INBREEDING AND RELATEDNESS

Predicted Values

Questions of Interest

If genotypic data are available, individual inbreeding and kinship values can be estimated:

- What is the Genetic Relatedness Matrix? (association mapping)
- How do social behaviors evolve?
- How should captive breeding programs be managed? (conservation genetics)
- Are these remains from a person in this family? (disaster victim identification)

Identity by Descent

The degree of dependence between a pair of alleles was described by correlation by Wright (1922) and by the probability of identity by descent (ibd) by Malécot (1948).

Two alleles are ibd if they have both descended from the same allele in a reference population. Distinct pairs of alleles in that reference population are not ibd. Therefore ibd is a relative, not an absolute, concept.

Wright S. 1922. Coefficients of inbreeding and relationship. Am Naturalist 56:330-338.

Malécot G. 1948. *The Mathematics of Heredity.* Translated by Yermanos DM (1960). Freeman, San Francisco.

Evolutionary Replication

The concept of ibd rests on descent from a reference population to the present generation, and this process is subject to genetic sampling variation. The probability of ibd for two alleles is an average over all possible evolutionary replicates of the history of those alleles from reference to present.

This means that the population sampled to provide observed genotypes is itself just one realization of an evolutionary process. The allele proportions p in that population are (evolutionary) sample values of underlying probabilities π.

Kinship vs Inbreeding

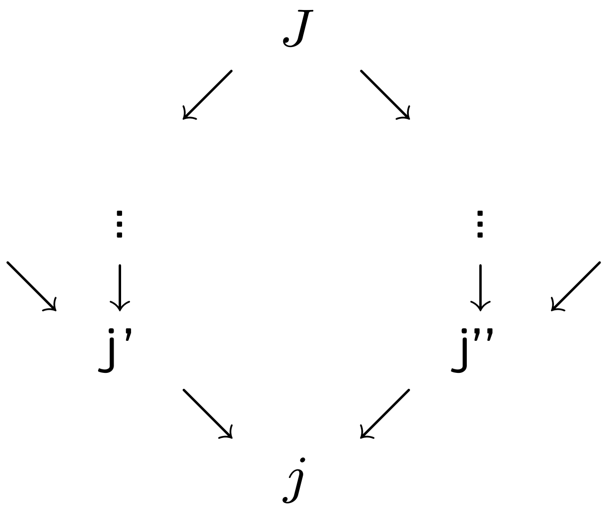
The *kinship* of individuals j, j' in a population is the probability an allele from j is ibd to an allele from j'. This is $\theta_{jj'}$.

The inbreeding of individual j in a population is the probability the two alleles in that individual are ibd. Write this as F_j.

Two alleles drawn from individual j are equally likely to be the same allele or different alleles:

$$\theta_{jj} \quad = \quad \frac{1}{2}\left(1 + F_j\right)$$

Predicted Values: Path Counting



If there are n individuals (including j', j'', J) in the path linking the parents through J, then the inbreeding F_j of j, or the kinship $\theta_{j'j''}$ of j' and j'', is

$$F_j = \theta_{j'j''} \;=\; \left(\frac{1}{2}\right)^n (1 + F_J)$$

If there are several ancestors, this expression is summed over all the ancestors.

Average Kinships

The average over all pairs of distinct individuals, $j \neq j'$, of the kinships $\theta_{jj'}$ is written as θ_S. The average of this over populations is θ_S. *These are probabilities for individuals.*

When there is random mating and Hardy-Weinberg equilibrium in a population, any pair of distinct alleles in a population (within or between individuals) is equivalent and then the average ibd probability for all these pairs is written as θ_W, where W means within populations. The average over populations is θ_W. *These are probabilities for distinct allele pairs.*

Within-population Inbreeding: F_{IS}

For a population, the inbreeding coefficient for individual j, *relative to* the identity of pairs of alleles between individuals in that population, is

$$f_j = \frac{F_j - \theta_S}{1 - \theta_S}$$

The average over individuals within this population is the population-specific f, and it compares within-individual ibd to between-individual ibd in the same population. It is the quantity being addressed by Hardy-Weinberg testing in the population.

Within-population Kinship

For a population, the kinship of individuals j, j' *relative to* the kinship for all pairs of individuals in that population is

$$\psi_{jj'} = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

and these average zero over all pairs of individuals in the population.

The average kinship for individual j is

$$\Psi_j = \frac{1}{n-1} \sum_{j' \neq j}^{n} \theta_{jj'}$$

and the average relative kinship is

$$\psi_j = \frac{1}{n-1} \sum_{j' \neq j}^{n} \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

Aside: Jacquard Coefficients

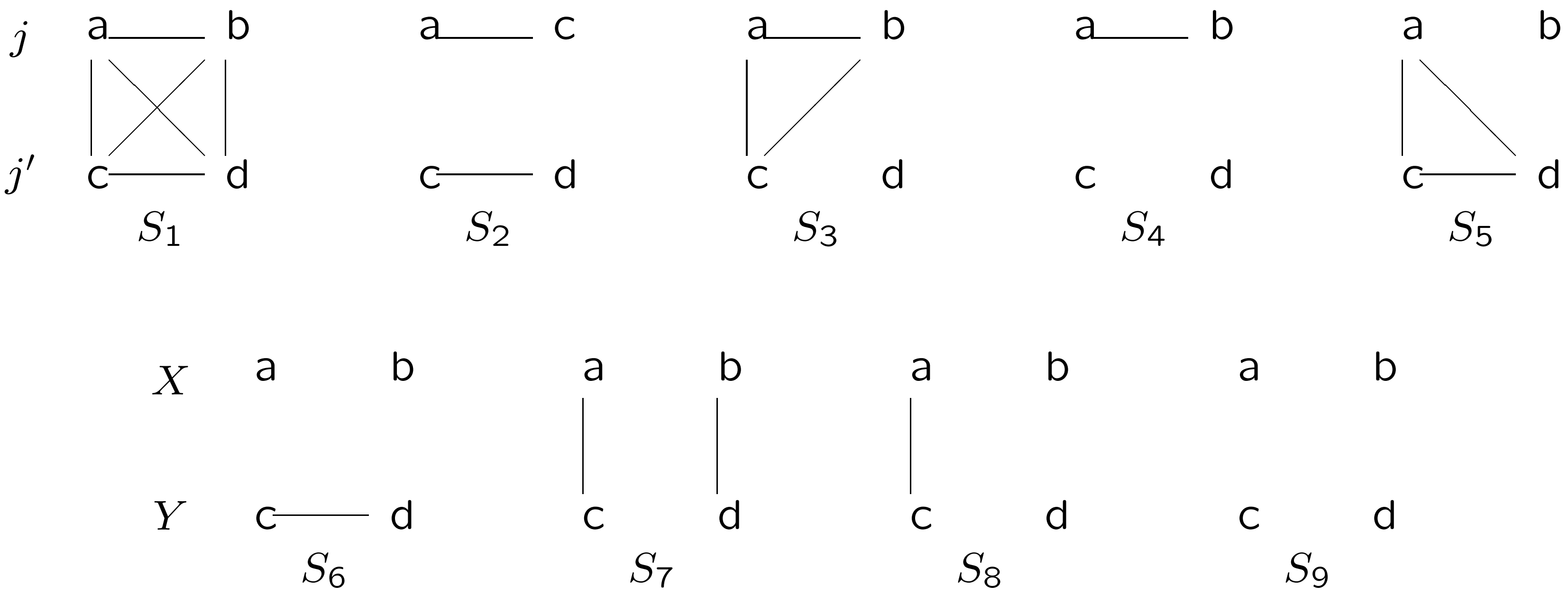
A complete description of the ibd status among four alleles a, b, c, d carried by two individuals requires 15 measures (as opposed to the two, F, $1-F$, for one individual):

