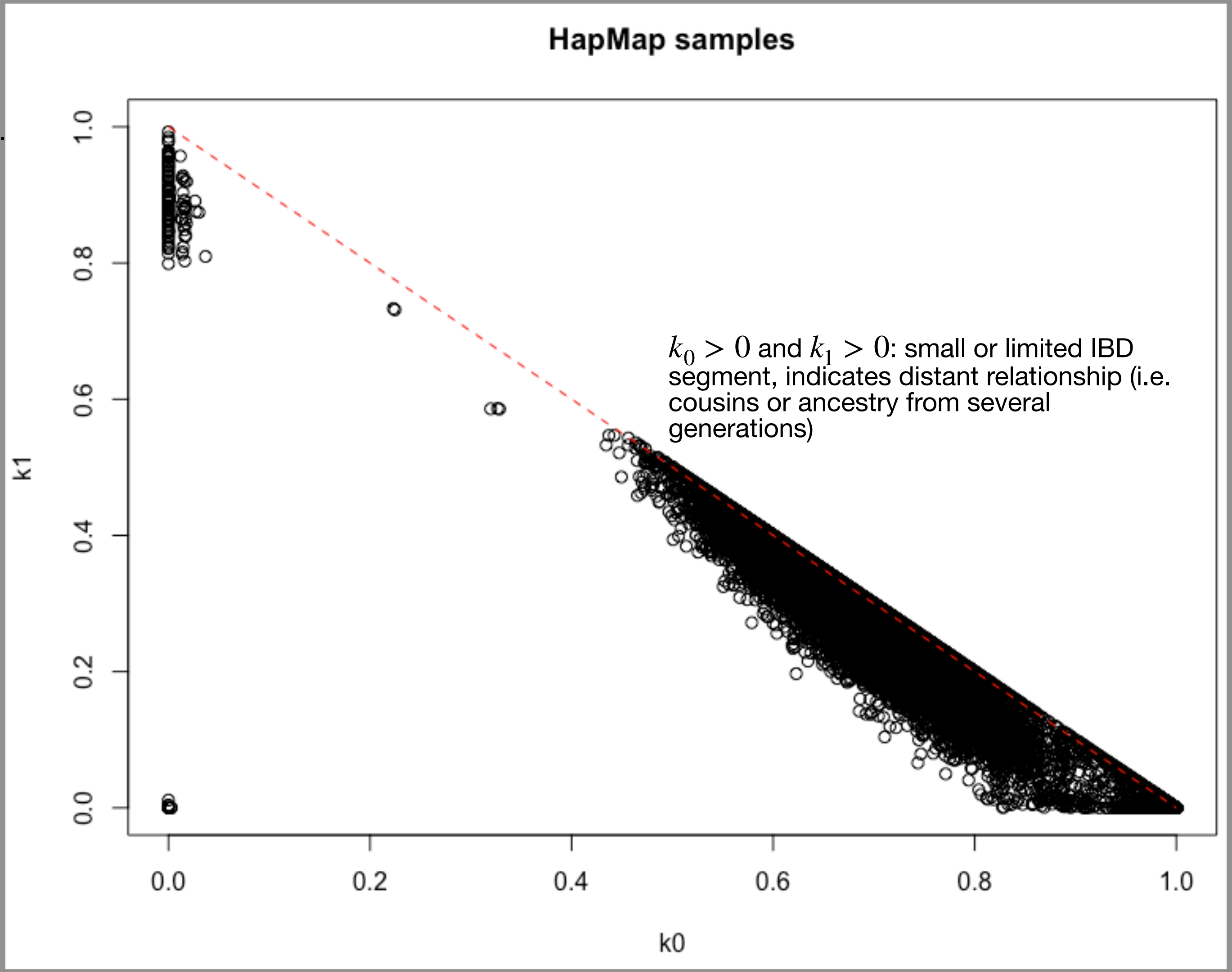
| Relationship | $k_0$ | $k_1$ | $k_2$ | $\theta$ |
|---|---|---|---|---|
| MZ | 0 | 0 | 1 | $\frac{1}{2}$ |
| PO | 0 | 1 | 0 | $\frac{1}{4}$ |
| FS | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |
| HS | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| AV | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| GG | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | $\frac{1}{8}$ |
| UN | 1 | 0 | 0 | 0 |

RECALL:

$k_0 = p(\#IBD = 0 \,|\, relationship)$
$k_1 = p(\#IBD = 1 \,|\, relationship)$
$k_2 = p(\#IBD = 2 \,|\, relationship)$

$k_0 = 0$ and $k_1 > 0$: there are IBD segments.
Usually:

