INBREEDING AND KINSHIP

Kinship and Inbreeding

The *kinship* $\theta^i_{jj'}$ of individuals j, j' in population i is the probability an allele from j is ibd to an allele from j'.

The inbreeding F^i_j of individual j in population i is the probability the two alleles in that individual are ibd.

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

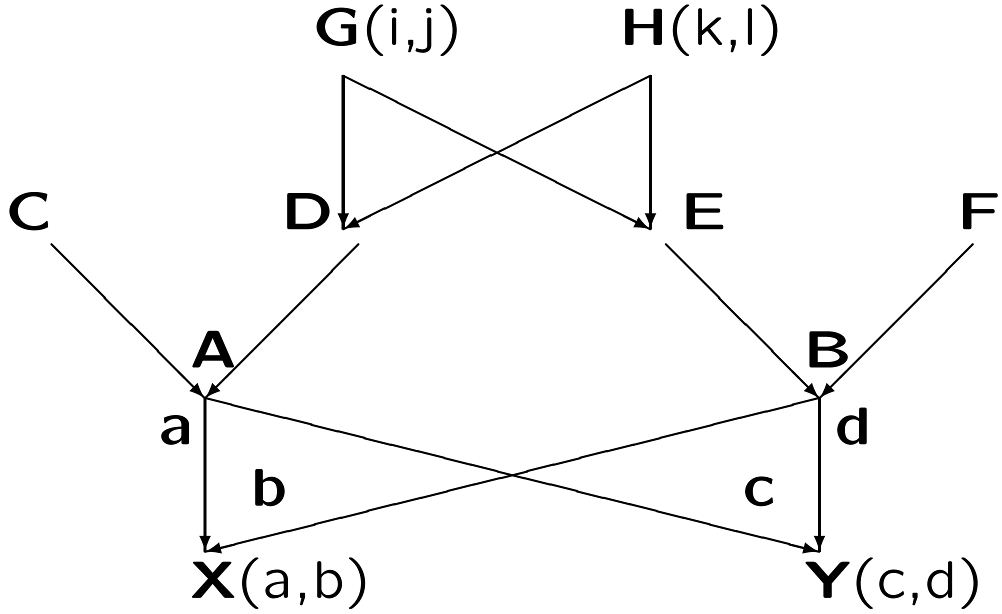
$$\theta^i_{jj} = \frac{1}{2}\left(1 + F^i_j\right)$$

If there are n individuals (including X, Y, A) in the path linking the parents through A, then

$$F_I = \theta_{XY} = \left(\frac{1}{2}\right)^n (1 + F_A)$$

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

Siblings whose parents are first cousins



What is the kinship coefficient for individuals X and Y?

Average Kinships

The average over all pairs of distinct individuals, $j \neq j'$, of the kinships $\theta^i_{jj'}$ is written as θ^i_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^i_W, where W means within populations. The average over populations is θ_W. *These are probabilities for distinct alleles.*

The ibd probability for any allele from population i and any allele from population i' is $\theta^{ii'}_B$, where B means between populations. Averaging over all pairs of distinct populations gives θ_B.

Within-population Inbreeding: F_{IS}

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

$$\beta_j^i = \frac{F_j^i - \theta_S^i}{1 - \theta_S^i}$$

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

If the reference set of alleles is for pairs of individuals within populations, averaged over populations, then the average relative inbreeding coefficient is $\beta_{IS} = (F_I - \theta_S)/(1 - \theta_S)$ where F_I is the average of F_j^i over individuals j and populations i. It is generally called F_{IS}.

Total Inbreeding: F_{IT}

For population i, the inbreeding coefficient for individual j, *relative to* the identity of pairs of alleles from different populations averaged over all pairs of populations, is

$$\beta_j^i \;=\; \frac{F_j^i - \theta_B}{1 - \theta_B}$$

The average over individuals within this population is the population-specific F_{IT}^i. The average of these over all populations is the total inbreeding coefficient $F_{IT} = (F_I - \theta_B)/(1 - \theta_B)$.