Alleles ibd*	Probability	Alleles ibd*	Probability
a, b, c, d	δ_{abcd}	a, b	δ_{ab}
a, b, c	δ_{abc}	a, c	δ_{ac}
a, b, d	δ_{abd}	a, d	δ_{ad}
a, c, d	δ_{acd}	b, c	δ_{bc}
b, c, d	δ_{bcd}	b, d	δ_{bd}
a, b and c, d	$\delta_{ab.cd}$	c, d	δ_{cd}
a, c and b, d	$\delta_{ac.bd}$	none	δ_0
a, d and b, c	$\delta_{ad.bc}$		

*Alleles not listed are not ibd to those listed

Aside: Nine-parameter IBD Set

In most applications there is no need to distinguish between maternal and paternal alleles and the 15 ibd states can be collapsed into nine $\{S_i\}$, whose probabilities $\{\Delta_i\}$ are the Jacquard coefficients. Solid lines join pairs of ibd alleles: top row is the pair of alleles for j, bottom row the pair of alleles for j'.



Aside: Coancestry Coefficient

The coancestry coefficient $\theta_{jj'}$ is the probability that a random allele from $j(ab)$ is ibd to a random allele from $j'(cd)$:

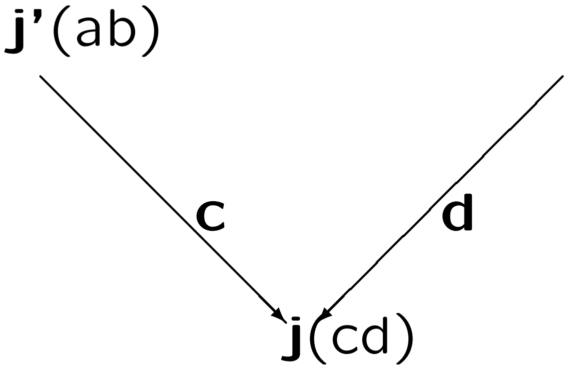
$$\begin{aligned}
\theta_{jj'} &= \frac{1}{4}[\Pr(a \equiv c) + \Pr(a \equiv d) + \Pr(b \equiv c) + \Pr(b \equiv d)] \\
&= \delta_{abcd} + \frac{1}{2}(\delta_{abc} + \delta_{abd} + \delta_{acd} + \delta_{bcd}) + \frac{1}{2}(\delta_{ac.bd} + \delta_{ad.bc}) \\
&\quad + \frac{1}{4}(\delta_{ac} + \delta_{ad} + \delta_{bc} + \delta_{bd}) \\
&= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8
\end{aligned}$$

Aside: κ-coefficients

If individuals $j(ab)$ and $j'(cd)$ are both not inbred, then $a \not\equiv b$ and $c \not\equiv d$ and the nine states $\{S_i\}$ reduce to three: S_7, S_8, S_9. Then $\Delta_7, \Delta_8, \Delta_9$ are the probabilities that they carry 0, 1, or 2 pairs of ibd alleles. For example: their two maternal alleles may be ibd or not ibd, and their two paternal alleles may be ibd or not.

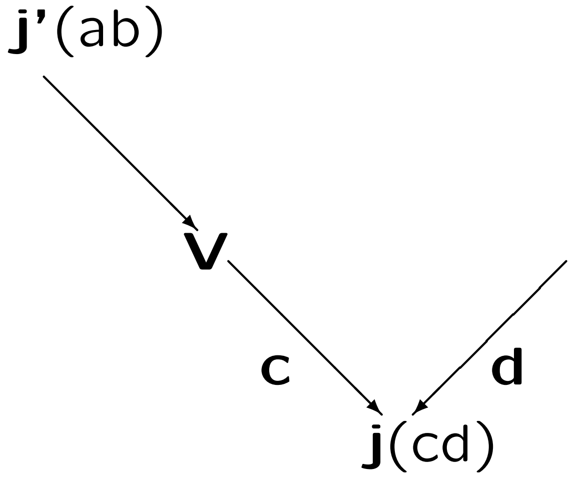
The probabilities of two individuals having 0, 1 or 2 pairs of ibd alleles are generally written as $\kappa_0, \kappa_1, \kappa_2$ and $\theta = \frac{1}{2}\kappa_2 + \frac{1}{4}\kappa_1$ for pairs of non-inbred individuals.

