INBREEDING AND RELATEDNESS

Bruce Weir
and
Jérôme Goudet

July 18, 2023

Predicted Values

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.

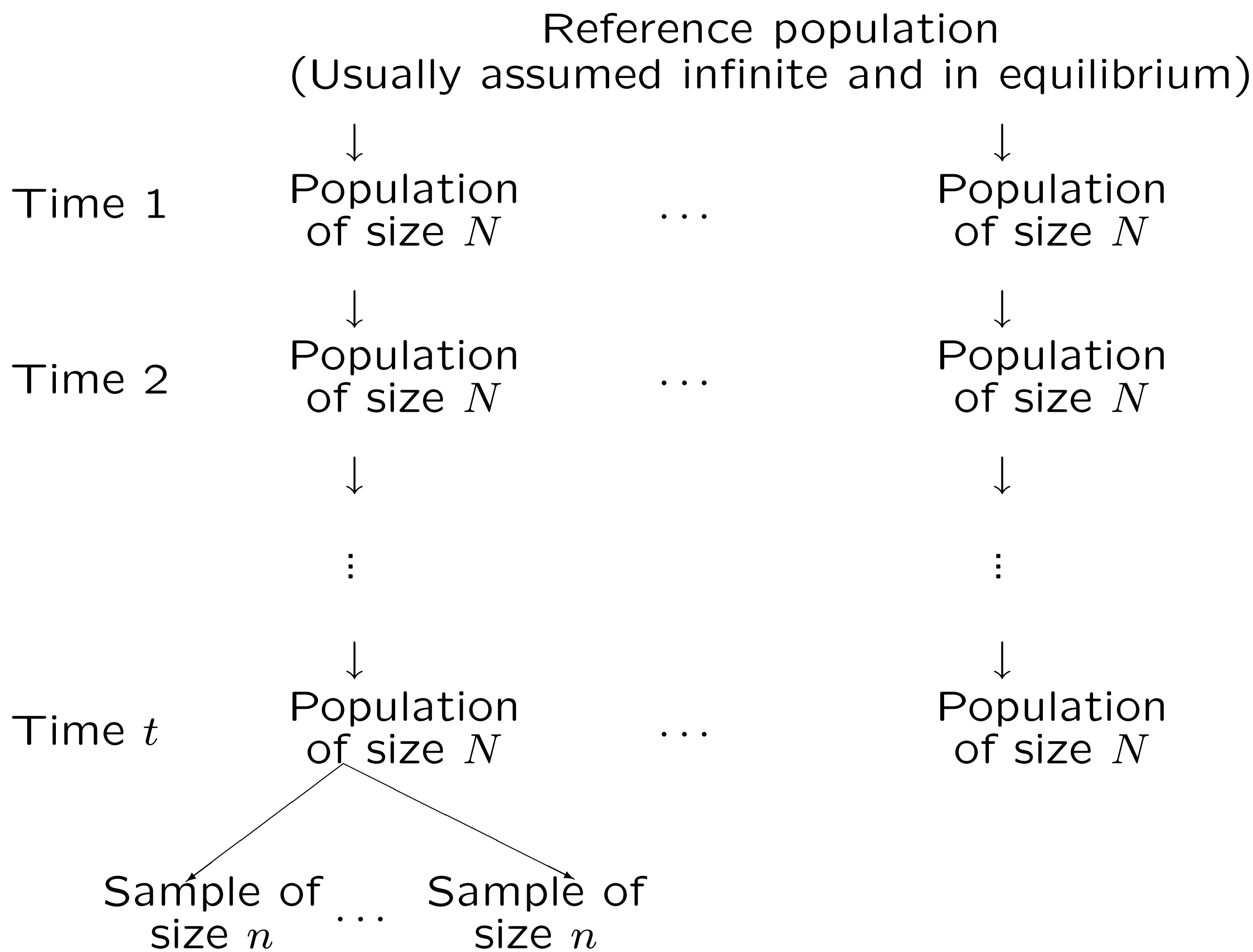
Malécot G. 1948. *The Mathematics of Heredity.* Translated by Yermanos DM (1960). Freeman, San Francisco.

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.

Classical Model



Kinship vs Inbreeding

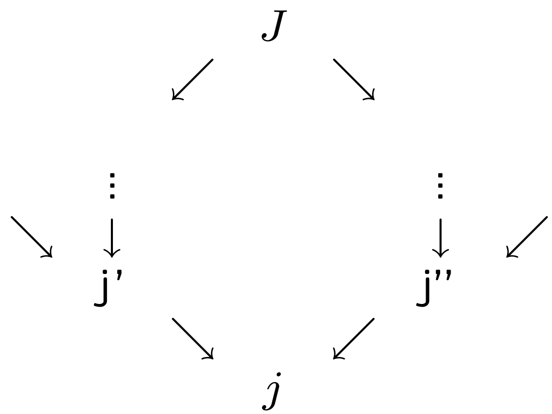
The *coancestry* of individuals j, j' in a population is the probability an allele from j is ibd to an allele from j'. This is written as $\theta_{jj'}$.

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:

$$\theta_{jj} = \frac{1}{2}\Big(1 + F_j\Big)$$

Predicted Values: Path Counting

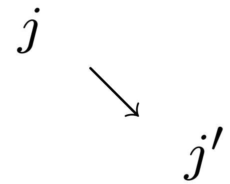


If there are n individuals (including j', j'', J) in the path linking the parents j', j'' through J, then the inbreeding F_j of j, or the coancestry $\theta_{j'j''}$ of j' and j'', is

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

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

Parent-Child



The common ancestor of parent j and child j' is j. The path linking j, j' to their common ancestor is jj' and this has $n = 2$ individuals. Therefore

$$\theta_{jj'} \;=\; \left(\frac{1}{2}\right)^2 = \frac{1}{4}$$

Full sibs

$$
\begin{array}{ccccc}
J & & & & J' \\
\downarrow & \searrow & & \swarrow & \downarrow \\
 & & & & \\
\downarrow & \swarrow & & \searrow & \downarrow \\
j & & & & j'
\end{array}
$$

The common ancestors of full sibs j and j' are J and $J;'$. The paths linking j, j' to their common ancestors are jJj' and $jJ'j'$ and these each have $n = 3$ individuals. Therefore

$$\theta_{jj'} \;=\; \left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^3 = \frac{1}{4}$$

Average Coancestries

The average over all pairs of distinct individuals, $j \neq j'$, of the coancestries $\theta_{jj'}$ is written as θ_S.

When information on individual genotypes is not available, the probability that any pair of distinct alleles (within one individual or in wto individuals) are ibd is θ_D. For a sample of size n, if the average inbreeding coefficient is $F_I = \sum_{j=1}^{n} F_j/n$:

$$\theta_D \;=\; \frac{1}{2n-1}F_I + \frac{2n-2}{2n-1}\theta_S$$

When two alleles are sampled completely independently (includes the same allele twice, different alleles within one individual and in different individuals) the average ibd probability is θ_W. For a sample of size n:

$$\theta_W \;=\; \frac{1}{2n}(1+F_I) + \frac{n-1}{n}\theta_S$$

When there is Hardy-Weinberg equilibrium, $\theta_S = F_I$ so then $\theta_D = \theta_S$ and $\theta_W = [1 + (2n-1)\theta_S]/(2n)$.

Within-population Inbreeding

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 $f = (F - \theta_S)/(1 - \theta_S)$, and it is the quantity being addressed by hardy-Weinberg etsting for that population. Averaging thus over populations gives what Wright called F_{IS}.

Within-population Kinship

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

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

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

The average coancestry for individual j is

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

and the average relative kinship is

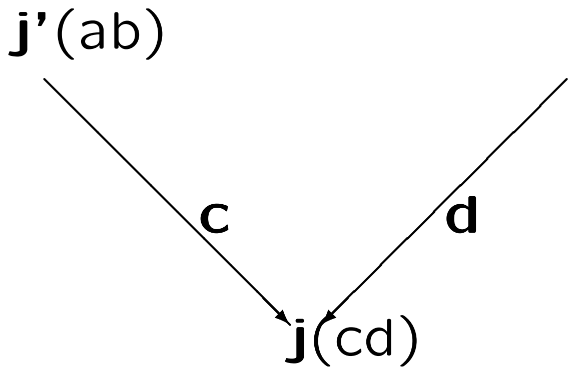
$$\psi_j = \frac{1}{n-1} \sum_{j' \neq j}^{n} \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

κ-coefficients

If individuals j and j' are both not inbred, their two maternal alleles may be ibd or not ibd, and their two paternal alleles may be ibd or not but their maternal and paternal alleles are not ibd.

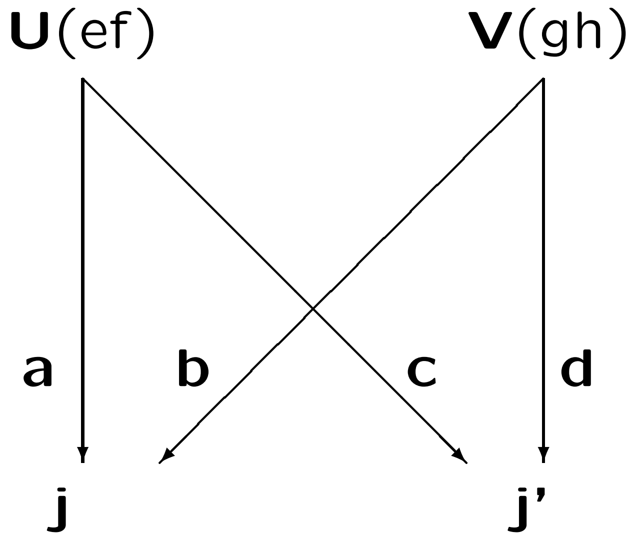
The probabilities of two individuals having 0, 1 or 2 pairs of ibd alleles are generally written as $\kappa_0, \kappa_1, \kappa_2$ and $\theta = \frac{1}{2}\kappa_2 + \frac{1}{4}\kappa_1$ for pairs of non-inbred individuals.