Within-population Kinship

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

$$\beta^i_{jj'} = \frac{\theta^i_{jj'} - \theta^i_S}{1 - \theta^i_S}$$

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

If the reference set is all pairs of alleles, one from each of two populations,

$$\beta^i_{jj'} = \frac{\theta^i_{jj'} - \theta_B}{1 - \theta_B}$$

The average β^i_{ST} over all pairs of individuals in population i is the population-specific F^i_{ST}, and averaging this over populations gives the global $F_{ST} = (\theta_S - \theta_B)/(1 - \theta_B)$. It is the ibd probability between individuals within populations relative to the ibd probability between populations.

Genotypic Measures

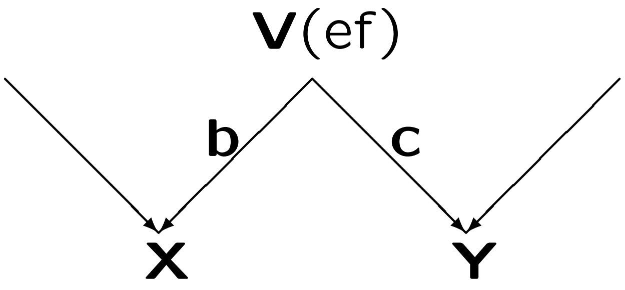
When individuals are distinguished:

$$\begin{aligned}(1-F_{IT}) &= (1-F_{IS})(1-F_{ST})\\ F_{IS} &= \frac{F_{IT}-F_{ST}}{1-F_{ST}}\end{aligned}$$

This classic result also holds for population-specific values

$$\begin{aligned}(1-F^i_{IT}) &= (1-F^i_{IS})(1-F^i_{ST})\\ F^i_{IS} &= \frac{F^i_{IT}-F^i_{ST}}{1-F^i_{ST}}\end{aligned}$$

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 X, Y has an expected value 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 WG, Weir BS. 2011. Genet Res 93:47-74

Allele Matching Approach

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

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

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

These are compared to the average matching for all pairs of individuals: $\tilde{M}_S$ for all pairs in the same sample or $\tilde{M}_B$ for all pairs from different samples.

Allele Matching

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

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

The estimates for inbreeding and kinship are

$$\hat{\beta}_j = \frac{\tilde{M}_j - \tilde{M}_S}{1 - \tilde{M}_S} \quad , \quad \hat{\beta}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_S}{1 - \tilde{M}_S}$$

Combine over SNPs as the ratio of averages

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

Allele Matching

The estimates behave well for estimating the parameters, as expected from Ochoa and Storey:

$$\beta_j = \frac{F_j - \theta_S}{1 - \theta_S} \quad , \quad \beta_{jj'} = \frac{\theta_{jj'} - \theta_S}{1 - \theta_S}$$

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

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

Ochoa A, Storey JD. 2019. bioRxiv https://doi.org/10.1101/083923. First published 2016-10-27.

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 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?

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

For large sample sizes, this reduces to

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

In other words, $\widehat{f}_{\mathrm{Hom}_j}$ is an (almost) unbiased estimate of $\beta_j = (F_j-\theta_S)/(1-\theta_S)$, the individual-specific version of Wright's F_{IS}.

Averaging over individuals gives the usual estimate for $f = F_{IS}$ for the population, and $F_{IS} = (F_{IT}-F_{ST})/(1-F_{ST})$.