Aide: Parent-Child



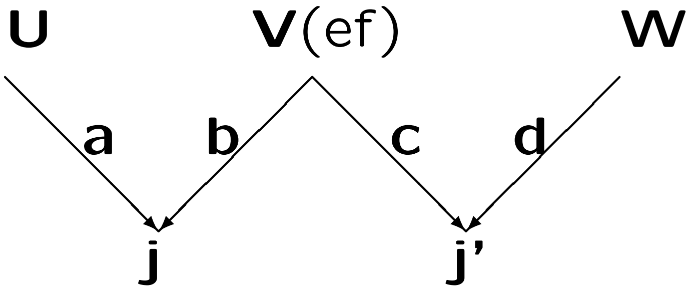
$$\text{Pr}(c \equiv a) = 0.5, \quad \text{Pr}(c \equiv b) = 0.5, \quad \kappa_1 = 1$$

Aside: Grandparent-grandchild



$$\Pr(c \equiv a) = 0.25, \quad \Pr(c \equiv b) = 0.25, \quad \kappa_1 = 0.5 \& \kappa_0 = 0.5$$

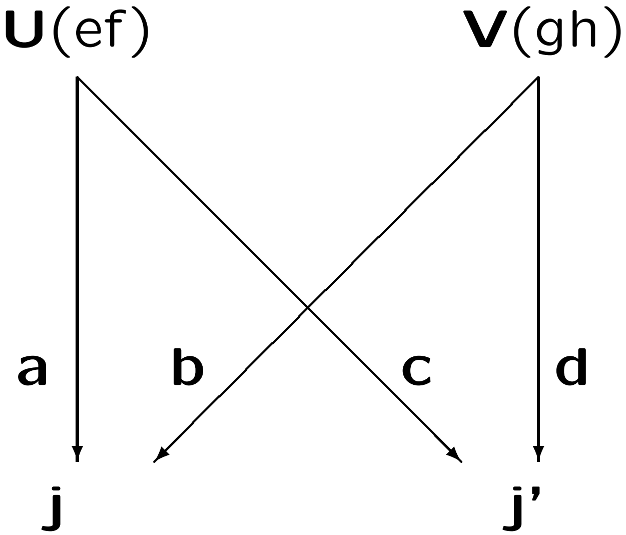
Aside: Half sibs



		0.5	0.5
		$c \equiv e$	$c \equiv f$
0.5	$b \equiv e$	0.25	0.25
0.5	$b \equiv f$	0.25	0.25

Therefore $\kappa_1 = 0.5$ so $\kappa_0 = 0.5$.

Aside: Full sibs



		0.5	0.5
		$b \equiv d$	$b \not\equiv d$
0.5	$a \equiv c$	0.25	0.25
0.5	$a \not\equiv c$	0.25	0.25

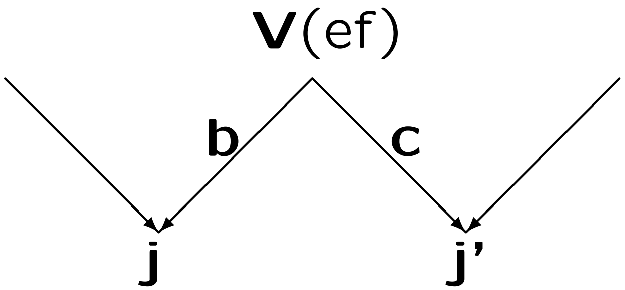
$$\kappa_0 = 0.25, \kappa_1 = 0.50, \kappa_2 = 0.25$$

Aside: Non-inbred Relatives

Relationship	κ_2	κ_1	κ_0	$\theta = \frac{1}{2}\kappa_2 + \frac{1}{4}\kappa_1$
Identical twins	1	0	0	$\frac{1}{2}$
Full sibs	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
Parent-child	0	1	0	$\frac{1}{4}$
Double first cousins	$\frac{1}{16}$	$\frac{3}{8}$	$\frac{9}{16}$	$\frac{1}{8}$
Half sibs*	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$
First cousins	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
Unrelated	0	0	1	0

* Also grandparent-grandchild and avuncular (e.g. uncle-niece).

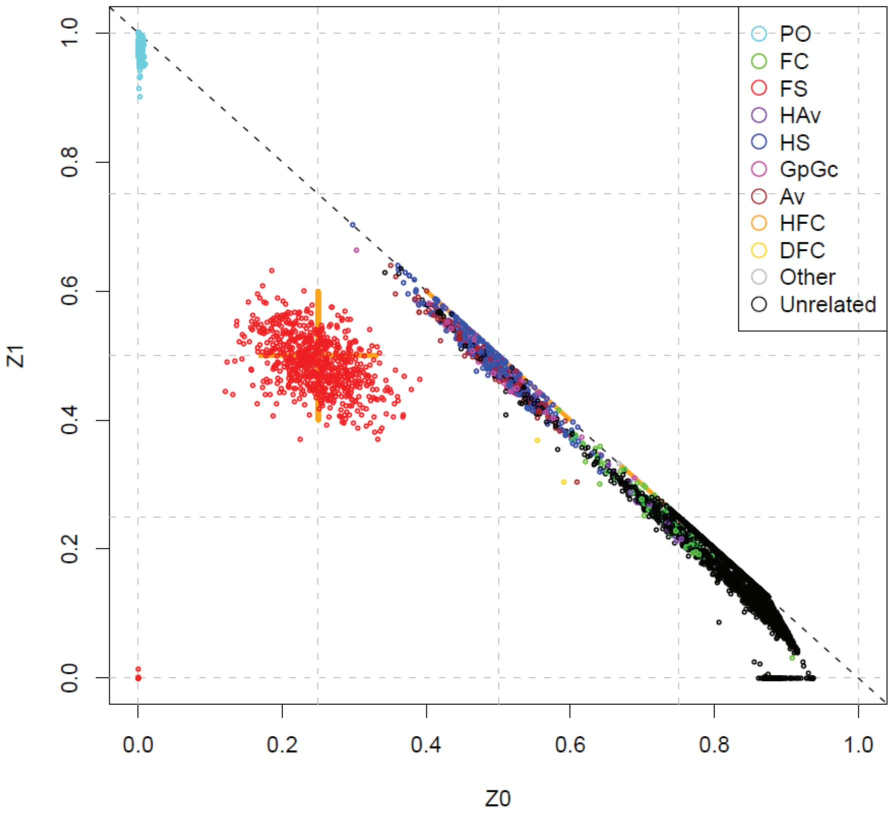
Predicted vs Actual Kinship



For half-sibs, for example, the predicted kinship, is $(1/2)^3 = 1/8$. However, alleles b, c are equally likely to be ibd or not ibd (ibd if they are both copies of e or f) so the actual kinship is either 0.25 (with probability 1/2) or 0 (with probability 1/2). The actual kinship of j, j' has an expected value (the average over evolutionary replicates of j, j') of 1/8 and a standard deviation of 1/8. Over the whole genome, the standard deviation is 0.013. The estimate from observed marker genotypes will be of the actual ("gold standard") kinship.