Parent-Child



$$\Pr(c \equiv a) = 0.5, \quad \Pr(c \equiv b) = 0.5, \quad \kappa_1 = 1$$
$$\kappa_0 = 0 \quad , \quad \kappa_2 = 0$$

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

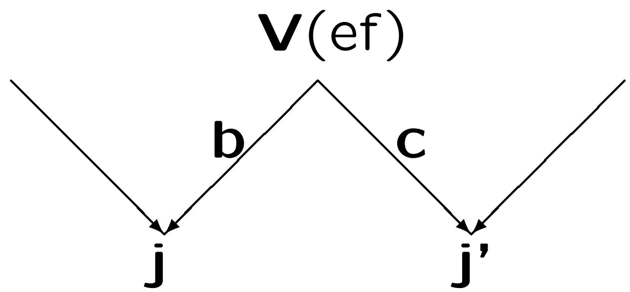
$$\kappa_0 = 0.25, \kappa_1 = 0.50, \kappa_2 = 0.25$$

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 coancestry is either 0.25 (with probability 1/2) or 0 (with probability 1/2). The actual coancestry 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") coancestry.

Hill and Weir, Genet Res 2011

Numerical Variation in Actual Kinship

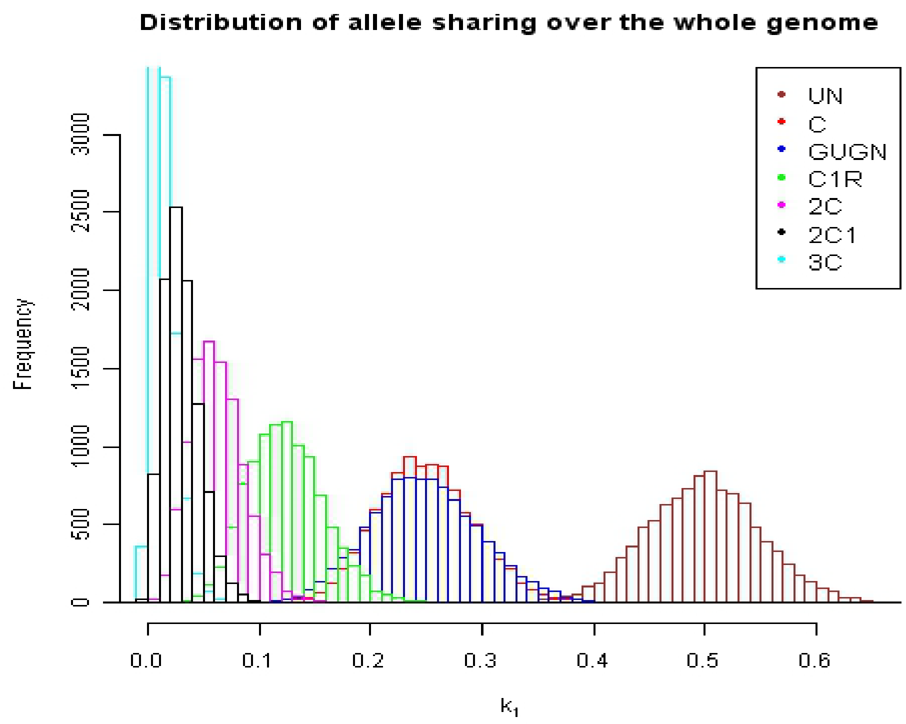


Figure 5 of Hill and Weir, 2011.

Empirical Variation in Actual Kinship

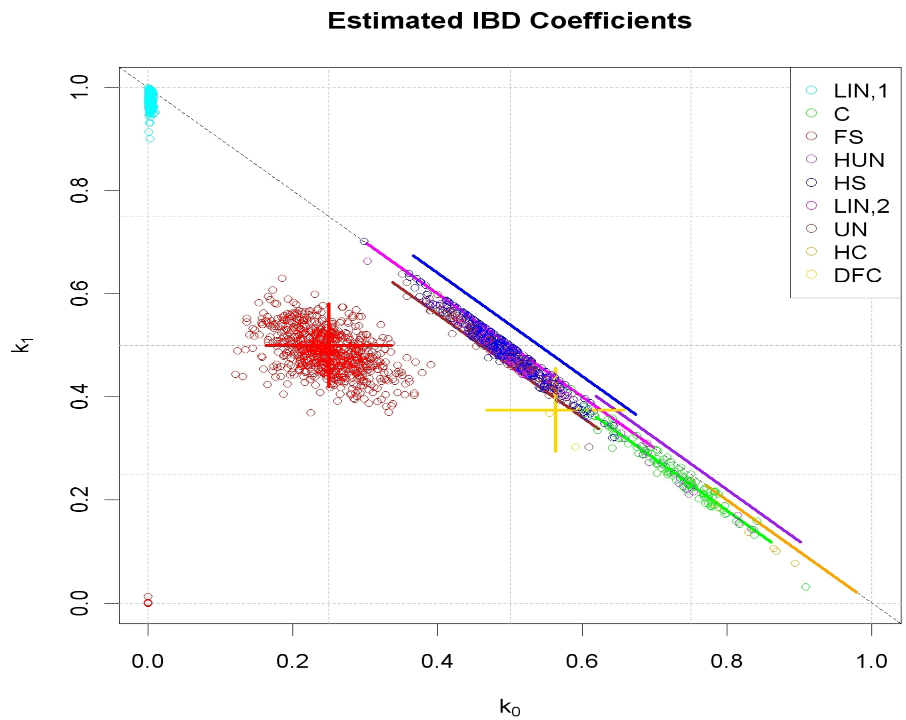


Figure 6 of Hill and Weir, 2011.

The Problem

We can predict various inbreeding and relatedness parameters if we have the pedigree of the individuals. Actual degree of inbreeding or relatedness can differ from predicted value.

How can we use genetic profiles to estimate the actual relatedness status?

Key Result

Two alleles are ibd if they have descended from the same allele in a reference population.

If θ is the probability two alleles are ibd, then the probability the alleles are both of type A is

$$\Pr(AA) = \theta\pi_A + (1-\theta)\pi_A^2$$

where π_A is the probability an allele is of type A. Suggests a translation of ibd state to an observable state.

Problem: the reference population is not observable and π_A is unknown.

Inbreeding Coefficient

If the two alleles are those for individual j, the ibd probability is F_j. From the previous slide

$$\Pr(Aa)_j \;=\; 2\pi_A(1-\pi_A)(1-F_j)$$

Define $\tilde{H}_j$ by $\tilde{H}_j = 1$ for Aa and $\tilde{H}_j = 0$ for AA or aa:

$$\mathcal{E}(\tilde{H}_j) \;=\; 2\pi_A(1-\pi_A)(1-F_j)$$

This relation suggests a moment estimator of F_j in terms of sample allele frequencies $\tilde{p}_A$:

$$\widehat{F}_j = 1 - \frac{\tilde{H}_j}{2\tilde{p}_A(1-\tilde{p}_A)}$$

This was given by Li and Horvitz (1953).

Problem with Simple Estimator

The simple estimator is sometimes written as

$$1 - \widehat{F}_j = \frac{H_{\text{Obs}}}{H_{\text{Exp}}}$$

which uses observed and 'expected' heterozygosities.

The problem is with the expected value:

$$\mathcal{E}[2\tilde{p}_A(1-\tilde{p}_A)] \;=\; 2\pi_A(1-\pi_A)\left[(1-\theta_S) - \frac{1}{2n}(1+F_I-2\theta_S)\right]$$

where F_I is the average inbreeding coefficient of n individuals in the sample providing $\tilde{p}_A$, and θ_S is the average coancestry coefficient for all pairs of individuals in that sample.

Aside: Derivation of Expected Heterozygosity

The sample frequency for allele A is the average of allelic indicators x_{jk} for allele $k, k = 1, 2$ in individual $j, j = 1, 2, \ldots n$. The indicators equal 1 for alleles of type A and 0 otherwise. They have expectations

$$\begin{aligned}
\mathcal{E}(x_{jk}) &= \pi_A \\
\mathcal{E}(x_{jk}x_{j'k'}) &= \begin{cases} \pi_A & j = j', k = k' \\ F_j\pi_A + (1 - F_j)\pi_A^2 & j = j', k \neq k' \\ \theta_{jj'}\pi_A + (1 - \theta_{jj'})\pi_A^2 & j \neq j' \end{cases}
\end{aligned}$$

The sample allele frequency, its mean and variance follow from

$$\begin{aligned}
\tilde{p}_A &= \frac{1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{2} x_{jk} \\
\mathcal{E}(\tilde{p}_A) &= \pi_A
\end{aligned}$$

Aside: Derivation of Expected Heterozygosity

$$\tilde{p}_A^2 = \frac{1}{4n^2}\left(\sum_{j=1}^{n}\sum_{k=1}^{2} x_{jk}^2 + \sum_{j=1}^{n}\sum_{k=1}^{2}\sum_{\substack{k'=1\\k\neq k'}}^{2} x_{jk}x_{jk'} + \sum_{j=1}^{n}\sum_{\substack{j'=1\\j\neq j'}}^{n}\sum_{k=1}^{2}\sum_{k'=1}^{2} x_{jk}x_{j'k'}\right)$$

$$\mathcal{E}(\tilde{p}_A^2) = \frac{1}{4n^2}\left\{2n\pi_A + 2\sum_{j=1}^{n}[F_j\pi_A + (1-F_j)\pi_A^2] + 4\sum_{j=1}^{n}\sum_{\substack{j'=1\\j\neq j'}}^{n}[\theta_{jj'}\pi_A + (1-\theta_{jj'})\pi_A^2]\right\}$$

$$= \frac{1}{4n^2}\left\{2n\pi_A + 2n[\pi_A^2 + \pi_A(1-\pi_A)F_I] + 4n(n-1)[\pi_A^2 + \pi_A(1-\pi_A)\theta_S]\right\}$$

$$= \pi_A^2 + \pi_A(1-\pi_A)\left[\theta_S + \frac{1}{2n}(1+F_I-2\theta_S)\right]$$

$$\mathcal{E}[\tilde{p}_A(1-\tilde{p}_A)] = \pi_A(1-\pi_A)\left[(1-\theta_S) + \frac{1}{2n}(1+F_I-2\theta_S)\right]$$

Coancestry Coefficient

The ibd probability for a random allele from j and one from j' is the coancestry coefficient $\theta_{jj'}$. If these two alleles are different, estimation could proceed as for the inbreeding coefficient, but with the same issue of having to estimate $2\pi_A(1-\pi_A)$.

