

7. Relatedness analysis: allele sharing

Iván Galván-Femenía¹, Jan Graffelman¹

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

December 22, 2016

ivan.galvan@upc.edu

Master in Innovation and Research in Informatics (MIRI)



UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Masters in **Computer Science** and **Engineering**

Contents

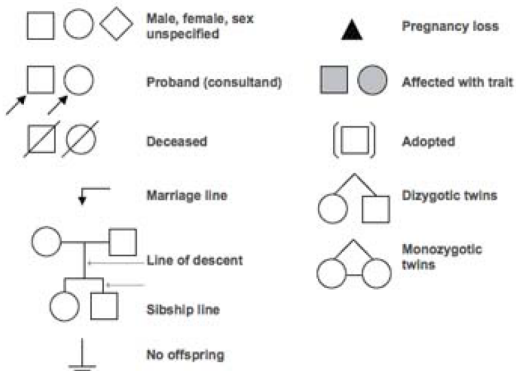
- 1 Introduction
- 2 Allele sharing
- 3 IBS studies
- 4 IBD studies
- 5 Computer exercises

Allele sharing

- This course focuses on population based genetic studies (unrelated individuals).
- Sometimes not all individuals are unrelated, and it is of interest to detect closely related individuals (parent-offspring, sister-brother, ...)
- Though not direct family members, groups of relatively more related individuals may exist in the data (e.g. population substructure) and it is of interest to detect this.
- Allele sharing analysis may help in both situations.

Family data and pedigrees

Standard Pedigree Nomenclature

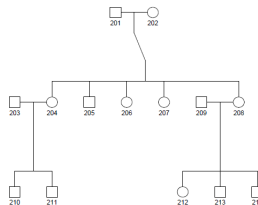


Coding of family data

- 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



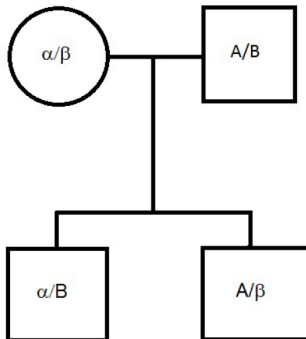
Allele sharing

- 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.

IBS and IBD

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

IBS and IBD



2 alleles IBS but 0 alleles IBD

Family relationships

- If family relationships are known (e.g. coded in a .ped file) then an allele sharing analysis can help to see if relations are correctly specified.
- This can be done graphically and numerically.
- Markers with many alleles (e.g. microsatellites) are more useful than just bi-allelic markers.

Family relationships

- Population-based studies that search for the genetic factors affecting disease usually assume individuals to be independent.
- It is therefore of interest to detect undocumented or misspecified relationships between individuals in genetic association studies.
- The huge amount of available genetic information will be helpful for detecting such relationships.
- Focus is on 1st and 2nd degree relationships:

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

Useful statistical tools in this context

- Graphical methods
 - Plots of summary statistics of the degree of allele sharing (matching) between individuals.
 - Multidimensional scaling (MDS).
 - Principal component analysis.
 - ...
- Likelihood-based methods
 - RELPAIR program: compute the likelihood of genetic marker data conditional on a given relationship (Epstein, et al. (2000)).
 - Good at classifying 1st relationships (MZ, FS, PO), poor at distinguishing 2nd relationships (HS, AV, GG).
 - Maximum likelihood estimation.

Allele sharing (matching)

- Each individual inherits two alleles for a particular genetic marker.
- A pair of individuals can share 0, 1 or 2 alleles.
- The closer the family relationship, the more alleles a pair will, on average, share.
- E.g., for an A/T single nucleotide polymorphism (SNP):

	AA	AT	TT
AA	2	1	0
AT	1	2	1
TT	0	1	2

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

Allele sharing (matching)

- Count:

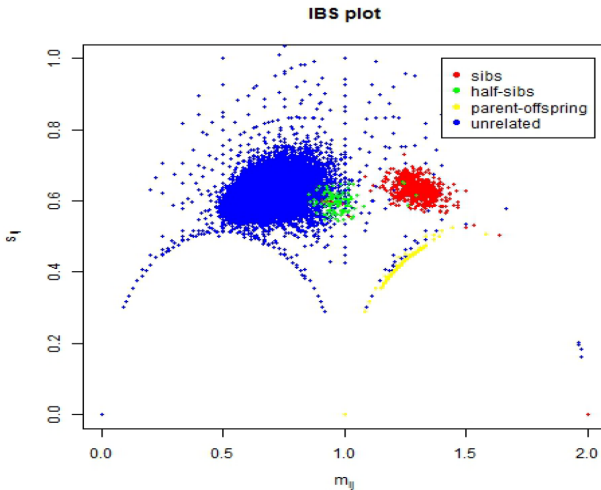
x_{ijk} = number of shared alleles between individual i and j for marker k (0,1,2)

- Compute:

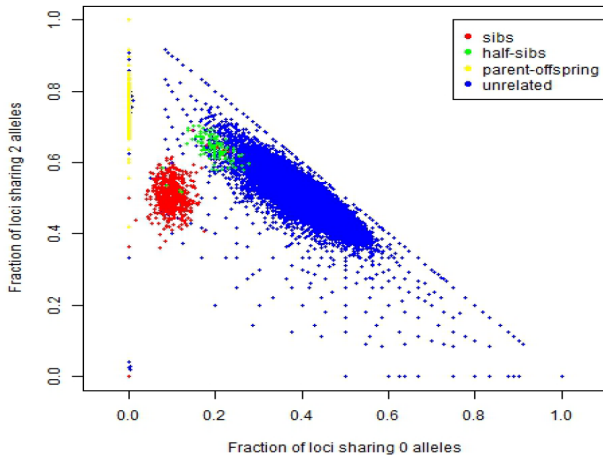
$$m_{ij} = \frac{1}{K} \sum_{k=1}^K x_{ijk} \quad \text{and} \quad s_{ij}^2 = \frac{1}{K-1} \sum_{k=1}^K (x_{ijk} - m_{ij})^2$$

- Plot m_{ij} against s_{ij} .
- This plot reveals characteristic clusters that correspond to the different family relationships.
- Popular alternative: scatter plot of the fraction of loci sharing 2 alleles against the fraction of loci sharing 0 alleles.

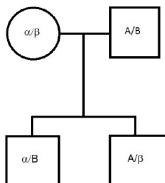
Example ($n = 1108$; $p = 320$ microsatellites, Abecasis, 2001)



Example ($n = 1108$; $p = 320$ microsatellites, Abecasis, 2001)



IBD probabilities for a given relationship



	α/A	α/B	β/A	β/B
α/A	2	1	1	0
α/B	1	2	0	1
β/A	1	0	2	1
β/B	0	1	1	2

Cotterman coefficients:

$$k_0 = P(\#IBD = 0|FS) = 0.25$$

$$k_1 = P(\#IBD = 1|FS) = 0.50$$

$$k_2 = P(\#IBD = 2|FS) = 0.25$$

Cotterman coefficients

Identity-by-descent probabilities for some standard relationships:

Relationship	k_0	k_1	k_2	θ
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.

Estimation of IBD probabilities

- Estimate IBD probabilities from the genotype data by Maximum Likelihood.
- 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**.

Maximum likelihood approach

- Let G_1 and G_2 be the pair of genotypes observed at a locus for two individuals.
- Let m ($m = 0, 1$ or 2) represent the number of alleles IBD.
- Probability of observed a pair of genotypes G_1 & G_2 , given k_0, k_1, k_2 :

$$\begin{aligned}P(G_1 \& G_2 | k_0, k_1, k_2) = & P(G_1 \& G_2 | m = 0)k_0 \\ & + P(G_1 \& G_2 | m = 1)k_1 \\ & + P(G_1 \& G_2 | m = 2)k_2\end{aligned}$$

$P(G_1 \& G_2 | m = 0)$, $P(G_1 \& G_2 | m = 1)$ and $P(G_1 \& G_2 | m = 2)$ depend on the genotypes of the individuals and are calculated from the allele frequencies in the population.

Calculating the joint genotype probabilities

- Suppose two genotypes $G_1 = i/i$ and $G_2 = i/i$, let p_i be the proportion of the allele i in the population, then:

$$P(G_1 = i/i \ \& \ G_2 = i/i | m = 0) = P(G_1 = i/i)P(G_2 = i/i) = p_i \cdot p_i \cdot p_i \cdot p_i = p_i^4$$

$$\begin{aligned} P(G_1 = i/i \ \& \ G_2 = i/i | m = 1) &= P(G_1 = i/i)P(G_2 = i/i | G_1 = i/i, m = 1) = \\ &= p_i^2 \left[p_i \frac{1}{2} + p_i \frac{1}{2} \right] = p_i^3 \end{aligned}$$

$$P(G_1 = i/i \ \& \ G_2 = i/i | m = 2) = P(G_1 = i/i) = P(G_2 = i/i) = p_i \cdot p_i = p_i^2$$

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	$4p_i p_j p_m p_l$		

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

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

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

Maximum likelihood 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

- R-package SNPrelate
- R-package GWASTools
- GRR (Abecasis Lab: <http://csg.sph.umich.edu/abecasis/GRR/>)
- RELPAIR
- PLINK (method of moments estimation)
- KING
- ...

References

-  Foulkes, A.S. (2009) *Applied statistical genetics with R*. Springer.
-  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.
-  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 exercises

The file `Yoruba10000.rda` contains 10,000 SNPs of a Yoruba population consisting of 90 parent-offspring trios (2 parents and 1 child). We wish to investigate if the genetic data is consistent with the specified relationships. The file contains two objects, `X.Fam` with pedigree information and `X.Geno` with genotype information.

- 1 Compute the mean m of the number of alleles shared for each pair of individuals.
- 2 Compute the standard deviation s of the number of alleles shared for each pair of individuals.
- 3 Plot all pairs in a scatterplot of s against m .
- 4 Do you think all relationships between all individuals were correctly specified?