Wright S. 1922. Am Nat

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

Expectations of allele frequencies in a sample of n individuals:

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

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: Standard

If X_{jl}, the allele dosage, is the number of copies of the reference allele for SNP l carried by individual j, Yang et al. 2011. Am J Hum Genet introduced $\widehat{F}^{III}$, called $\widehat{F}_{\text{Uni}}$ by Yengo et al. and f_{GCTA3} by Gazal et al:

$$\widehat{F}^{u}_{\text{Uni}_j} \; = \; \frac{1}{L}\sum_{l=1}^{L}\left(\frac{X_{jl}^2-(1+2\tilde{p}_l)X_{jl}+2\tilde{p}_l^2}{\tilde{p}_l(1-\tilde{p}_l)}\right)$$

For large samples this has an expected value under genetic sampling of

$$\mathcal{E}(\widehat{F}_{\text{Uni}_j}) \; = \; \frac{F_j-2\psi_j+\theta_S}{1-\theta_S}$$

where ψ_j is the average kinship of individual j with other members of the study sample,

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

Alternative Estimators: Standard

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

Yang et al. also discussed

$$\text{STD}_j = \frac{1}{L}\sum_{l=1}^{L}\frac{(X_{jl}-2\tilde{p}_l)^2}{2\tilde{p}_l(1-\tilde{p}_l)} - 1$$

For large samples, these estimates have expected values

$$\mathcal{E}(\text{STD}_j) = \frac{F_j - 4\psi_j + 3\theta_S}{1-\theta_S}$$

This has behavior close to that of $\widehat{F}_{\text{Uni}_j}$.

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

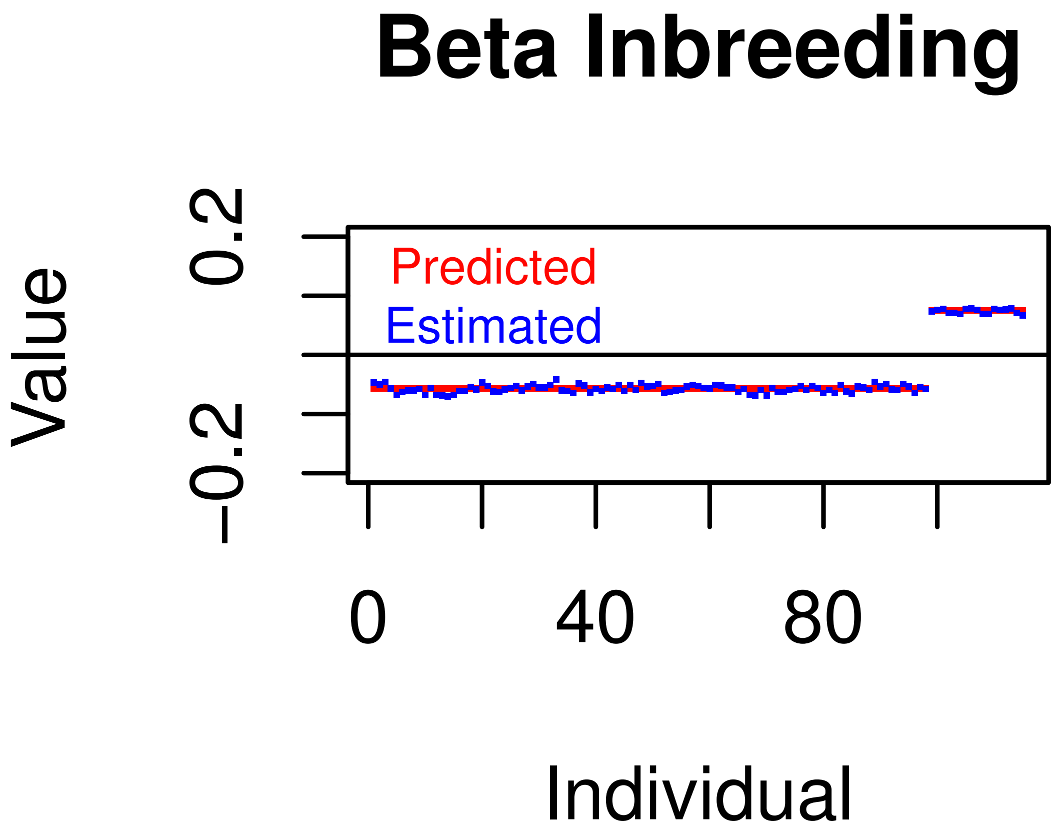
Example

The β inbreeding estimator was applied to a set of 115 individuals simulated and typed at 79,069 polymorphic SNPs Weir BS, Goudet J. 2017. Genetics.