Define $\tilde{H}_{jj'}$ for the proportion of pairs of alleles, one from j and one from j' that are different (between-individual "heterozygosity").

$$\mathcal{E}(\tilde{H}_{jj'}) = 2\pi_A(1-\pi_A)(1-\theta_{jj'})$$

Averaging over all pairs of individuals, $j \neq j'$:

$$\mathcal{E}(\tilde{H}_S) = 2\pi_A(1-\pi_A)(1-\theta_S)$$

Within-population Inbreeding and Coancestry

Estimates of inbreeding and coancestry relative to average coancestry are

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

$$\mathcal{E}(\hat{f}_j) = \frac{F_j - \theta_S}{1 - \theta_S} \quad , \quad \mathcal{E}(\hat{\psi}_{jj'}) = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

$$\mathcal{E}(1 - \hat{f}_j) = \frac{1 - F_j}{1 - \theta_S} \quad , \quad \mathcal{E}(1 - \hat{\psi}_{jj'}) = \frac{1 - \theta_{jj'}}{1 - \theta_S}$$

The unknown π's did not have to be estimated: sample allele frequencies $\tilde{p}_A$ not used. In practice, sum numerators and denominators over loci.

Note that f_j is individual-specific value of Wright's F_{IS}, and $\psi_{jj'}$ is its analog for two individuals. With data from only one population, $F_I = F_{IT}$ and $\theta_S = F_{ST}$ are not estimable.

Multiple SNPs

Single-SNP estimates for one individual would not be useful: the $\hat{f}_j$ values are 1 for homozygotes and negative for heterozygotes. Averaging over individuals would reflect the proportion of SNPs that are homozygous, but would still have high variances.

Averaging over L SNPs $l, l = 1, 2 \ldots L$, could be with an average of ratios:

$$\hat{f}_j = 1 - \frac{1}{L}\sum_{l=1}^{n} \frac{\tilde{H}_{jl}}{2\tilde{p}_l(1-\tilde{p}_l)}$$

but this is unstable because the denominator can be zero or close to zero.

Using the ratio of averages gives an unbiased estimator for a large number of SNPs (Ochoa and Storey, 2021):

$$\hat{f}_j = 1 - \frac{\sum_{l=1}^{L} \tilde{H}_{jl}}{\sum_{l=1}^{L}[2\tilde{p}_l(1-\tilde{p}_l)]}$$

Allele sharing Estimators

The inbreeding and kinship estimators $\hat{f}$ and $\hat{\psi}$ use the observed identity in state of pairs of alleles:

$$\hat{f}_{\mathrm{AS}_j} = 1 - \frac{\sum_l \tilde{H}_{jl}}{\sum_l \tilde{H}_{Sl}}$$

$$\hat{\psi}_{\mathrm{AS}_{jj'}} = 1 - \frac{\sum_l \tilde{H}_{jj'l}}{\sum_l \tilde{H}_{Sl}}$$

$$\hat{\psi}_{\mathrm{AS}_j} = 1 - \frac{\sum_l \frac{1}{n-1} \sum_{j'=1, j' \neq j}^{n} \tilde{H}_{jj'l}}{\sum_l \tilde{H}_{Sl}}$$

For a large number of SNPs, but for all sample sizes,

$$\mathcal{E}(\hat{f}_{\mathrm{AS}_j}) = f_j = \frac{F_j - \theta_S}{1 - \theta_S}$$

$$\mathcal{E}(\hat{\psi}_{\mathrm{AS}_{jj'}}) = \psi_{jj'} = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

$$\mathcal{E}(\hat{\psi}_{\mathrm{AS}_j}) = \psi_j = \frac{\Psi_j - \theta_S}{1 - \theta_S}$$

Alternative Notation

Using "heterozygosity" for pairs of individuals is somewhat of an abuse of terminology. Looking forward to the Population Structure section suggests we work instead with allele sharing measures: A_j for the two alleles carried by individual j, A_{jj} for two alleles drawn randomly from individual j and $A_{jj'}$ for two alleles, one drawn randomly from individual j and one from individual j'. Their sample values are:

		$\tilde{A}_j$	$\tilde{A}_{jj}$
	AA	1	1
j	Aa	0	$\frac{1}{2}$
	aa	1	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

Value of 0.5 for two heterozygotes is different from the value 1 used in usual "number of pairs of alleles ibs."

Alternative Notation

In terms of allele dosages:

$$\tilde{A}_j = (X_j - 1)^2 \quad , \quad \mathcal{E}(\tilde{A}_j) = A + (1-A)F_j$$

$$\tilde{A}_{jj} = \frac{1}{2}[1 + (X_j - 1)^2] \quad , \quad \mathcal{E}(\tilde{A}_{jj}) = A + (1-A)\theta_{jj}$$

$$\tilde{A}_{jj'} = \frac{1}{2}[1 + (X_j - 1)(X_{j'} - 1)] \quad , \quad \mathcal{E}(\tilde{A}_{jj'}) = A + (1-A)\theta_{jj'}$$

$$\tilde{A}_S = \frac{1}{n(n-1)} \sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} \tilde{A}_{jj'} \quad , \quad \mathcal{E}(\tilde{A}_S) = A + (1-A)\theta_S$$

where

$$A = 1 - 2\pi(1-\pi) \quad , \quad \theta_{jj} = \frac{1}{2}(1 + Fj)$$

and, for large sample sizes,

$$1 - \tilde{A}_S \;=\; 2\tilde{p}(1-\tilde{p})$$

Allelic Matching Proportions for Individuals

Averaging over pairs of individuals:

$$\begin{aligned}
\tilde{A}_S &= \frac{1}{n(n-1)} \sum_{\substack{j=1 \\ j \neq j'}}^{n} \sum_{j'=1}^{n} \tilde{A}_{jj'} \\
\mathcal{E}(\tilde{A}_S) &= A + (1-A)\theta_S
\end{aligned}$$

The allele sharing kinship estimators and their expected values are

$$\hat{\psi}_{\mathsf{AS}_{jj'}} = \frac{\sum_l (\tilde{A}_{jj'l} - \tilde{A}_S)}{\sum_l (1 - \tilde{A}_S)}\ ,\ \mathcal{E}(\hat{\psi}_{\mathsf{AS}_{jj'}}) = \psi_{jj'} = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

The standard kinship estimators and their expected values are

$$\hat{\psi}_{\mathsf{STD}_{jj'}} = \frac{\sum_l (X_{jl} - 2\tilde{p}_l)(X_{j'l} - 2\tilde{p}_l)}{\sum_l 4\tilde{p}_l(1 - \tilde{p}_l)}\ ,\ \mathcal{E}(\hat{\psi}_{\mathsf{STD}_{jj'}}) = \psi_{jj'} - \psi_j - \psi_{j'}$$

Allelic Matching Proportions Within Populations

When the genotypic structure of data is not known, or is ignored, allelic data can be used to characterize population structure.

What is the proportion $\tilde{A}^i_{Wl}$ of random pairs of alleles in a sample from population i that are the same allelic type at SNP l?

If $\tilde{p}_{il}$ is the sample frequency for the SNP l reference allele:

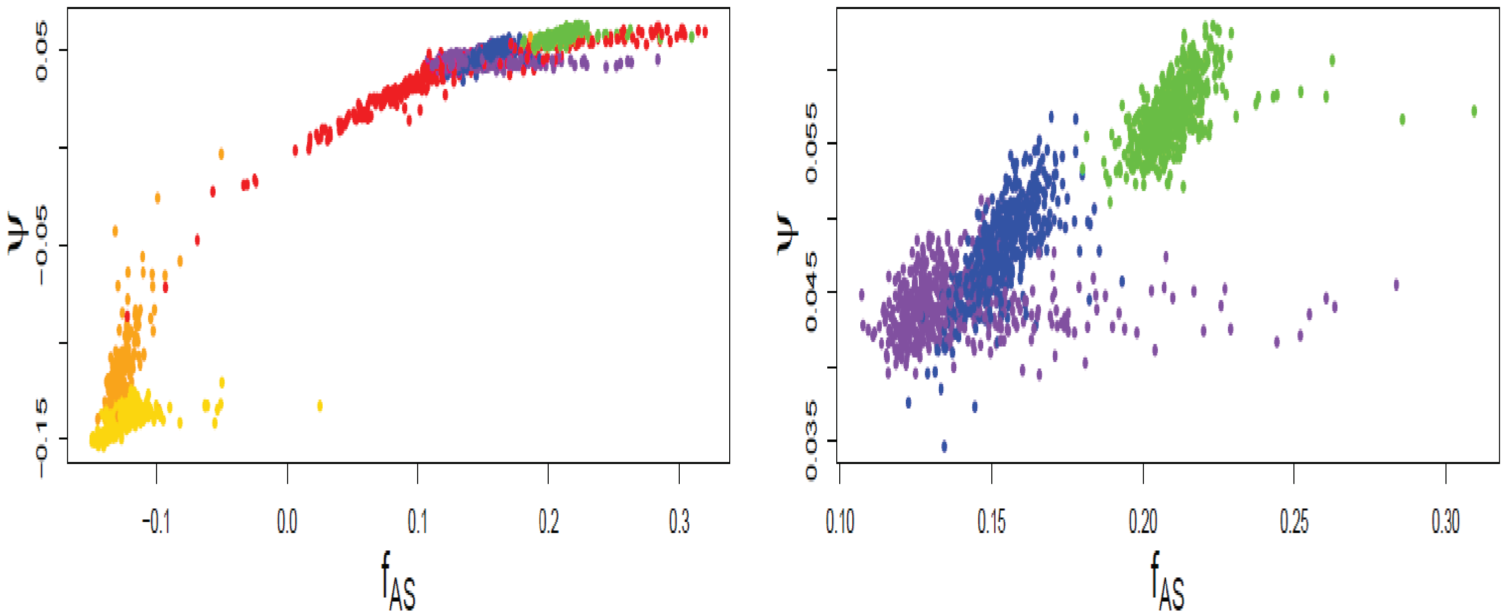
$$\tilde{A}^i_{Wl} \;=\; \tilde{p}^2_{il} + (1-\tilde{p}_{il})^2 = 1 - 2\tilde{p}_{il}(1-\tilde{p}_{il})$$