Hill and Weir, Genet Res 2011

Aside: PLINK Example



Shows variation of estimated κ's around predicted κ's.

Inbreeding and Kinship Estimation

Allele Sharing Approach

Write the observed allelic matching as $\tilde{A}_j$ within individual j, and as $\tilde{A}_{jj'}$ between individuals j, j'. For SNPs, these proportions are:

		$\tilde{A}_j$
	AA	1
j	Aa	0
	aa	1

$\tilde{A}_{jj'}$		j'		
		AA	Aa	aa
	AA	1	0.5	0
j	Aa	0.5	0.5	0.5
	aa	0	0.5	1

These are compared to the average matching for all pairs of individuals: $\tilde{A}_S$ for all pairs in the same sample.

Allele Sharing

The model specifies that the expectation over evolutionary replicates for a matching proportion $\tilde{A}_l$, at SNP l, is $A_l + (1 - A_l)\theta$ where θ is the ibd probability for the pair(s) of alleles being matched and A_l is a nuisance parameter:

$$A_l \;=\; \pi_l^2 + (1-\pi_l)^2 = 1 - 2\pi_l(1-\pi_l)$$

The allele-sharing estimates for inbreeding and kinship are

$$\hat{f}_j = \frac{\tilde{A}_j - \tilde{A}_S}{1 - \tilde{A}_S} \quad , \quad \hat{\psi}_{jj'} = \frac{\tilde{A}_{jj'} - \tilde{A}_S}{1 - \tilde{A}_S}$$

Combine over SNPs with as the ratio of averages

$$\hat{f}_j = \frac{\sum_l(\tilde{A}_{j_l} - \tilde{A}_{S_l})}{\sum_l(1 - \tilde{A}_{S_l})} \quad , \quad \hat{\psi}_{jj'} = \frac{\sum_l(\tilde{A}_{jj'_l} - \tilde{A}_{S_l})}{\sum_l(1 - \tilde{A}_{S_l})}$$

Ratio of Average vs Average of Ratios

As the number of SNPs increases, Ochoa and Storey (2021) showed that the ratio of averages estimates converge almost surely to the parameters

$$f_j = \frac{F_j - \theta_S}{1 - \theta_S} \quad , \quad \psi_{jj'} = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

However, convergence is not guaranteed for the average of ratios estimates

$$\frac{1}{L}\sum_l \frac{\tilde{A}_{j_l} - \tilde{A}_{S_l}}{1 - \tilde{A}_{S_l}} \quad , \quad \frac{1}{L}\sum_l \frac{\tilde{A}_{jj'_l} - \tilde{A}_{S_l}}{1 - \tilde{A}_{S_l}}$$

When the π_l are unknown, it is not possible to estimate the ibd probabilities F_j and $\theta_{jj'}$.

Ochoa A, Storey JD. 2021. PLoS Genetics 17:Article 1009241

Allele Sharing

The estimates behave well for estimating the parameters

$$f_j = \frac{F_j - \theta_S}{1 - \theta_S} \quad , \quad \psi_{jj'} = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

Individuals less inbred than the average kinship have negative f values.

The average over pairs of individuals j, j' in one population, of either the estimates $\widehat{\psi}_{jj'}$ or the parameters $\psi_{jj'}$, gives zero. Some estimates and parameters are negative and some are positive.

Goudet J, Kay T, Weir BS. 2018. Mol Ecol 27:4121-4135.

Weir BS, Goudet J. 2017. Genetics 206:2085-2103.

Zhang Q, Goudet J, Weir BS. 2021. Submitted.

Alternative Estimators: Heterozygosity

The heterozygosity indicator $\tilde{H}_{jl}$ at SNP l for individual j is 1 if the individual is heterozygous and 0 if it is homozygous. Hall et al. 2012. Genet Res and Yengo et al. 2017. PNAS gave individual-specific estimates:

$$\hat{f}_{\text{Hom}_j} = 1 - \frac{\tilde{H}_{jl}}{2\tilde{p}_l(1-\tilde{p}_l)}$$

and used weighted averages over SNPs:

$$\begin{aligned}\hat{f}_{\text{Hom}_j} &= 1 - \frac{\sum_l \tilde{H}_{jl}}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)} \\ &= 1 - \frac{H_{\text{Obs}}}{H_{\text{Exp}}}\end{aligned}$$

This estimator was called f_{PLINK} by Gazal et al. 2014. Hum Hered. Note the similarity to the MLE $\hat{f}_{\text{LH1}}$ for the within-population inbreeding coefficient f given earlier - that quantity is the average over individuals of the $\hat{f}_{\text{Hom}_j}$ quantities.

Alternative Estimators: Heterozygosity

What do the usual inbreeding estimators actually estimate under genetic sampling?

$$\mathcal{E}(\widehat{f}_{\mathsf{Hom}_j}) \;=\; 1 - \frac{1 - F_j}{(1-\theta_S) - \frac{1}{2n}(1 + F_W - 2\theta_S)}$$

For large sample sizes, this reduces to

$$\mathcal{E}(\widehat{f}_{\mathsf{Hom}_j}) \;=\; \frac{F_j - \theta_S}{1 - \theta_S} = f_j$$

Aside: Expectation of $2\tilde{p}_l(1-\tilde{p}_l)$

Expectations of allele frequencies in a sample of n individuals:

$$\begin{aligned}
\mathcal{E}(\tilde{p}_l) &= \pi_l \\
\mathcal{E}(\tilde{p}_l^2) &= \pi_l^2 + \pi_l(1-\pi_l)\left[\theta_S + \frac{1}{2n}(1+F_W-2\theta_S)\right] \\
\mathcal{E}[2\tilde{p}_l(1-\tilde{p}_l)] &= 2\pi_l(1-\pi_l]\left[(1-\theta_S) - \frac{1}{2n}(1+F_W-2\theta_S)\right] \\
&\approx 2\pi_l(1-\pi_l](1-\theta_S) \\
&= \mathcal{E}(1-\tilde{A}_S)
\end{aligned}$$

It is not the case that $2\tilde{p}_l(1-\tilde{p}_l)$ is an unbiased estimator for $2\pi_l(1-\pi_l)$, even if the sample size is large.