- $k_1 \approx 0.25$ half siblings
- $k_1 \approx 0.5$ full siblings
- $k_1 > 0.5$ parent-offspring



$k_0 > 0$ and $k_1 > 0$: small or limited IBD segment, indicates distant relationship (i.e. cousins or ancestry from several generations)

$k_0 = 0$ and $k_1 = 0$: indicative of monozygotic twins

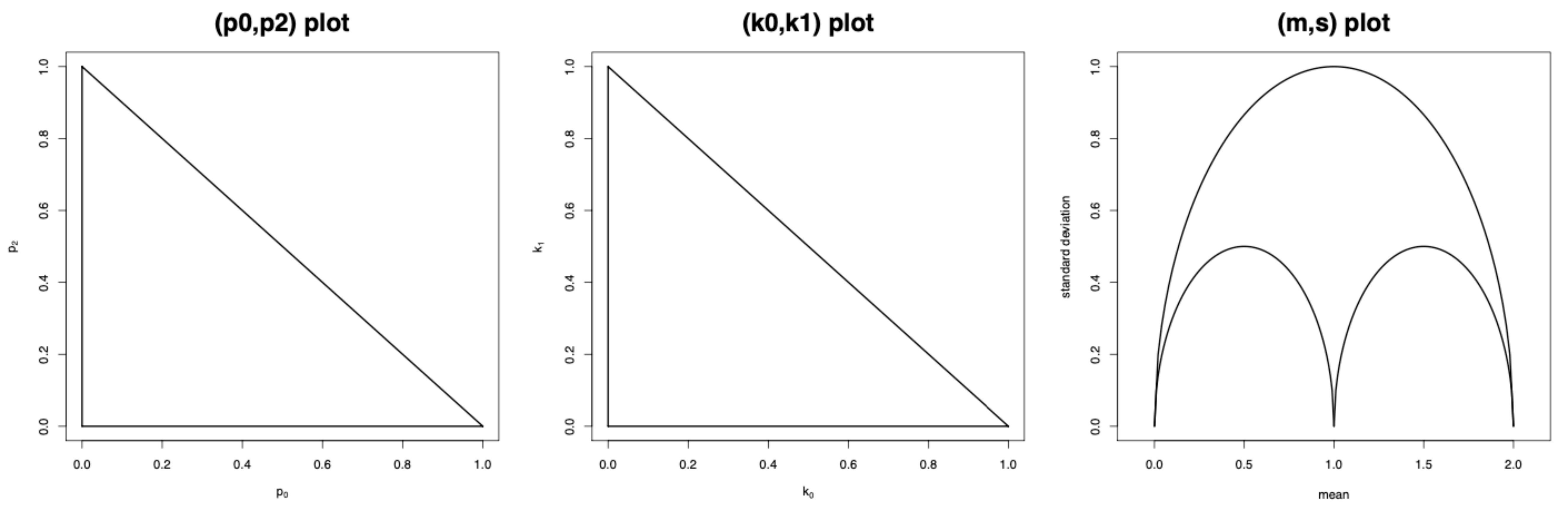
$k_0 > 0$ and $k_1 \approx 0$: no detected IBD segments, any genetic similarity observed is due to chance.

# Relatedness analysis

## Introduction - IBD methods

**Restrictions on estimators**

- All estimates live in a constrained space
  - $(p_0, p_1, p_2)$ is a composition with $p_0 + p_1 + p_2 = 1$
  - $(k_0, k_1, k_2)$ is a composition with $k_0 + k_1 + k_2 = 1$



(p0,p2) plot

(k0,k1) plot

(m,s) plot

# Content

Relatedness analysis (allele sharing)

1. Introduction
2. IBS methods
3. IBD methods
4. Computer exercise

# Relatedness analysis

## References

- Abecasis, G.R., Cherny, S.S., Cookson W.O.C. and Cardon, L. R. (2001) GRR: graphical representation of relationship errors. Bioinformatics, 17(8) pp. 742–743.

- Graffelman, J., Galv´an-Femen´ıa, I., De Cid, R., and Barcel´o-Vidal, C. (2019) A log-ratio biplot approach for exploring genetic relatedness based on identity by state. Frontiers in Genetics doi: 10.3389/fgene.2019.00341

- Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Annals of Human Genetics, 70: 841-847.

- Thompson, E.A. (1975). The estimation of pairwise relationships. Annals of Human Genetics, 39(2): 173-188.

- Weir, B.S., Anderson, A.D., Hepler, A.B. (2006) Genetic relatedness analysis: modern data and new challenges. Nature Review Genetics 7(10) pp. 771–780.