The expected value of this over replicates of the population is

$$\mathcal{E}(\tilde{A}^i_{Wl}) \;=\; A_l + (1-A_l)\theta^i_W$$

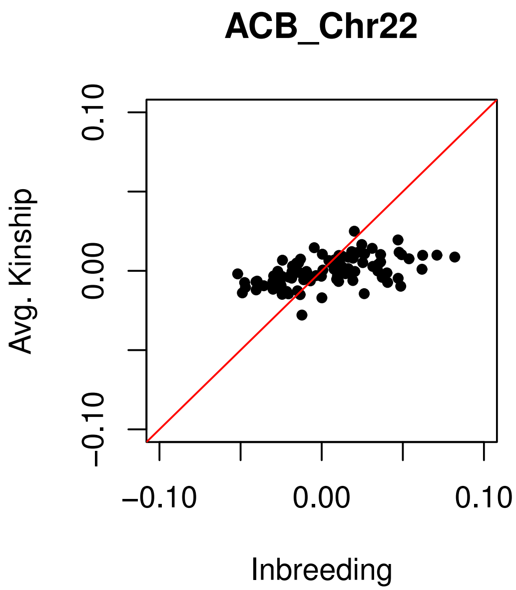
where $A_l = 2\pi_l(1-\pi_l)$. This is the key result: sample matching proportions for pairs of alleles depend on the probability of identity by descent for those pairs. There is an unknown function A_l of allele probabilities.

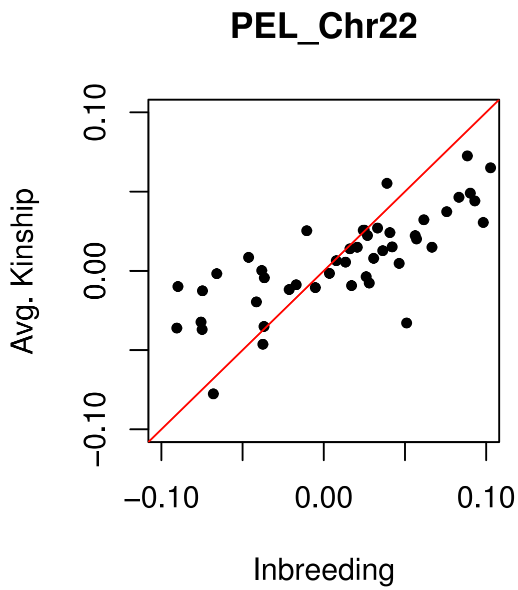
1000 Genomes Data



Estimates of within-population individual-specific average kinships vs estimates of within-population individual-specific inbreeding coefficients for 1000 Genomes data. Y-axis: $\widehat{\psi}_j$; X-axis: $\widehat{f}_j$. Left: All populations; Right: Excluding AMR and AFR. Gold: AFR (not ACB or ASW); Orange: AFR (ACB and ASW); Red: AMR; Purple: SAS; Blue: EUR; Green: EAS.

1000 Genomes Data





ACB
African Caribbean
in Barbados

PEL
Peruvian in
Lima, Peru

Many estimators of inbreeding assume no coancestry in a sample, and many estimators of coancestry assume no inbreeding. Inbreeding and coancestry should both be considered.

Allele Frequencies

Allele-sharing estimates avoid the need to estimate allele probabilities.

Could regard sample allele frequencies as estimates of allele probabilities: often gives similar estimates of inbreeding and coancestry coefficients to the allele-sharing estimates. But, there can be changes of rank as the scope of the study changes.

Alternatively, could estimate allele probabilities jointly with ibd probabilities. Iterative methods update allele probabilities and ibd probabilities in turn. Difficult to count ibd alleles at each stage: which alleles should be considered – just within individuals, plus between pairs of individuals, ...

Alternative Estimators

Other estimators use sample allele frequencies and sample allele dosages: the number of copies of one of the two alleles carried by an individual at a locus. If X_{jl} is the dosage for the allele at SNP l for individual j, the ratio of averages form of the standard estimator (e.g. in GCTA package) is

$$\hat{f}^{\mathsf{W}}_{\mathsf{Std}_j} = \frac{\sum_l (X_{jl} - 2\tilde{p}_l)^2}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)}$$

although it is common to see the average of ratios form

$$\hat{f}^{\mathsf{u}}_{\mathsf{Std}_j} = \frac{1}{L}\sum_{l=1}^{L} \frac{(X_{jl} - 2\tilde{p}_l)^2}{2\tilde{p}_l(1-\tilde{p}_l)}$$

An alternative form (Yang et al, 2011) is

$$\hat{f}^{\mathsf{W}}_{\mathsf{Uni}_j} = \frac{\sum_l [X_{jl}^2 - (1+2\tilde{p}_l)X_{jl} + 2\tilde{p}_l^2]}{\sum_l 2\tilde{p}_l(1-\tilde{p}_l)}$$

Expectations of Alternative Estimators

Although, for all sample sizes:

$$\mathcal{E}(\tilde{H}_S) \;=\; (1-\theta_S)\sum_l [2\pi_l(1-\pi_l)]$$

it is only for large sample sizes that:

$$\mathcal{E}[\sum_l 2\tilde{p}_l(1-\tilde{p}_l)] \;=\; (1-\theta_S)\sum_l [2\pi_l(1-\pi_l)]$$

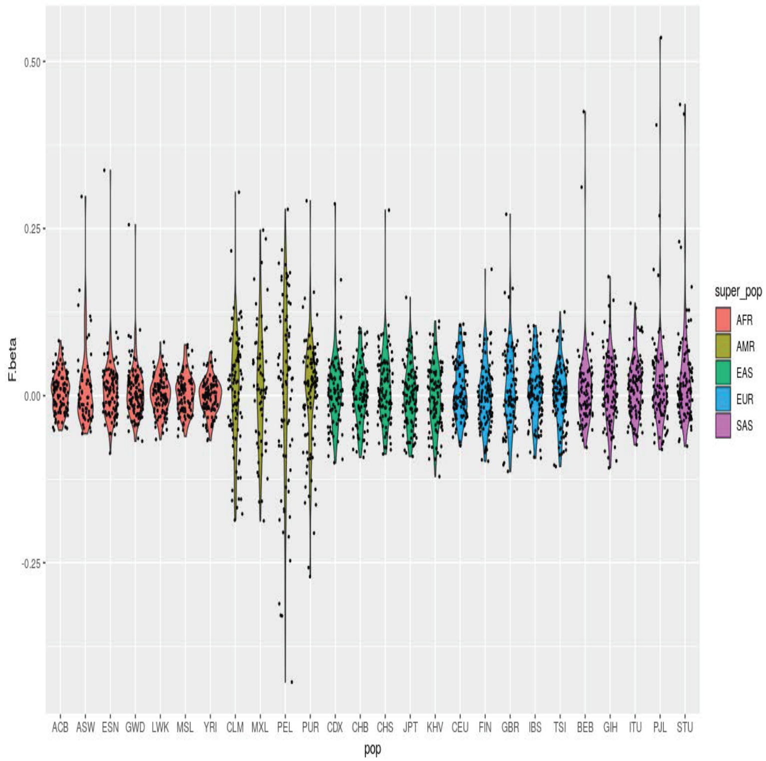
In the large-sample case

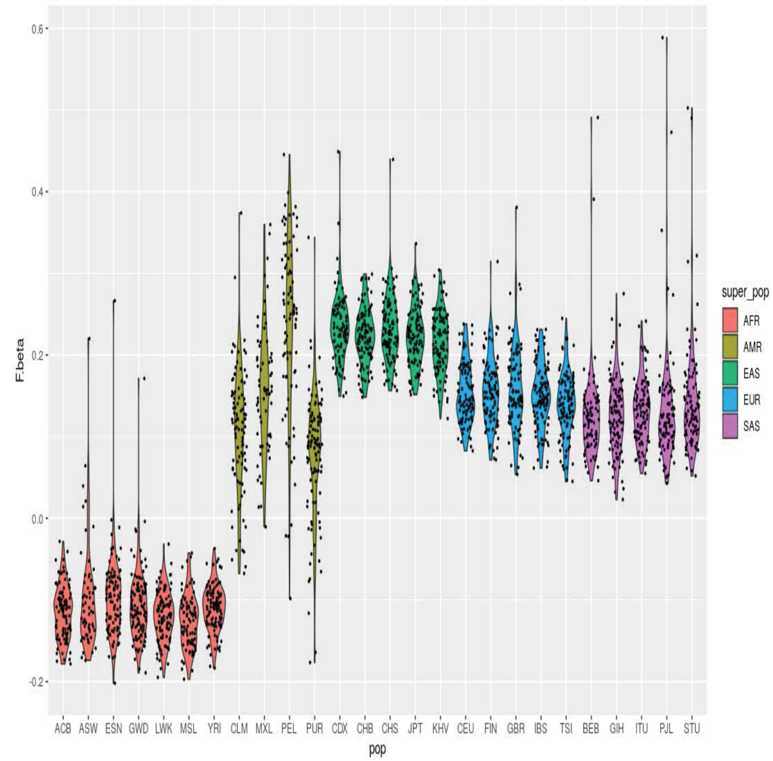
$$\mathcal{E}(\widehat{f}^{\mathrm{W}}_{\mathrm{Uni}_j}) \;=\; \frac{f_j - \Psi_j + \theta_S}{1-\theta_S} = f_j - 2\psi_j$$

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

The ranks of $\widehat{f}^{\mathrm{W}}_{\mathrm{Uni}_j}$ and $\widehat{f}^{\mathrm{W}}_{\mathrm{Std}_j}$ may be different from the ranks of f_j and ψ_j i.e. of F_j and Ψ_j.