Alternative Estimators: GCTA

If X_{jl}, the allele dosage, is the number of copies of the reference allele for SNP l carried by individual j, Yengo et al. used

$$\hat{f}_{\mathrm{Uni}_j} \;=\; \frac{1}{L}\sum_{l=1}^{L}\left(\frac{X_{jl}^2-(1+2\tilde{p}_l)X_{jl}+2\tilde{p}_l^2}{\tilde{p}_l(1-\tilde{p}_l)}\right)$$

For large samples the ratio of averages version of this has an expected value under genetic sampling of

$$\mathcal{E}(\hat{f}_{\mathrm{Uni}_j}) \;=\; \frac{F_j-2\Psi_j+\theta_S}{1-\theta_S}=f_j-2\psi_j$$

where Ψ_j,ψ_j are the average coancestries or kinships of individual j with other members of the study sample,

$$\Psi_j=\frac{1}{n-1}\sum_{\substack{j'=1\\ j\neq j'}}^{n}\theta_{jj'}\quad,\quad \psi_j=\frac{1}{n-1}\sum_{\substack{j'=1\\ j\neq j'}}^{n}\psi_{jj'}$$

The average over individuals of $\hat{f}_{\mathrm{Uni}_j}$ is f_{LH5} described earlier.

Alternative Estimators: GCTA

The inclusion of the ψ term means that the ranking of $\widehat{F}_{\mathrm{Uni}_j}$ expected values can be different from the ranking of F_j values. The rankings of $\widehat{f}_{\mathrm{Hom}_j}$ expected values are the same as those for F_j.

Yang et al. also discussed

$$\widehat{f}_{\mathrm{GCTA}_j} \;=\; \frac{1}{L}\sum_{l=1}^{L}\frac{(X_{jl}-2\tilde{p}_l)^2}{2\tilde{p}_l(1-\tilde{p}_l)}-1$$

For large samples, the ratio of averages versions of these estimates have expected values

$$\mathcal{E}(\widehat{f}_{\mathrm{GCTA}_j}) \;=\; \frac{F_j-4\psi_j+3\theta_S}{1-\theta_S}=f_j-4\psi_j$$

Alternative Estimators: MLE

Hall et al. used EM to give MLEs for f_j, assuming π_l's were known (and equal to $\tilde{p}_l$), using

$$\begin{aligned}\Pr(\tilde{H}_{jl}=1) &= 2\tilde{p}_l(1-\tilde{p}_l)(1-f_j)\\ \Pr(\tilde{H}_{jl}=0) &= 1-2\tilde{p}_l(1-\tilde{p}_l)(1-f_j)\end{aligned}$$

but it is easier to use a grid search to maximize the likelihood $\mathrm{Lik}(f_j)$, or its logarithm:

$$\mathrm{Lik}(f_j)] = \prod_l [1-2\tilde{p}_l(1-\tilde{p}_l)(1-f_j)]^{1-\tilde{H}_{jl}}[2\tilde{p}_l(1-\tilde{p}_l)(1-f_j)]^{\tilde{H}_{jl}}$$

These estimates are close in value to $\widehat{f}_{\mathrm{Hom}_j}$.

Alternative Estimators: Runs of Homozygosity

Estimators so far use single SNP statistics and average over SNPs.

Runs of homozygosity, with a large number of SNPs, are likely to represent regions of identity by descent. The inbreeding coefficient can be estimated as the proportion of windows of SNPs that are completely homozygous.

Requires judgment in deciding window length, degree of window overlap, allowance for some heterozygotes, and (possibly) minor allele frequency. The quantity being estimated depends on these values.

McQuillan et al. 2006. Am J Hum Genet; Joshi et al. 2015. Nature

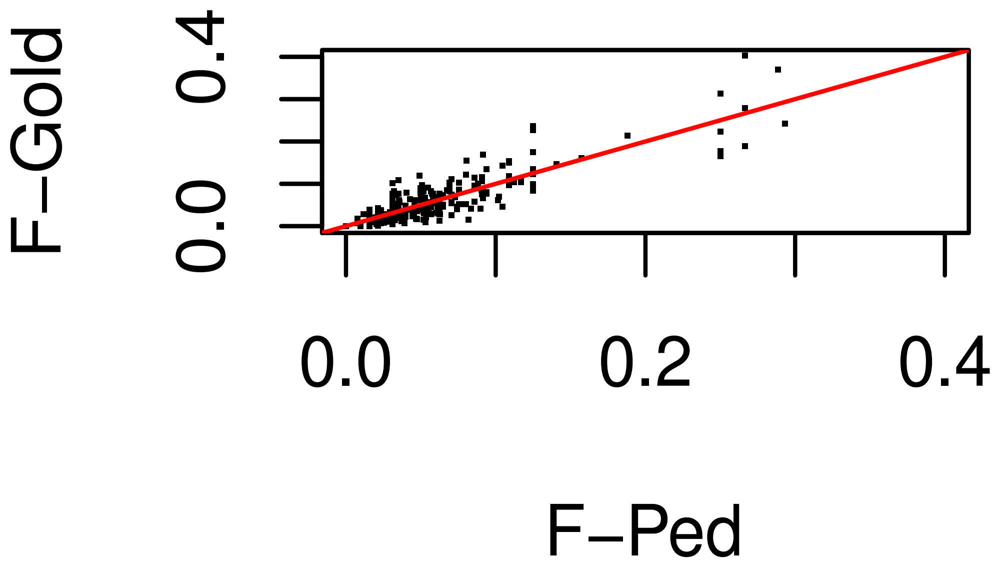
Comparison of Estimators: Simulations

Simulation of 50 founder individuals, with 100,000 SNPs over a 20 Morgan map.