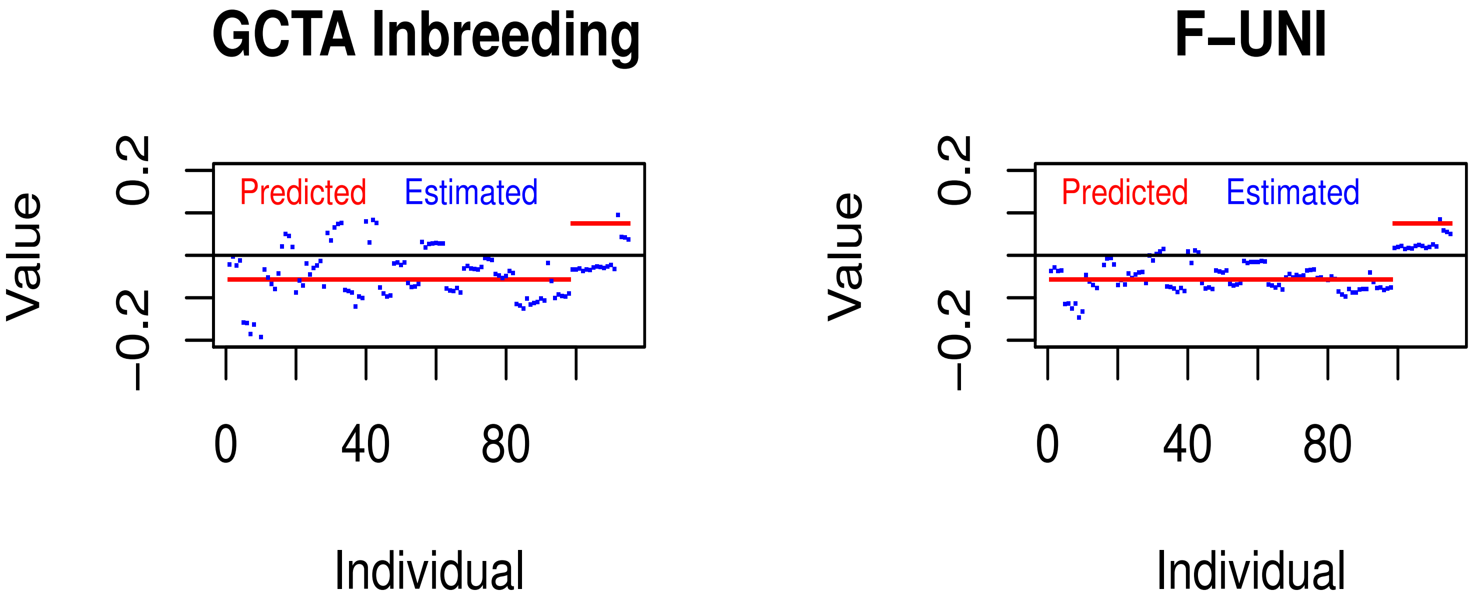
Among the 6,555 pairs of individuals the kinship values have an average value of $\theta_S = 0.0427$. There are 17 individuals with values of $F = 0.125, \beta = 0.0860$ and 98 with $F = 0, \beta = -0.0446$ predicted from the pedigree.