1000 Genomes Data





Local Population Reference　　Whole World Reference

Chromosome 22 data from 1000 Genomes.

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

Expectations of Alternative Kinship Estimators

For kinship:

$$\hat{\psi}^{\text{W}}_{\text{Std}_{jj'}} = \frac{\sum_l (X_{jl} - 2\tilde{p}_l)(X_{jj'l} - 2\tilde{p}_l)}{\sum_l 4\tilde{p}_l(1 - \tilde{p}_l)}$$

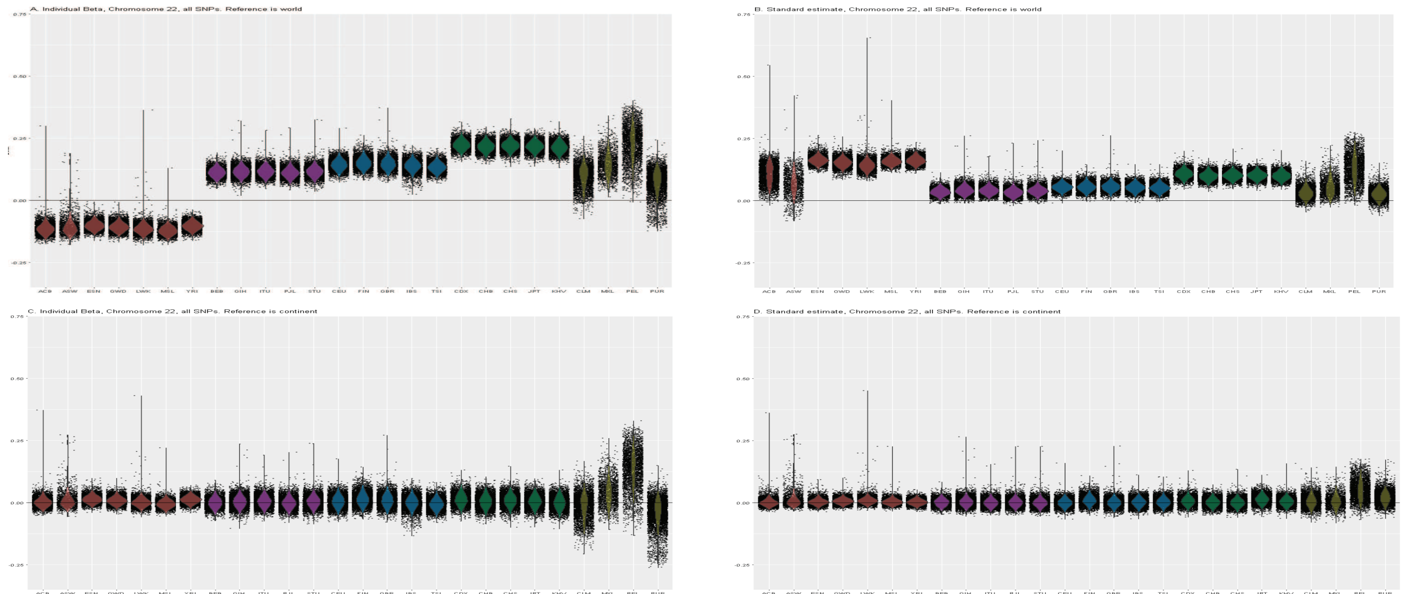
and this has an expected value, for large sample sizes, of

$$\begin{aligned} \mathcal{E}(\hat{\psi}^{\text{W}}_{\text{Std}_{jj'}}) &= \frac{\theta_{jj'} - \psi_j - \psi_{jj'} + \theta_S}{1 - \theta_S} \\ &= \psi_{jj'} - \psi_j - \psi_{j'} \end{aligned}$$

Unlike $\hat{\psi}_{\text{AS}_{jj'}}$, the standard kjnship estimates are not expected to have the same ranks as the $\theta_{jj'}$'s.

1000 Genomes Data

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



Allele sharing estimates Standard 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

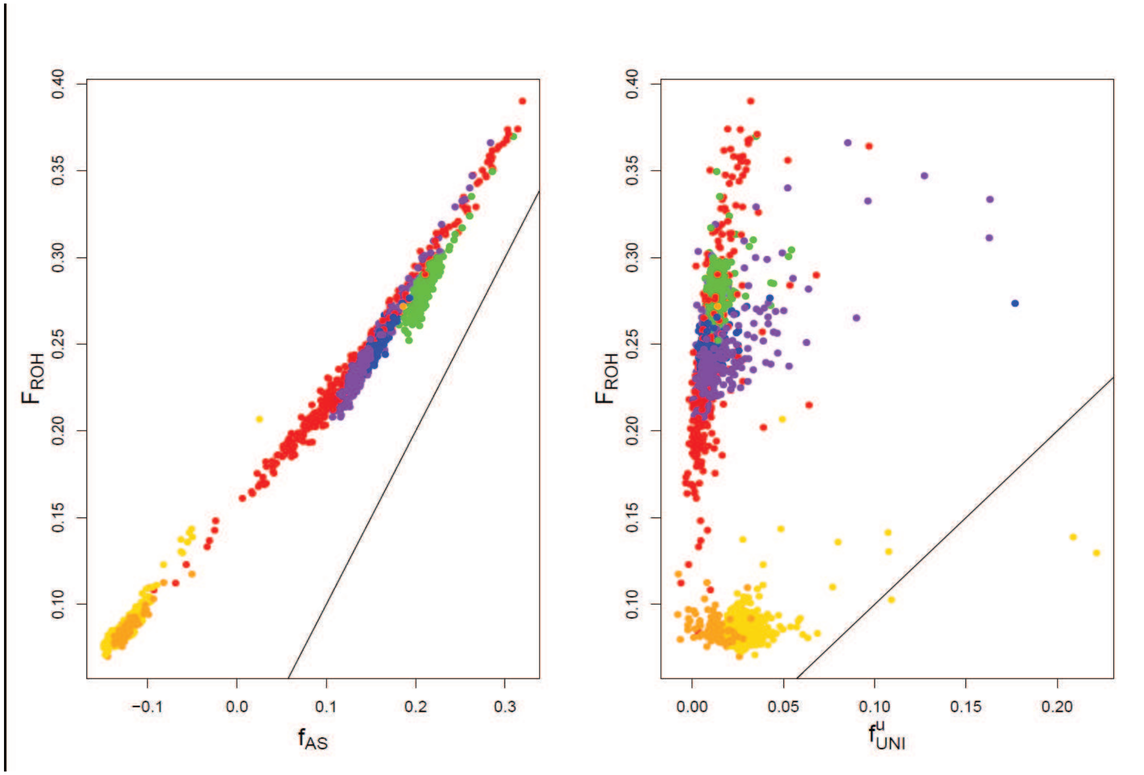
Alternative Estimators: Runs of Homozygosity

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

1000 Genomes Data

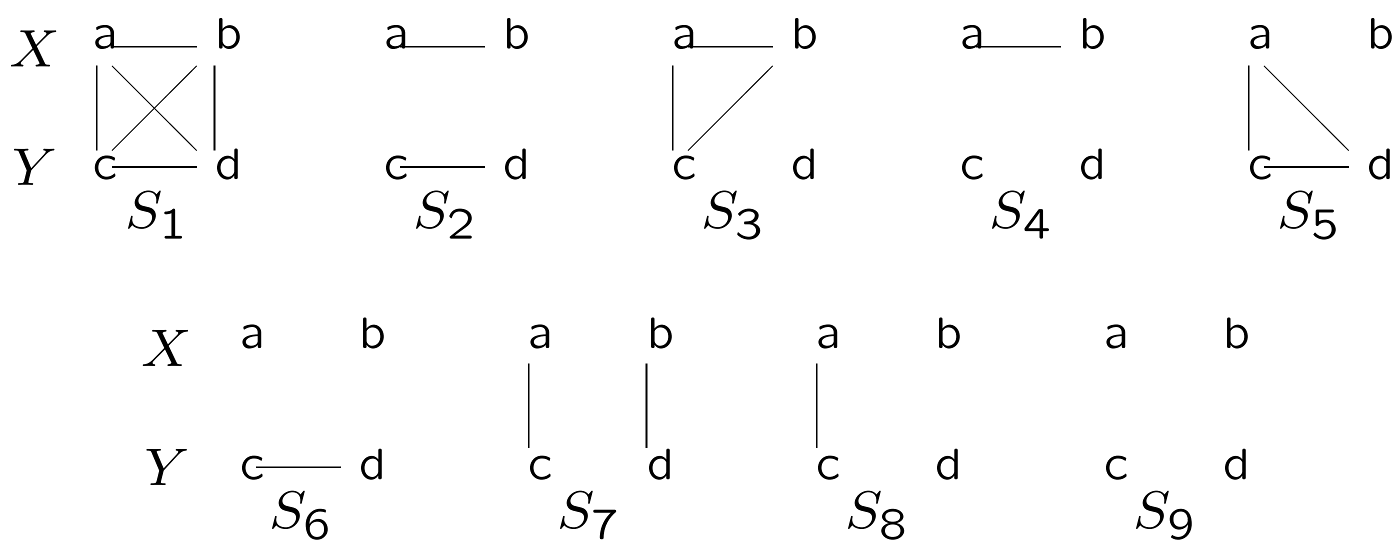


ROH/PLINK estimates vs SNP by SNP estimates for 1000 Genomes data, with the World as a reference set. Left: $\widehat{F}_{\mathrm{ROH}}$ vs $\widehat{f}_{\mathrm{AS}}$; Right: F_{ROH} vs $\widehat{f}^{u}_{\mathrm{Uni}}$. Solid line $X = Y$. Gold: AFR (not ACB or ASW); Orange: AFR (ACB and ASW); Red: AMR; Purple: SAS; Blue: EUR; Green: EAS.