Software quantiNemo software Neuenschwander et al. 2008. Bioinformatics to generate eight subsequent generations of 50 individuals per generation and it is these 400 descendants that were used for subsequent analysis.

The mating system was 80% monogamous and 20% random mating. Each of the 100 alleles per SNP among the founders was given a unique identifier so that subsequent identity by descent could be tracked. The average ibd proportion over loci, within individuals and between each pair of individuals, provided "gold standard" or actual inbreeding and kinship coefficients, as opposed to the pedigree-based values from path counting.

Simulated Pedigree vs Actual Inbreeding



Comparison of Estimators: Notation

Fped, fped: pedigree values of F and f.

Fgold, fgold: actual values of F and f.

Froh: runs of homozgosity estimate.

fMLE: maximum likelihood estimate of F.

fHom: $1 - \tilde{H}/2\tilde{p}(1 - \tilde{p})$

f: allele-matching estimates of f,

fUni: f_{Uni} estimate.

fGcta: f_{Gcta} estimate..

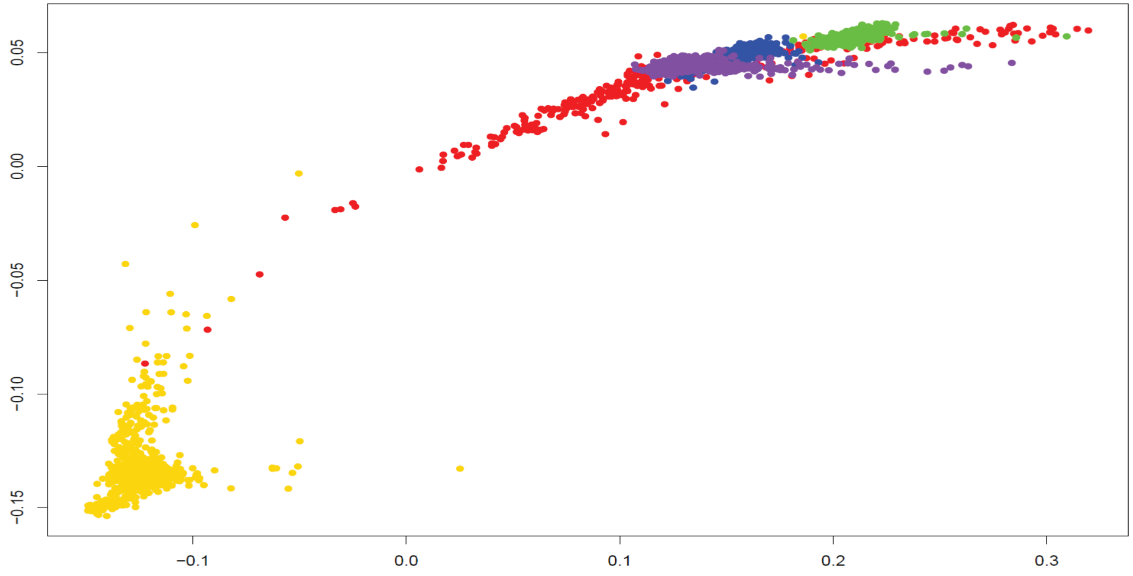
Comparison of Estimators: Correlations

```
       Fped  fped Fgold fgold  Froh  fMLE  fHom     f  Ugold  fGcta
Fped  1.000 1.000 0.902 0.901 0.879 0.790 0.836 0.836  0.707  0.642
fped  1.000 1.000 0.902 0.902 0.879 0.790 0.836 0.836  0.707  0.642
Fgold 0.902 0.902 1.000 1.000 0.975 0.889 0.918 0.918  0.829  0.743
fgold 0.901 0.902 1.000 1.000 0.975 0.889 0.918 0.918  0.829  0.743
Froh  0.879 0.879 0.975 0.975 1.000 0.929 0.952 0.952  0.819  0.779
fMLE  0.790 0.790 0.889 0.889 0.929 1.000 0.976 0.976  0.838  0.876
fHom  0.836 0.836 0.918 0.918 0.952 0.976 1.000 1.000  0.747  0.781
f     0.836 0.836 0.918 0.918 0.952 0.976 1.000 1.000  0.747  0.781
Ugold 0.707 0.707 0.829 0.829 0.819 0.838 0.747 0.747  1.000  0.917
fGcta 0.642 0.642 0.743 0.743 0.779 0.876 0.781 0.781  0.917  1.000
```