The $\widehat{\beta}_j$ values are very close to the $\beta_j = (F_j - \theta_S)/(1 - \theta_S)$ values, as shown on the next slide:

Example: Beta values



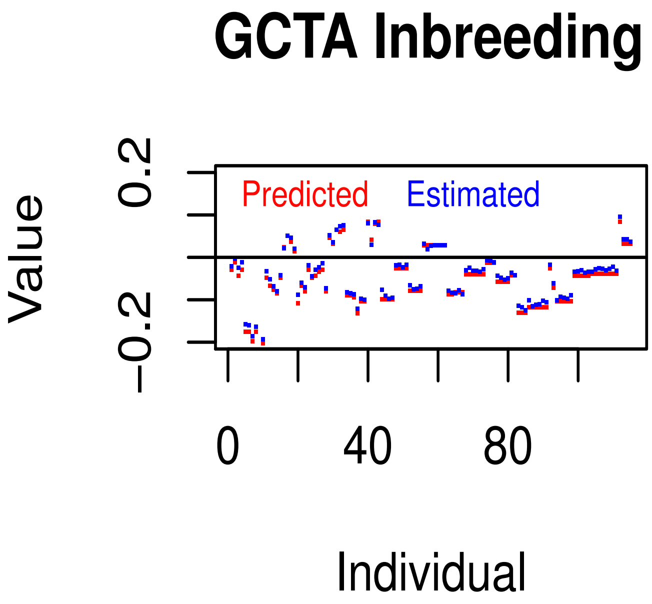
Example: Standard values

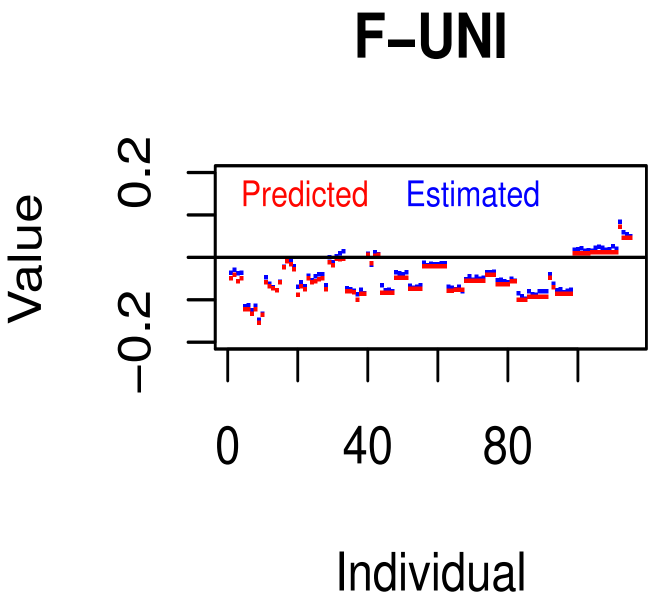


The problem is that these estimates use $\tilde{p}$'s instead of π's.

Example: Standard Expected values

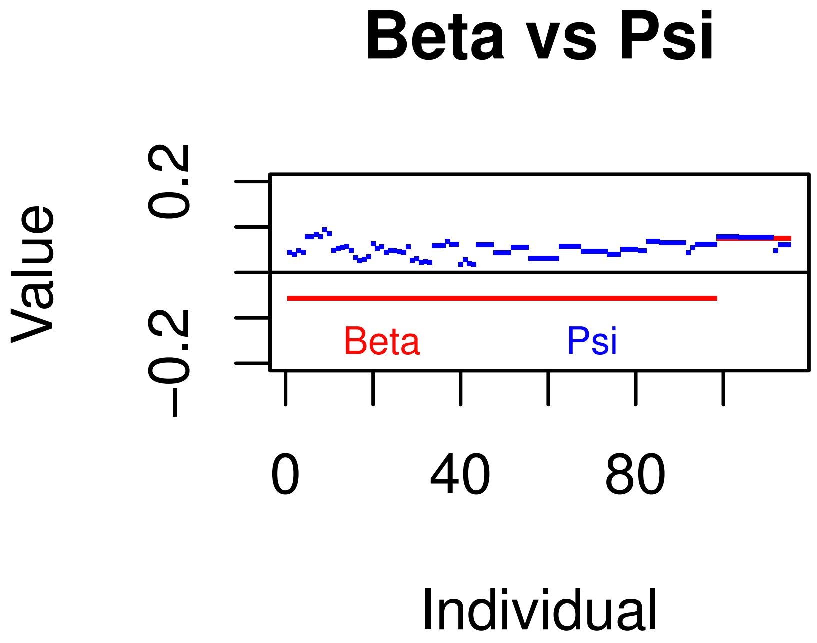
The GCTA estimators are close to their expected vales, but not to F or to β.