Nine ibd States S_i

Full set of nine ibd states S_i need to be considered for natural populations with mixed mating systems and for quantitative genetic analyses of non-additive gene action.

Solid lines join pairs of ibd alleles: top row shows alleles a, b for individual X, bottom row shows alleles c, d for individual Y.



Genotype Probabilities for Two Individuals

X, Y	Δ_1	Δ_2	Δ_3	Δ_5	Δ_4	Δ_6	Δ_7	Δ_8	Δ_9
$G_1 : AA, AA$	π_A	π_A^2	π_A^2	π_A^2	π_A^3	π_A^3	π_A^2	π_A^3	π_A^4
$G_2 : aa, aa$	π_a	π_a^2	π_a^2	π_a^2	π_a^3	π_a^3	π_a^2	π_a^3	π_a^4
$G_3 : Aa, Aa$							$2\pi_A\pi_a$	$\pi_A\pi_a$	$4\pi_A^2\pi_a^2$
$G_4 : AA, Aa$			$\pi_A\pi_a$		$2\pi_A^2\pi_a$			$\pi_A^2\pi_a$	$2\pi_A^3\pi_a$
$G_5 : aa, Aa$			$\pi_A\pi_a$		$2\pi_A\pi_a^2$			$\pi_A\pi_a^2$	$2\pi_A\pi_a^3$
$G_6 : Aa, AA$				$\pi_A\pi_a$		$2\pi_A^2\pi_a$		$\pi_A^2\pi_a$	$2\pi_A^3\pi_a$
$G_7 : Aa, aa$				$\pi_A\pi_a$		$2\pi_A\pi_a^2$		$\pi_A\pi_a^2$	$2\pi_A\pi_a^3$
$G_8 : AA, aa$		$\pi_A\pi_a$			$\pi_A\pi_a^2$	$\pi_A^2\pi_a$			$\pi_A^2\pi_a^2$
$G_9 : aa, AA$		$\pi_A\pi_a$			$\pi_A^2\pi_a$	$\pi_A\pi_a^2$			$\pi_A^2\pi_a^2$

π_A, π_a are allele probabilities. Two dependencies among genotype-pair probabilities:

$$G_1 + G_2 + G_3 + G_4 + G_5 + G_6 + G_7 + G_8 + G_9 = 1$$

$$G_4 + 2G_8 + G_7 = G_6 + 2G_9 + G_5$$

Probabilities for Unordered Individuals

$X\&Y$	Δ_1	Δ_2	$\Delta_3+\Delta_5$	$\Delta_4+\Delta_6$	Δ_7	Δ_8	Δ_9
$G_1 : AA\&AA$	π_A	π_A^2	π_A^2	π_A^3	π_A^2	π_A^3	π_A^4
$G_2 : aa\&aa$	π_a	π_a^2	π_a^2	π_a^3	π_a^2	π_a^3	π_a^4
$G_3 : Aa\&Aa$					$2\pi_A\pi_a$	$\pi_A\pi_a$	$4\pi_A^2\pi_a^2$
$G_4+G_6 : AA\&Aa$			$\pi_A\pi_a$	$2\pi_A^2\pi_a$		$2\pi_A^2\pi_a$	$4\pi_A^3\pi_a$
$G_5+G_7 : aa\&Aa$			$\pi_A\pi_a$	$2\pi_A\pi_a^2$		$2\pi_A\pi_a^2$	$4\pi_A\pi_a^3$
$G_8+G_9 : AA\&aa$		$2\pi_A\pi_a$		$\pi_A\pi_a$			$2\pi_A^2\pi_a^2$

No distinction now between Δ_3 and Δ_5 or between Δ_4 and Δ_6.

Summary ibd Measures

The 9 Δ's can be summarized with 8 linear functions of them that refer to pairs, trios, two-pairs and quadruples of alleles:

Summary	Jacquard/Cockerham
$F_X = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$	$\Delta_4 = F_X - 2\gamma_{\ddot{X}Y} - \Delta_{\ddot{X}.\ddot{Y}} + 2\delta_{\ddot{X}\ddot{Y}}$
$F_Y = \Delta_1 + \Delta_2 + \Delta_5 + \Delta_6$	$\Delta_6 = F_Y - 2\gamma_{X\ddot{Y}} - \Delta_{\ddot{X}.\ddot{Y}} + 2\delta_{\ddot{X}\ddot{Y}}$
$\theta_{XY} = \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4}\Delta_8$	$\Delta_8 = 4\theta_{XY} - 4\gamma_{\ddot{X}Y} - 4\gamma_{X\ddot{Y}} - 4\Delta_{\ddot{X}+\ddot{Y}} + 8\delta_{\ddot{X}\ddot{Y}}$
$\gamma_{\ddot{X}Y} = \Delta_1 + \frac{1}{2}\Delta_3$	$\Delta_3 = 2(\gamma_{\ddot{X}Y} - \delta_{\ddot{X}\ddot{Y}})$
$\gamma_{X\ddot{Y}} = \Delta_1 + \frac{1}{2}\Delta_5$	$\Delta_5 = 2(\gamma_{X\ddot{Y}} - \delta_{\ddot{X}\ddot{Y}})$
$\Delta_{\ddot{X}.\ddot{Y}} = \Delta_1 + \Delta_2$	$\Delta_2 = \Delta_{\ddot{X}.\ddot{Y}} - \delta_{\ddot{X}\ddot{Y}}$
$\Delta_{\ddot{X}+\ddot{Y}} = \Delta_1 + \frac{1}{2}\Delta_7$	$\Delta_7 = 2(\Delta_{\ddot{X}+\ddot{Y}} - \delta_{\ddot{X}\ddot{Y}})$
$\delta_{\ddot{X}\ddot{Y}} = \Delta_1$	$\Delta_1 = \delta_{\ddot{X}\ddot{Y}}$
	$\Delta_9 = 1 - F_X - F_Y - 4\theta_{XY} + 4\gamma_{\ddot{X}Y} + 4\gamma_{X\ddot{Y}}$
	$+\Delta_{\ddot{X}.\ddot{Y}} + 2\Delta_{\ddot{X}+\ddot{Y}} - 6\delta_{\ddot{X}\ddot{Y}}$

Unordered Individuals Summary Measures

$$\begin{aligned}
\frac{1}{2}(F_X + F_Y) &= \Delta_1 + \Delta_2 + \frac{1}{2}(\Delta_3 + \Delta_5) + \frac{1}{2}(\Delta_4 + \Delta_6) \\
\theta_{XY} &= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5) + \frac{1}{2}\Delta_7 + \frac{1}{4}\Delta_8 \\
\frac{1}{2}(\gamma_{\ddot{X}Y} + \gamma_{X\ddot{Y}}) &= \Delta_1 + \frac{1}{2}(\Delta_3 + \Delta_5) \\
\Delta_{\ddot{X}.\ddot{Y}} &= \Delta_1 + \Delta_2 \\
\Delta_{\ddot{X}+\ddot{Y}} &= \Delta_1 + \frac{1}{2}\Delta_7 \\
\delta_{\ddot{X}\ddot{Y}} &= \Delta_1
\end{aligned}$$

$$\begin{aligned}
(\Delta_4 + \Delta_6) &= (F_X + F_Y) - 2(\gamma_{\ddot{X}Y} + \gamma_{X\ddot{Y}}) - 2\Delta_{\ddot{X}.\ddot{Y}} + 4\delta_{\ddot{X}\ddot{Y}} \\
\Delta_8 &= 4\theta_{XY} - 4(\gamma_{\ddot{X}Y} + \gamma_{X\ddot{Y}}) - 4\Delta_{\ddot{X}+\ddot{Y}} + 8\delta_{\ddot{X}\ddot{Y}} \\
(\Delta_3 + \Delta_5) &= 2(\gamma_{\ddot{X}Y} + \gamma_{X\ddot{Y}}) - 4\delta_{\ddot{X}\ddot{Y}}) \\
\Delta_2 &= \Delta_{\ddot{X}.\ddot{Y}} - \delta_{\ddot{X}\ddot{Y}} \\
\Delta_7 &= 2(\Delta_{\ddot{X}+\ddot{Y}} - \delta_{\ddot{X}\ddot{Y}}) \\
\Delta_1 &= \delta_{\ddot{X}\ddot{Y}} \\
\Delta_9 &= 1 - (F_X + F_Y) - 4\theta_{XY} + 4(\gamma_{\ddot{X}Y} + \gamma_{X\ddot{Y}}) + \Delta_{\ddot{X}.\ddot{Y}} + 2\Delta_{\ddot{X}+\ddot{Y}} - 6\delta_{\ddot{X}\ddot{Y}}
\end{aligned}$$

Application of Summary Measures

For traits with dominance, the covariance of genetic effects G_X, G_Y for inbred and related individuals X, Y is

$$\begin{aligned}\text{Cov}(G_X, G_Y) &= 2\theta_{XY}\sigma_A^2 + 2\Delta_{\ddot{X}+\ddot{Y}}\sigma_D^2 \\ &\quad + 2(\gamma_{\ddot{X}Y} + \gamma_{X\ddot{Y}})C_1 + \delta_{\ddot{X}\ddot{Y}}C_2 + (\Delta_{\ddot{X}.\ddot{Y}} - F_X F_Y)C_3\end{aligned}$$

The additive and dominance variance components are σ_A^2, σ_D^2 and C_1 is the covariance between additive and homozygous dominance deviations, C_2 is the variance of homozygous dominance effects, and C_3 is the squared sum of homozygous dominance effects.

