# Relatedness analysis (allele sharing)

Jan Graffelman[1]

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

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

jan.graffelman@upc.edu

December 17, 2019

# Contents

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

# Motivation

The detection of (closely) related individuals in genetic studies is of interest in various contexts.

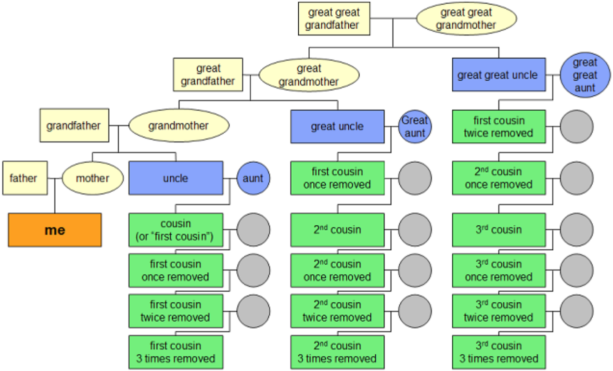
- In association studies, many methods assume independent individuals. Closely related individuals will not be independent.
- In conservation genetics, breeding programs are set up for preferably unrelated individuals.
- In quality control of the data, samples can be accidentally duplicated, and it is 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.
- ...

# Close and remote relatedness

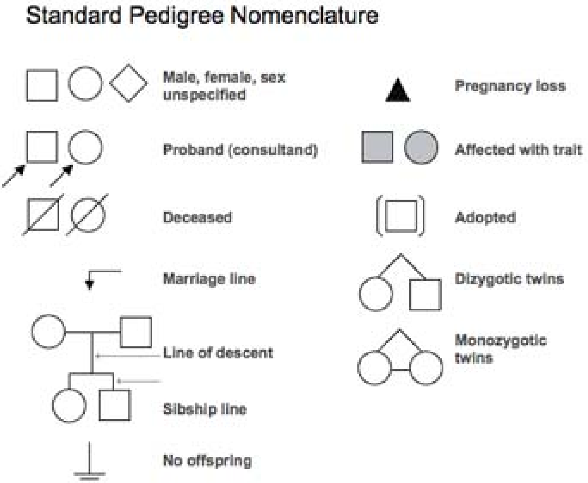
- A distinction is generally drawn between
  - Close or recent relatedness: family relationships (MZ, PO, FS, HS, AV, FC, ...)
  - Distant or remote relatedness: population substructure (non-homogeneous genetic data)
- Here we mostly address recent relatedness
- Focus is on 1° and 2° relationships:

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

# Close relatedness: family relationships



# Close relatedness: family data and pedigrees

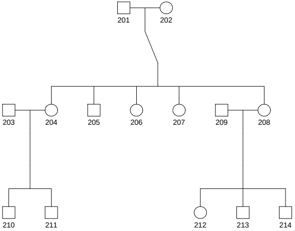


# Coding of family data

- A database of related individuals is typically coded in .ped file format.
- Besides the genotype information, Family ID, Sample ID, Paternal ID, Maternal ID, Sex (1=male; 2=female; other=unknown) and Affection status (1=affected; 0=unaffected) are registered.

Example:

| 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 |



# Allele Sharing

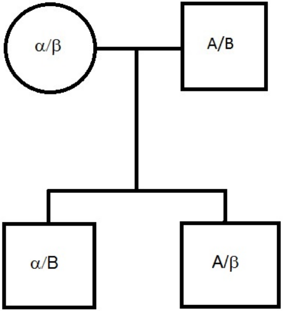
Much of relatedness research is based on the principle of allele sharing

- 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.

# IBD and IBS

- 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.

# IBD and IBS



2 alleles IBS but 0 alleles IBD

# IBS alleles

- 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.
- E.g., 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.
- An average of 2 would mean that the two individuals are identical (monozygotic twins) or that a sample has been accidentally duplicated.
- This principle can be used to uncover closely related individuals, or to detect sample heterogeneity (individuals from different populations).

# Allele sharing

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
- By plotting estimates of IBD probabilities with 0, 1 or 2 IBS alleles: $(k_0, k_1)$ plot

Notes:

- The $(p_0, p_2)$ plot and $(k_0, k_1)$ plot leave out one of the three proportions. The three proportions can be explicitly visualized simultaneously in a ternary diagram
- Variants with multiple alleles (e.g. microsatellites) are more informative for discriminating relationship categories than bi-allelic variants (SNP data).
- High MAF variants are more informative for discriminating relationship categories.