Example: Beta vs Psi

Individuals with the same F_j will have the same β_j but can have quite different ψ_j values:



Comparison of Estimators: Simulations

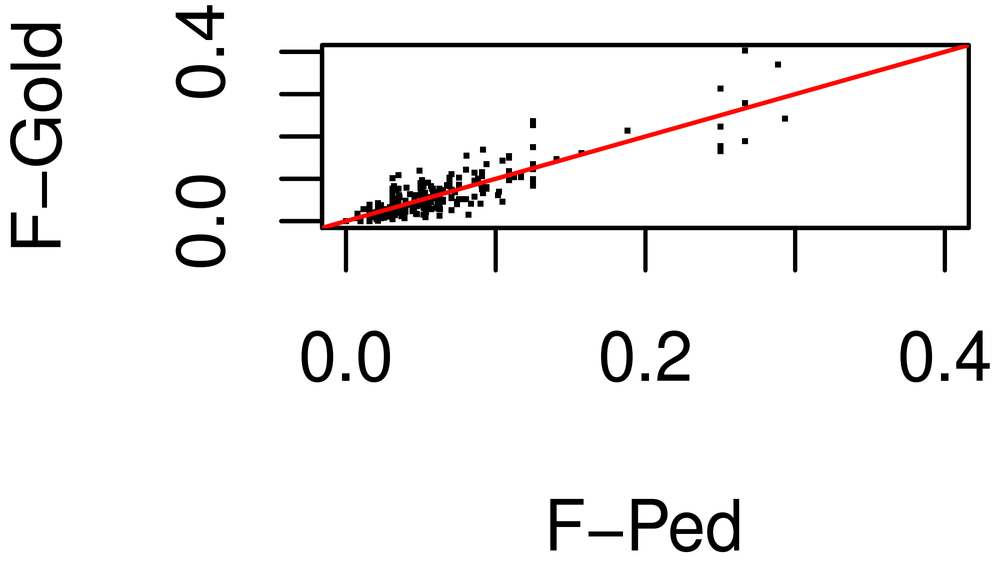
Simulation of 50 founder individuals, with 100,000 SNPs over a 20 Morgan map.

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

100K SNPs



Comparison of Estimators: Notation

Fped, Bped: pedigree values of F and β.

Fgold, Bgold: actual values of F and β.

Froh: runs of homozgosity estimate.

Fmle: maximum likelihood estimate of F.