For additive traits

$$\text{Cov}(G_X, G_Y) = 2\theta_{XY}\sigma_A^2 \quad , \quad \text{Var}(G_X) = (1 + F_X)\sigma_A^2$$

Predicted Values of Summary Measures

The summary measures for a pedigree can be calculated by tracing alleles back to the founders.

For a random mating population at drift/mutation equilibrium:

$$\theta = \frac{1}{1+4N\mu} \quad \text{any pair of alleles}$$

$$\gamma = \frac{2\theta^2}{1+\theta} \quad \text{any three alleles}$$

$$\delta = \frac{6\theta^3}{(1+\theta)(1+2\theta)} \quad \text{any four alleles}$$

$$\Delta = \frac{\theta^2(1+5\theta)}{(1+\theta)(1+2\theta)} \quad \text{any two pairs of alleles}$$

We have little knowledge about actual values of these quantities.

Estimation of Summary Measures

By analogy to the two-allele case, consider the ibs states for two genotypes.

Genotypes	ibs alleles
AA,AA and aa,aa	All four alleles ibs
AA,aa and aa,AA	ibs within both indivs, no ibs between indivs
Aa,Aa	no ibs within indivs, two ibs pairs between indivs
AA,Aa; aa,Aa	ibs within first indiv, one ibs pair between indivs
Aa,AA; Aa,aa	ibs within second indiv, one ibs pair between indivs

The five states are consistent with the claim of there being five identifiable ibd states for loci with two alleles:
Csűrös M. 2014. Non-identifiability of identity coefficients at biallelic loci. Theoretical Population Biology 92:22-29.

Combine last two rows if individuals not ordered.

Estimation of Summary Measures

X, Y	Δ_1	Δ_2	Δ_3	Δ_5	Δ_4	Δ_6	Δ_7	Δ_8	Δ_9
$G_1 + G_2$	1	$1 - H$	$1 - H$	$1 - H$	$1 - \frac{3}{2}H$	$1 - \frac{3}{2}H$	$1 - H$	$1 - \frac{3}{2}H$	$1 - 2H(1 - H) - 3K$
$G_8 + G_9$		H			$\frac{1}{2}H$	$\frac{1}{2}H$			K
G_3							H	$\frac{1}{2}H$	$2K$
$G_4 + G_5$			H		H			$\frac{1}{2}H$	$H(1 - H)$
$G_6 + G_7$				H		H		$\frac{1}{2}H$	$H(1 - H)$

$H = 2\pi_A\pi_a, K = 2\pi_A^2\pi_a^2$

Note that

$$\begin{aligned} \mathcal{E}[(\tilde{G}_4 + \tilde{G}_5) - (\tilde{G}_6 + \tilde{G}_7)] &= H([\Delta_3 - \Delta_5) + (\Delta_4 - \Delta_6)] \\ &= H(F_X - F_Y) \end{aligned}$$

The table with summary measures does not have a simple structure.

Estimation of Summary Measures

X,Y	1	F_X	F_Y	θ_{XY}	$\gamma_{\ddot{X}Y}$	$\gamma_{X\ddot{Y}}$	$\Delta_{\ddot{X}.\ddot{Y}}+$ $2\Delta_{\ddot{X}+\ddot{Y}}$	$\delta_{\ddot{X}\ddot{Y}}$
G_1	p^4	p^3q	p^3q	$4p^3q$	$2p^2q-4p^3q$	$2p^2q-4p^3q$	x	y
G_2	q^4	pq^3	pq^3	$4pq^3$	$2pq^2-4pq^3$	$2pq^2-4pq^3$	x	y
G_3	$4p^2q^2$	$-4p^2q^2$	$-4p^2q^2$	$4pq-16p^2q^2$	$-4pq+16p^2q^2$	$-4pq+16p^2q^2$	$4x$	$4y$
G_4	$2p^3q$	$2p^2q^2$	$-2p^3q$	$4p^2q-8p^3q$	$2pq-8p^2q^2$	$-4p^2q+8p^3q$	$-2x$	$-2y$
G_5	$2pq^3$	$2p^2q^2$	$-2pq^3$	$4pq^2-8pq^3$	$2pq-8p^2q^2$	$-4pq^2+8pq^3$	$-2x$	$-2y$
G_6	$2p^3q$	$-2p^3q$	$2p^2q^2$	$4p^2q-8p^3q$	$-4p^2q+8p^3q$	$2pq-8p^2q^2$	$-2x$	$-2y$
G_7	$2pq^3$	$-2pq^3$	$2p^2q^2$	$4pq^2-8pq^3$	$-4pq^2+8pq^3$	$2pq-8p^2q^2$	$-2x$	$-2y$
G_8	p^2q^2	pq^3	p^3q	$-4p^2q^2$	$-2pq^2+4p^2q^2$	$-2p^2q+4p^2q^2$	x	y
G_9	p^2q^2	p^3q	pq^3	$-4p^2q^2$	$-2p^2q+4p^2q^2$	$-2pq^2+4p^2q^2$	x	y

$p=\pi_A, q=\pi_a, x=p^2q^2, y=pq-6p^2q^2$

Would like to estimate Jacquard or Summary probabilities from the observed proportions $\tilde{G}_i$ of SNPs that fall in genotype-pair categories $i=1,2,\ldots,9$. From genotypic data for loci with two alleles, $\Delta_{\ddot{X}.\ddot{Y}}$ cannot be distinguished from $\Delta_{\ddot{X}+\ddot{Y}}$.

Without Inbreeding

For two non-inbred individuals only $\kappa_2 = \Delta_7, \kappa_1 = \Delta_8, \kappa_0 = \Delta_9$ are not zero. Simple moment estimators use the number N_i of SNPs for which two individuals share i pairs in alleles ibs. If $H = \sum_{l=1}^{L} 2\pi_l(1-\pi_l), K = \sum_{l=1}^{L} 2\pi_l^2(1-\pi_l)^2$:

$$
\begin{aligned}
\mathcal{E}_T(N_0) &= \frac{1}{2}(2\Delta_2 + \Delta_4 + \Delta_6)H + \Delta_9 K \\
\mathcal{E}_T(N_1) &= (\Delta_3 + \Delta_4 + \Delta_5 + \Delta_6 + \Delta_8 + 2\Delta_9)H - 4\Delta_9 K \\
\mathcal{E}_T(N_2) &= L - \frac{1}{2}(2\Delta_2 + 2\Delta_3 + 3\Delta_4 + 2\Delta_5 + 3\Delta_6 + 2\Delta_8 + 4\Delta_9)H + 3\Delta_9 K
\end{aligned}
$$

It is usual to replace π_l for the SNP l reference allele by $\tilde{p}_l$, set $\Delta_i = 0, i \leq 6$ and and solve for the κ's:

$$
\widehat{\kappa}_0 = \frac{N_0}{\tilde{K}}, \; \widehat{\kappa}_1 = \frac{N_1 + 4N_0}{\tilde{H}} - \frac{2N_0}{\tilde{K}}, \; \widehat{\kappa}_2 = 1 - \frac{N_1 + 4N_0}{\tilde{H}} + \frac{N_0}{\tilde{K}}
$$

$$
\widehat{\theta} = \frac{1}{2}\widehat{\kappa}_2 + \frac{1}{4}\widehat{\kappa}_1 = \frac{1}{2} - \frac{N_1 + 4N_0}{4\tilde{H}}
$$

Without Inbreeding

Problem 1: The expected values of $\tilde{H}$ and $\tilde{K}$, for large sample sizes, are

$$\begin{aligned}\mathcal{E}(\tilde{H}) &= (1-\theta_S)H \\ \mathcal{E}(\tilde{K}) &= (1-6\theta_S+8\gamma_S+3\Delta_S-6\delta_S)K-(\theta_S-2\gamma_S+\delta_S)H\end{aligned}$$

where $\theta_S, \gamma_S, \delta_S, \Delta_S$ are the ibd probabilities for random pairs, triples, quadruples and two-pairs of alleles from distinct individuals in the sampled population.

Even if there is no inbreeding, $\widehat{\theta}$ is biased for either θ or ψ:

$$\mathcal{E}(\widehat{\theta}) = \frac{\theta-\frac{1}{2}\theta_S}{1-\theta_S}$$

Problem 2: There is inbreeding.

Three Alleles in Two Individuals

For two alleles, two in individual X and one taken randomly from individual Y, there are four states of identity by descent:

Three ibd	$\gamma_{\ddot{X}Y}$
Two ibd within X	$F_X - \gamma_{\ddot{X}Y}$
Two ibd between X, Y	$2(\theta_{XY} - \gamma_{\ddot{X}Y})$
No ibd	$\gamma_{0_{\ddot{X}Y}} = 1 - F_X - 2\theta_{XY} + 2\gamma_{\ddot{X}Y}$

The probabilities of all six sets of allelic states are

	At least two ibd alleles			No ibd alleles
	$\Delta_1 + \frac{1}{2}\Delta_3$	$\Delta_2 + \frac{1}{2}\Delta_3 + \Delta_4$	$\Delta_5 + \Delta_7 + \frac{1}{2}\Delta_8$	$\Delta_6 + \frac{1}{2}\Delta_8 + \Delta_9$
X, Y	$\gamma_{\ddot{X}Y}$	$(F_X - \gamma_{\ddot{X}Y})$	$2(\theta_{XY} - \gamma_{\ddot{X}Y})$	$(1 - F_X - 2\theta_{XY} + 2\gamma_{\ddot{X}Y})$
AA, A	π_A	π_A^2	π_A^2	π_A^3
AA, a		$\pi_A\pi_a$		$\pi_A^2\pi_a$
Aa, A			$\pi_A\pi_a$	$2\pi_A^2\pi_a$
Aa, a			$\pi_A\pi_a$	$2\pi_A\pi_a^2$
aa, A		$\pi_A\pi_a$		$\pi_A\pi_a^2$
aa, a	π_a	π_a^2	π_a^2	π_a^3
$Aa, a - Aa, A$				$-2\pi_A\pi_a(\pi_A - \pi_a)$

Three Alleles in Two Individuals

The observed value for $(Aa, a - Aa, A)$ is $(\tilde{G}_7 - \tilde{G}_6)$ and

$$\mathcal{E}(\tilde{G}_7 - \tilde{G}_6) \;=\; 2\pi_A(1-\pi_A)(1-2\pi_A)(1 - F_X - 2\theta_{XY} + 2\gamma_{\ddot{X}Y})$$

Csűrös (2014) assumed that

$$\mathcal{E}[2\tilde{p}_A(1-\tilde{p}_A)(1-2\tilde{p}_A)] \;=\; 2\pi_A(1-\pi_A)(1-2\pi_A)$$

in order to estimate $(1 - F_X - 2\theta_{XY} + 2\gamma_{\ddot{X}Y})$. He also assumed that F_X, θ_{XY} could be estimated and therefore he thought that $\gamma_{\ddot{X}Y}$ was estimable.