1000 Genomes Data

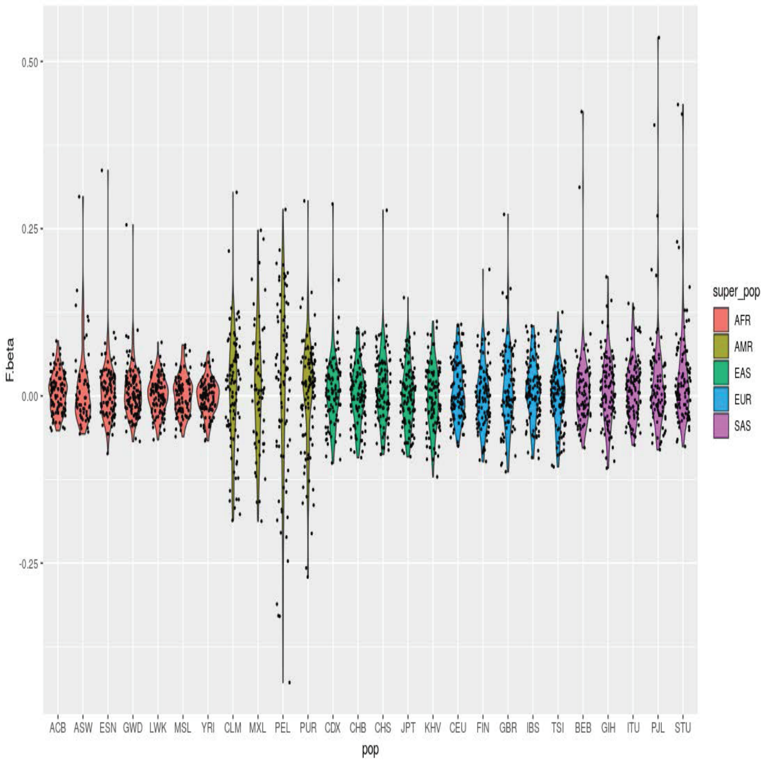
	AFR	661	AFRICAN
1	ACB	96	African Caribbeans in Barbados
2	ASW	61	Americans of African Ancestry in SW USA
3	ESN	99	Esan in Nigeria
4	GWD	113	Gambian in Western Divisions in the Gambia
5	LWK	99	Luhya in Webuye, Kenya
6	MSL	85	Mende in Sierra Leone
7	YRI	108	Yoruba in Ibadan, Nigeria
	AMR	347	ADMIXED AMERICAN
8	CLM	94	Colombians from Medellin, Colombia
9	MXL	64	Mexican Ancestry from Los Angeles USA
10	PEL	85	Peruvians from Lima, Peru
11	PUR	104	Puerto Rican from Puerto Rico
	EAS	504	EAST ASIAN
12	CDX	93	Chinese Dai in Xishuangbanna, China
13	CHB	103	Han Chinese in Beijing, China
14	CHS	105	Southern Han Chinese
15	JPT	104	Japanese in Tokyo, Japan
16	KHV	99	Kinh in Ho Chi Minh City, Vietnam
	EUR	503	EUROPEAN
17	CEU	99	Utah Residents (CEPH) with Northern and Western European Ancestry
18	FIN	99	Finnish in Finland
19	GBR	91	British in England and Scotland
20	IBS	107	Iberian Population in Spain
21	TSI	107	Toscani in Italia
	SAS	489	SOUTH ASIAN
22	BEB	86	Bengali from Bangladesh
23	GIH	103	Gujarati Indian from Houston, Texas
24	ITU	102	Indian Telugu from the UK
25	PJL	96	Punjabi from Lahore, Pakistan
26	STU	102	Sri Lankan Tamil from the UK

ψ vs f in 1000 Genomes Data

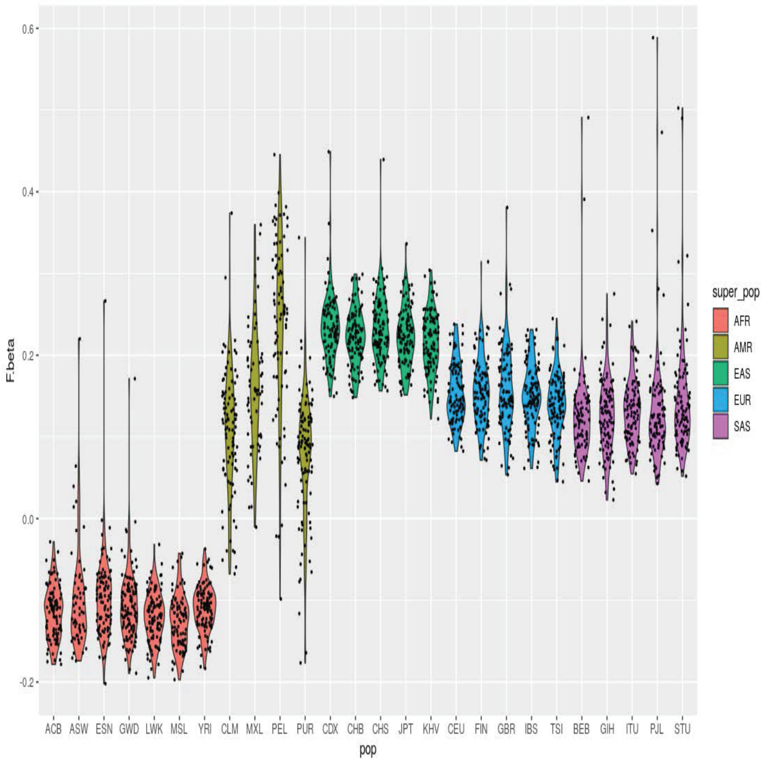


Estimates $\widehat{\psi}_j$ of within-population individual-specific average kinships (Y-axis) vs estimates $\widehat{f}_j$ of within-population individual-specific inbreeding coefficients (X-axis) for 1000 Genomes data, with the World as reference set. Gold: AFR; Red: AMR; Purple: SAS; Blue: EUR; Green: EAS.

Inbreeding is Relative: Not Absolute



Local Population Reference



Whole World Reference

Chromosome 22 data from 1000 Genomes.

Continents (left to right): AFR, AMR, EAS, EUR, SAS

Estimation of Kinship

A general allele sharing estimator for the kinship of individuals j, j' in the same sample:

$$\hat{\psi}_{jj'} = \frac{\tilde{A}_{jj'} - \tilde{A}_R}{1 - \tilde{A}_R}$$

Here $\tilde{A}_{jj'}$ is the allele sharing for the target pair of individuals, and $\tilde{A}_R$ is for a reference set.

- if R is all pairs of individuals in the same sample, $\tilde{A}_R$ is the average sharing over jj' pairs, and the estimates have an average of zero.

Estimation of Kinship

- if R is a set of populations, say in the continent to which the target pair of individuals belong, $\tilde{A}_R$ is the average sharing for all pairs of alleles, one from each of two populations in this same set of populations. (Continental Reference)

- if R is all populations for which data are available, $\tilde{A}_R$ is the average sharing for all pairs of alleles, one from each of any two of these populations. (World Reference)

The averages of these two sets of estimates over all pairs of individuals in one population can be positive or negative.