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

Fbet: allele-matching estimates of β,

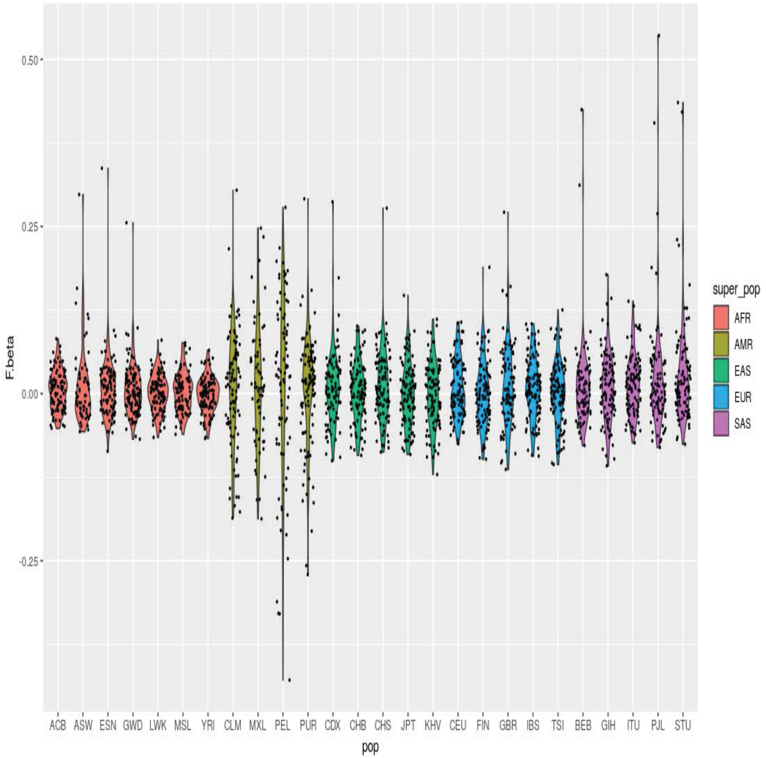
Ugold: actual value of F_{Uni}.

Funi: GCTA estimates of F_{Uni}.

Comparison of Estimators: Correlations

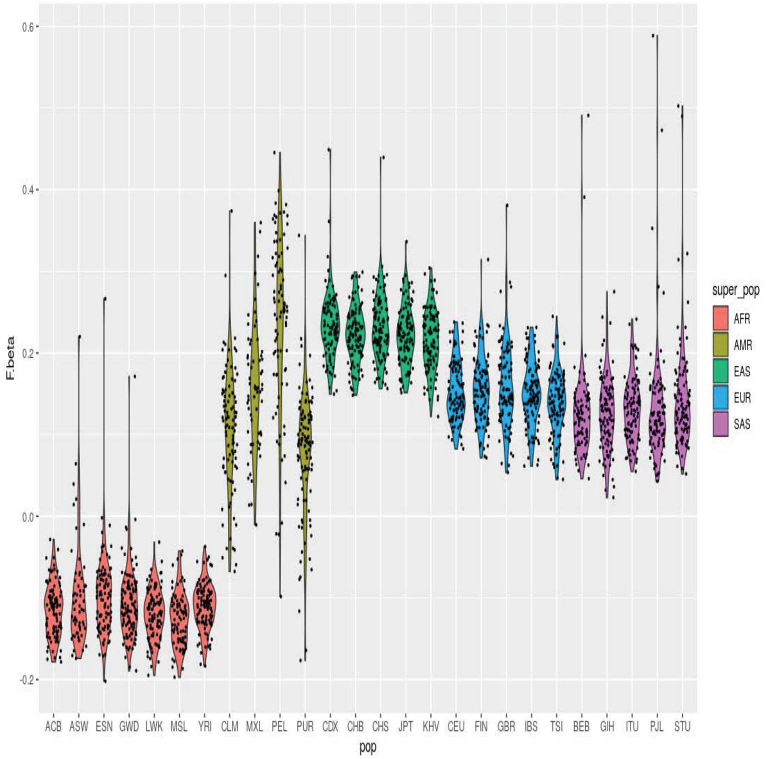
```
       Fped  Bped Fgold Bgold  Froh  Fmle  Fhom  Fbet  Ugold  Funi
Fped  1.000 1.000 0.902 0.901 0.879 0.790 0.836 0.836  0.707  0.642
Bped  1.000 1.000 0.902 0.902 0.879 0.790 0.836 0.836  0.707  0.642
Fgold 0.902 0.902 1.000 1.000 0.975 0.889 0.918 0.918  0.829  0.743
Bgold 0.901 0.902 1.000 1.000 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
Fbet  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
Funi  0.642 0.642 0.743 0.743 0.779 0.876 0.781 0.781  0.917  1.000
```