# $(m, s)$ plot (Abecasis et al, 2001)

- Let

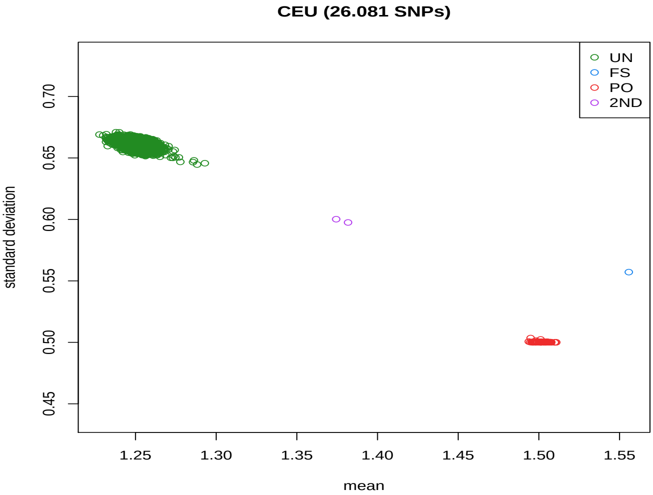
  $x_{ijk}$ = number of shared alleles between individual $i$ and $j$ for marker $k$ (0,1,2)

- Compute:

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

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

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



# $(p_0, p_2)$ plot (Rosenberg et al, 2001)

- Compute for each pair *ij*

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

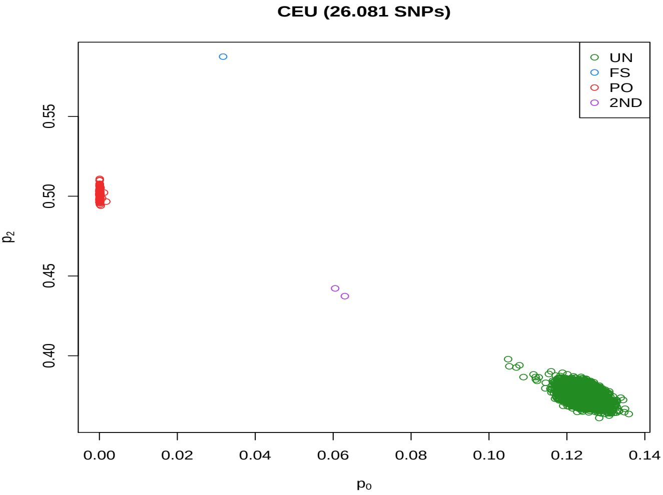
$$p_1 = \frac{1}{K}\sum_{k=1}^{K} I_{(x_{ijk}=1)}$$

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

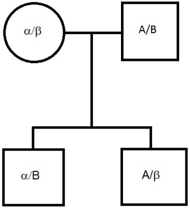
- Plot $p_0$ against $p_2$ (or other combinations).
- This plot also reveals clusters that correspond to the different family relationships.
- $(p_0, p_2)$ and $(m, s)$ are mathematically related:

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

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



# IBD probabilities for a given relationship



| | $\alpha/A$ | $\alpha/B$ | $\beta/A$ | $\beta/B$ |
|---|---|---|---|---|
| $\alpha/A$ | 2 | 1 | 1 | 0 |
| $\alpha/B$ | 1 | 2 | 0 | 1 |
| $\beta/A$ | 1 | 0 | 2 | 1 |
| $\beta/B$ | 0 | 1 | 1 | 2 |

Cotterman coefficients:

$$\begin{aligned} k_0 &= P(\#IBD = 0|FS) = 0.25 \\ k_1 &= P(\#IBD = 1|FS) = 0.50 \\ k_2 &= P(\#IBD = 2|FS) = 0.25 \end{aligned}$$

# Cotterman coefficients

Identity-by-descent probabilities for some standard relationships:

| 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 |

- Kinship or coancestry coefficient: $\theta = \frac{1}{4} k_1 + \frac{1}{2} k_2$
- Probability that two alleles at a locus, one taken at random from two individuals, are identical-by-descent.

# Procedure

- IBD probabilities can be estimated from the genotype data by maximum likelihood (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.

# ML approach

- 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.
- By the law of total probability:

$$P\left(G_1 \cap G_2 | k0, k1, k2\right) = P\left(G_1 \cap G_2 | m = 0\right) k_0 + P\left(G_1 \cap G_2 | m = 1\right) k_1 + P\left(G_1 \cap G_2 | m = 2\right) k_2$$

- The probabilities $P\left(G_1 \cap G_2 | m = 0\right)$ depend on the genotypes of the individuals and are calculated from the allele frequencies in the population.

# Calculating the joint genotype probabilities

Let $G_1 = i/i$ and $G_2 = i/i$, and let $p_i$ be the $i$th allele frequency.

$$
\begin{aligned}
P\left(G_1 = i/i \cap G_2 = i/i | m = 0\right) &= P\left(G_1 = i/i\right) P\left(G_2 = i/i\right) = p_i^2 p_i^2 = p_i^4 \\
P\left(G_1 = i/i \cap G_2 = i/i | m = 2\right) &= P\left(G_1 = i/i\right) = P\left(G_2 = i/i\right) = p_i^2 \\
P\left(G_1 = i/i \cap G_2 = i/i | m = 1\right) &= P\left(G_1 = i/i\right) P\left(G_2 = i/i | G_1 = i/i | m = 1\right) = p_i^2 p_i = p_i^3
\end{aligned}
$$