Kinship is relative, not absolute

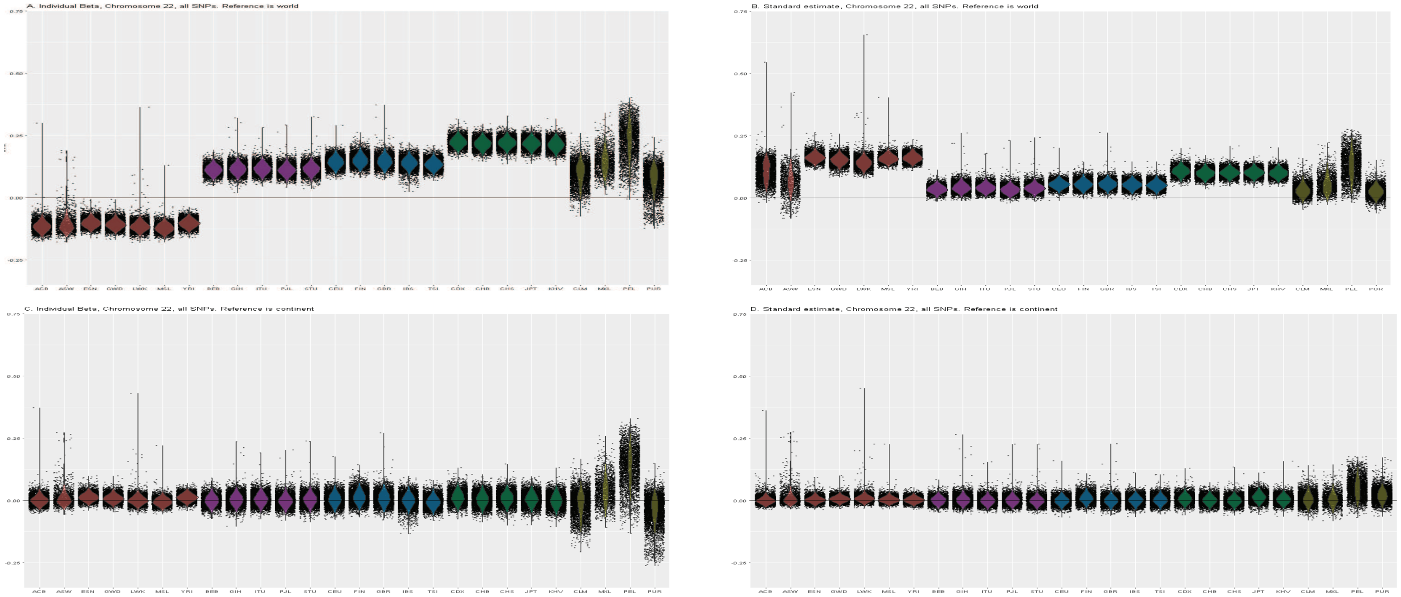
The ψ kinship estimates have been applied to 1000 Genomes data, and compared to standard estimates, shown on next slide.

For the whole world, all 26 populations, as reference the ψ estimates show a relatively narrow range of values within each African population (AFR) and lower African values than in the rest of the world, as expected from our understanding of higher genetic diversity within African than non-African populations from the migration history of modern humans. This pattern was not shown by the GCTA estimates - those estimates showed higher kinship among African individuals than among non-Africans.

The wide plots for the Admixed American populations (AMR) reflect the admixture within those populations, with greater relatedness reflecting more ancestral commonality. When each continental group is used as a reference, all populations show low kinship, except for the admixed AMR.

Kinship is relative, not absolute

Top row: Whole world reference. Bottom row: Continental group reference.



A. Individual Beta, Chromosome 22, all SNPs. Reference is world

B. Standard estimate, Chromosome 22, all SNPs. Reference is world

C. Individual Beta, Chromosome 22, all SNPs. Reference is continent

D. Standard estimate, Chromosome 22, all SNPs. Reference is continent

Allele sharing estimates

GCTA estimates

Chromosome 22 data from 1000 Genomes.

Continents (left to right): AFR, SAS. EUR, EAS, AMR

Populations (l to r):**AFR**: ACB, ASW, ESN, GWD, LWK, MSL, YRI;
SAS: BEB, GIH, ITU, PJL, STU; **EUR**: CEU, FIN, GBR, IBS, TSI;
EAS: CDX, CHB, CHS, JPT; **AMR**: KHV, CLM, MXL, PEL, PUR

Aside: κ-estimates

Given the observed inbreeding levels in the 1000 Genomes data, it is not clear that assuming non-inbred individuals and estimating κ coefficients is appropriate. However, they are often estimated.

It is usual, as in the PLINK and KING approaches, for example, to characterize pairs of individuals by the number N_i of loci at which they carry i pairs of ibs alleles. For L SNPs, the expected values of these counts are written most simply in terms of $H = \sum_{l=1}^{L} 2\pi_l(1-\pi_l)$ and $K = \sum_{l=1}^{L} 2\pi_l^2(1-\pi_l)^2$ where π_l is the probability for one of the alleles at SNP l:

$$\begin{aligned}
\mathcal{E}(N_2) &= \kappa_0(L - 2H + 3K) + \kappa_1(L - H) + \kappa_2 L \\
\mathcal{E}(N_1) &= \kappa_0(2H - 4K) + \kappa_1 H \\
\mathcal{E}(N_0) &= \kappa_0 K
\end{aligned}$$

It is usual to replace the expected counts by their observed values and replace H, K by $\tilde{H} = \sum_{l=1}^{L} 2\tilde{p}_l(1-\tilde{p}_l), \tilde{K} = \sum_{l=1}^{L} 2\tilde{p}_l^2(1-\tilde{p}_l)^2$ to obtain moment estimates of the κ_i's by rearranging these equations.

Aside: κ-estimates

It is the use of sample allele frequencies $\tilde{p}$ that causes problems: even for large sample sizes $\mathcal{E}(\tilde{H}) \neq H$.

Both PLINK and KING estimate coancestry as $\hat{\kappa}_2/2+\hat{\kappa}_1/4$, which is $[0.5-(4N_0+N_1)/(4\tilde{H})]$, but this has an expected value of $(\theta-\theta_S/2)/(1-\theta_S)$ which is neither θ for the target pair of individuals nor the within-population quantity ψ for that pair.

An allele-sharing estimate of the within-population relative value of κ_0 is

$$\hat{\kappa}_{\mathrm{AS0}_{jj'}} = \frac{N_{0_{jj'}} - N_{0_S}}{1 - N_{0_S}} \quad , \quad \mathcal{E}(\hat{k}_{\mathrm{AS0}_{jj'}}) = \frac{\kappa_{0_{jj'}} - \kappa_{0_S}}{1 - \kappa_{0_S}}$$

where N_{0_S} is the average over pairs of individuals in the study of the numbers of loci at which each pair share zero pairs of alleles ibs, and κ_{0_S} is the average of the κ_0 values for each pair of individuals.