Inbreeding is Relative: Not Absolute



Local Population Reference

$$\frac{\sum_{[}(\tilde{M}_{jl}-\tilde{M}_{Sl})}{\sum_{l}(1-\tilde{M}_{Sl})}$$



Whole World Reference

$$\frac{\sum_{[}(\tilde{M}_{jl}-\tilde{M}_{Bl})}{\sum_{l}(1-\tilde{M}_{Bl})}$$

Chromosome 22 data from 1000 Genomes.
Continents (left to right): AFR, AMR, EAS, EUR, SAS

Estimation of Kinship

Estimation of Kinship

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

$$\hat{\beta}_{jj'} = \frac{\tilde{M}_{jj'} - \tilde{M}_R}{1 - \tilde{M}_R}$$

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

- if R is all pairs of individuals in the same sample, $\tilde{M}_R$ is the average matching 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, $\tilde{M}_R$ is the average matching 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, $\tilde{M}_R$ is the average matching 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

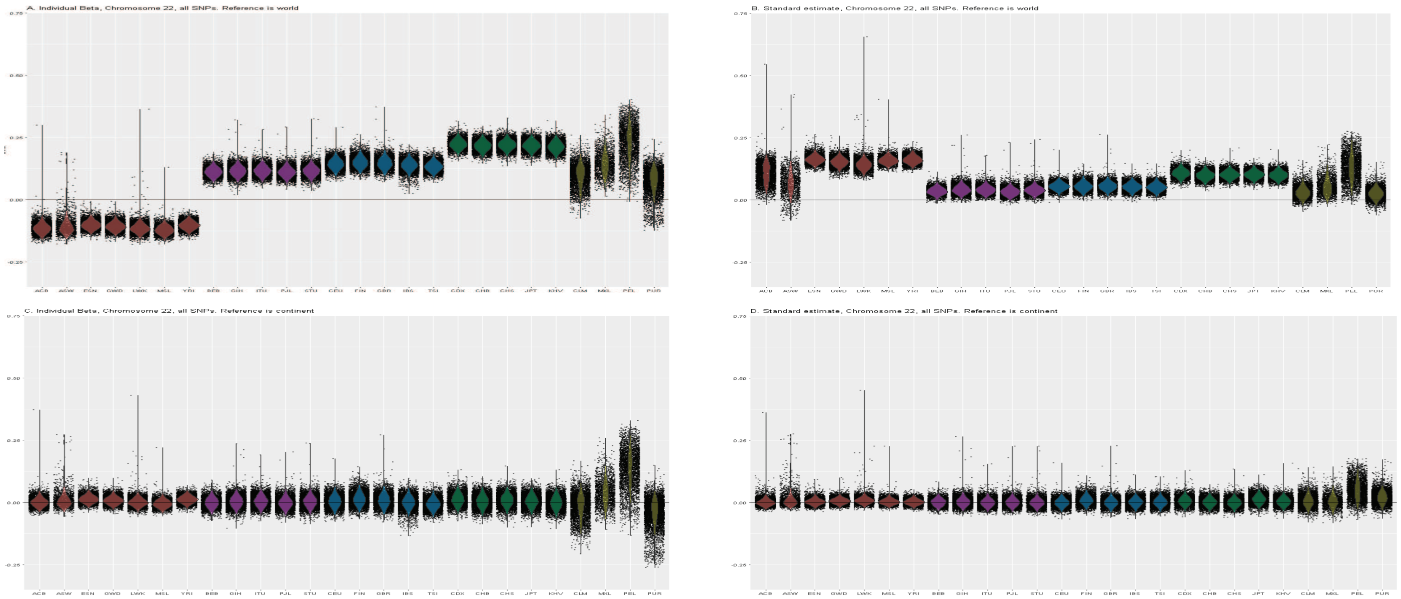
The β kinship estimates have been applied to 1000 Genomes data, and compared to standard estimates, shown on next slide.

For the whole world, all 26 populations, as reference the β 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.



Beta 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