# 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 | $4p_i p_j p_m p_l$ | | |

$$P\left(G_1 \cap G_2 | k_0, k_1, k_2\right) = d_0 k_0 + d_1 k_1 + d_2 k_2$$

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

Assumptions:

- Hardy-Weinberg equilibrium
- Known population allele frequencies
- Independent variants

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

| 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 |

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 ($l$)

# Software for relatedness research

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

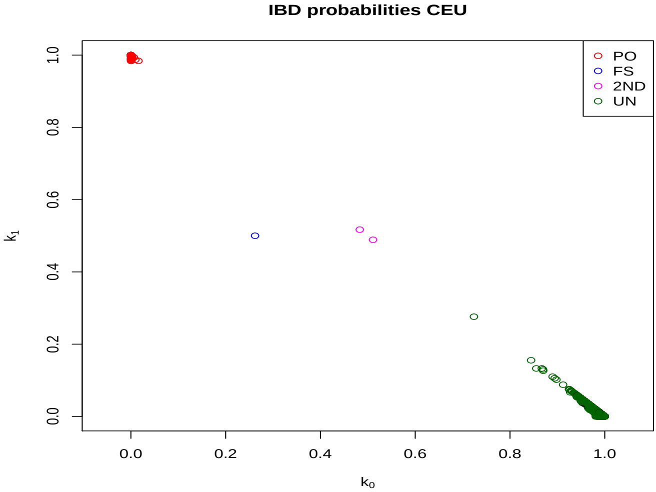
# Estimation of IBD probabilities with PLINK

```
#
# convert .ped to .bed and .fam files
#
runstring01 <- paste("plink -file hapmap3_r3_b36_fwd.consensus.qc.poly",
                     " -make-bed -out hapmap",sep="")
system(runstring01)

#
# Select the CEU individuals
#

runstring02 <- "plink -bfile hapmap -keep CEUsubset.txt -make-bed -out CEU"
system(runstring02)

#
# exclude the X chromosome
#

runstring03 <- "plink --bfile CEU --chr 1-22 --make-bed -out CEU2"
system(runstring03)

#
# Selecting complete SNPs with MAF > 0.40 only
#

runstring05 <- "plink --bfile CEU2 --geno 0 --maf 0.40 -make-bed -out CEU3"
system(runstring05)

#
# HWE filter
#

runstring06 <- "plink --bfile CEU3 --hwe 0.05 midp -make-bed -out CEU4"
system(runstring06)
```

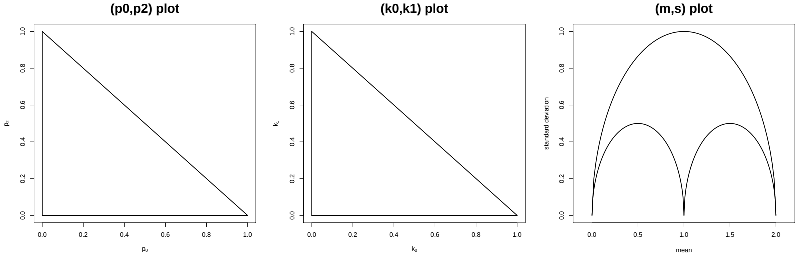
# Estimation of IBD probabilities with PLINK

```
#
# LD pruning
#

runstring07 <- "plink --bfile CEU4 --indep-pairwise 50 5 0.2  -make-bed -out CEU5"
system(runstring07)

runstring08 <- "plink --bfile CEU5 --extract CEU5.prune.in --make-bed --out CEU6"
system(runstring08)

#
# Calculate IBD probabilities
#

runstring09 <- "plink --bfile CEU6 --genome --genome-full --out CEU7"
system(runstring09)

#
# Read the IBD probabilities in R
#

X <- read.table("CEU7.genome",header=TRUE)

#
# Make a k_0 versus k_1 plot
#

plot(X$Z0,X$Z1,asp=1,xlab=expression(k[0]),ylab=expression(k[1]),main="IBD probabilities CEU")
```

# IBD probabilities CEU sample



# Restrictions on estimators



All estimators 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 $k_0 + k_1 + k_2 = 1$

# Some 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., Galván-Femenía, I., De Cid, R., and Barceló-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.

# Computer exercise

The filed `YRI.raw` contains SNPs of a Yoruba population consisting of parent-offspring trios (2 parents and 1 child). We wish to investigate if the genetic data is consistent with the specified relationships. The PLINK files `YRI.fam`, `YRI.bed` and `YRI.bim` are also available.

- Load the data in YRI.raw
- Compute the mean $m$ of the number of alleles shared for each pair of individuals.
- Compute the standard deviation $s$ of the number of alleles shared for each pair of indiduals.
- Plot all pairs in a scatterplot of $s$ against $m$.
- Plot the fraction of variants for which the individuals share 0 alleles against the fraction of variants for which the individuals share 2 alleles, and try to interpret the results.
- Use PLINK to estimate the IBD probabilities, and plot the probabilities of sharing 0 and 1 IBD alleles ($k_0$ and $k_1$) for all pairs of individuals.
- Do you think all relationships between all individuals were correctly specified?