However,

$$\mathcal{E}[2\tilde{p}_A(1-\tilde{p}_A)(1-2\tilde{p}_A)] \;=\; 2\pi_A(1-\pi_A)(1-2\pi_A)(1 - 3\theta_S + 2\gamma_S)$$

where γ_S is the ibd probability for any three alleles, one from each of three individuals in the sample.

Three Alleles in Two Individuals

The unknown allele probabilities do not enter into the expectation of the ratio $\tilde{T}3_{\ddot{X}Y} = (\tilde{G}_7 - \tilde{G}_6)/[2\tilde{p}_A(1-\tilde{p}_A)(1-2\tilde{p}_A)]$:

$$\mathcal{E}(\tilde{T}3_{\ddot{X}Y}) \;=\; \frac{1 - F_X - 2\theta_{XY} + 2\gamma_{\ddot{X}Y}}{1 - 3\theta_S + 2\gamma_S} = \frac{\gamma 0_{\ddot{X}Y}}{\gamma 0_S}$$

The numerator is the probability of no identity by descent among the two alleles of X and a random allele of Y. The denominator is the probability of no identity by descent from any three alleles, one from each of three individuals, in the sample.

For large sample sizes, the denominator is the same as the average of the numerator for all sets of three alleles from three distinct individuals.

Ignoring order for Individuals

Ignoring the order of two individuals, by adding in the observed value $(\tilde{G}_5 - \tilde{G}_4)$ of $(a, Aa - A, Aa,)$ and averaging with $(\tilde{G}_7 - \tilde{G}_6)$:

$$\tilde{T}_3 = \frac{(\tilde{G}_7 - \tilde{G}_6) + (\tilde{G}_5 - \tilde{G}_4)}{4\tilde{p}_A(1 - \tilde{p}_A)(1 - 2\tilde{p}_A)}$$

$$\mathcal{E}(\tilde{T}_3) = \frac{1 - \frac{1}{2}(F_X + F_Y) - 2\theta_{XY} + (\gamma_{\ddot{X}Y} + \gamma_{X\ddot{Y}})}{1 - 3\theta_S + 2\gamma_S}$$

The numerator is the probability of no identity by descent among the two alleles of one individual and a random allele of the other. The denominator is the probability of no identity by descent from any three alleles, one from each of three individuals, in the sample.

This estimator, but not its expectation, was given by Csűrös (2014).

Two Alleles in Two Individuals

This is analogous to the case for two alleles. There are two ibd states for a pair of alleles within or between individuals:

ibd within X	F_X	ibd between X, Y	θ_{XY}
no ibd within$X$	$F_{0_X} = 1 - F_X$	No ibd between X, Y	$\theta_{0_{XY}} = 1 - \theta_{XY}$

There are three sets of allelic states:

X	F_X	$1 - F_X$	X, Y	θ_{XY}	$1 - \theta_{XY}$
AA	π_A	π_A^2	A, A	π_A	π_A^2
Aa		$2\pi_A\pi_a$	A, a		$2\pi_A\pi_a$
aa	π_a	π_a^2	a, a	π_a	π_a^2

Two Alleles in Two Individuals

The observed values for Aa or A, a are $\tilde{H}_X$ and $\tilde{H}_{XY}$, with large-sample expectations

$$\mathcal{E}(\tilde{H}_X) = 2\pi_A(1-\pi_A)F_{0_X} \quad , \quad \mathcal{E}(\tilde{H}_{XY}) = 2\pi_A(1-\pi_A)\theta_{0_{XY}}$$
$$\mathcal{E}[2\tilde{p}_A(1-\tilde{p}_A)] = 2\pi_A(1-\pi_A)\theta_{0_S} \quad , \quad \mathcal{E}[2\tilde{p}_A(1-\tilde{p}_A)] = 2\pi_A(1-\pi_A)\theta_{0_S}$$

Large-sample estimators for the estimable functions do not depend on unknown allele probabilities:

$$\tilde{T}_{2_X} = \frac{\tilde{H}_X}{2\tilde{p}_A(1-\tilde{p}_A)} \quad , \quad \tilde{T}_{2_{XY}} = \frac{\tilde{H}_{XY}}{2\tilde{p}_A(1-\tilde{p}_A)}$$

$$\mathcal{E}(\tilde{T}_{2_X}) = \frac{F_{0_X}}{\theta_{0_S}} \quad , \quad \mathcal{E}(\tilde{T}_{2_{XY}}) = \frac{\theta_{0_{XY}}}{\theta_{0_S}}$$

Ignoring Order for Two Individuals

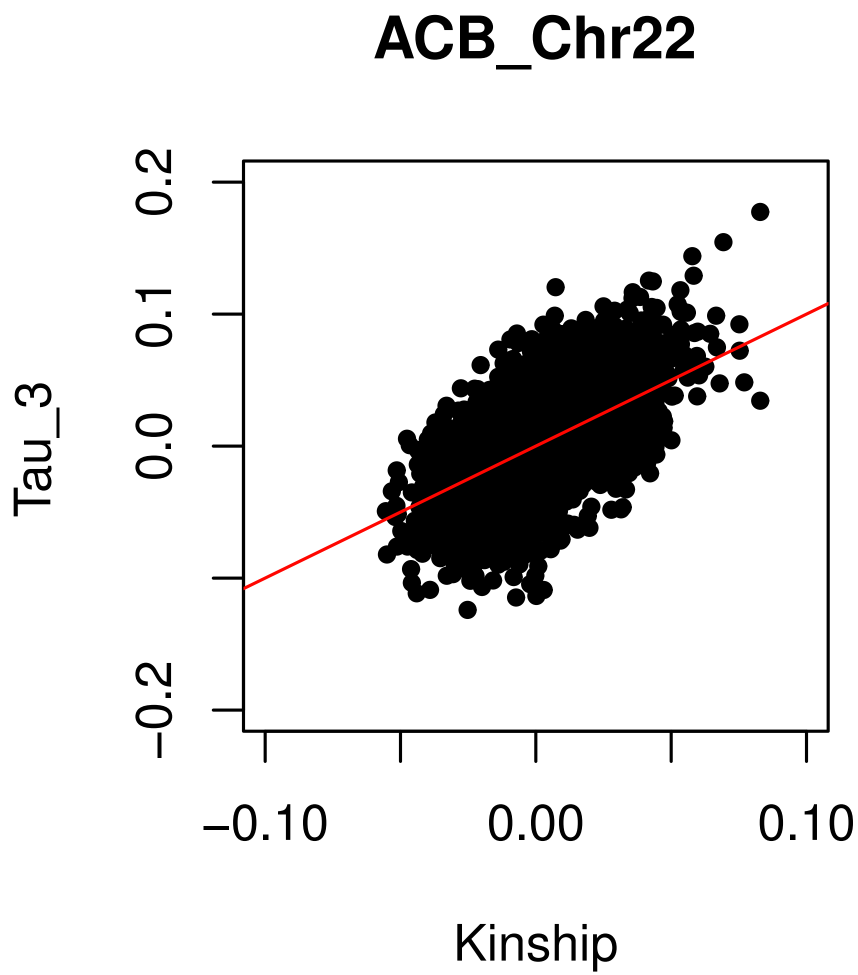
Values of a combined two-allele estimator $\tilde{T}_2$ might be compared with values of the combined three-allele estimator $\tilde{T}_3$.

$$\tilde{T}_2 = \frac{\frac{1}{2}(\tilde{H}_X + \tilde{H}_Y) + 2\tilde{H}_{XY}}{2\tilde{p}_A(1 - \tilde{p}_A)}$$

The expected value of this is a linear function of $[(F_X + F_Y)/2 + 2\theta_{XY}]$ whereas $\tilde{T}_3$ has an expected value that is a linear function of $[(F_X + F_Y)/2 + 2\theta_{XY} - (\gamma_{\ddot{X}Y} + \gamma_{X\ddot{Y}})]$.

There will be evidence for non-zero three-allele ibd probabilities if $\tilde{T}_3$ is not linearly related to $\tilde{T}_2$.

1000 Genomes Data



References

Cockerham CC. 1971. Higher order probability functions of identity of alleles by descent. Genetics 69:235-246.

Csűrös M. 2014. Non-identifiability of identity coefficients at biallelic loci. Theoretical Population Biology 92:22-29.

Jacquard A. 1974. *Genetics of Populations.* Springer, New York.

Li CC, Horvitz DG (1953). Some methods of estimating the inbreeding coefficient. *Am J Hum Genet* **5**:107-117.

Weir BS, Cockerham CC. 1977. Two-locus theory in quantitative genetics. In: Proceedings of the International Conference on Quantitative Genetics, (Eds.) E. Pollak, O. Kempthorne and T.B. Bailey. Iowa State University Press, Ames, IA. pp 247-269.

Weir BS, Goudet J. 2017. A unified characterization for population structure and relatedness. Genetics 206:2085-2103.