A unified characterization of population structure and relatedness.

Bruce S. Weir[1] and Jérôme Goudet[2,3]

[1] Department of Biostatistics, University of Washington, Seattle WA, USA

[2] Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

[3] Swiss Institute of Bioinformatics, Lausanne, Switzerland

November 16, 2016

Abstract

Many population genetic activities, ranging from evolutionary studies to association mapping to forensic identification, rely on appropriate estimates of population structure or relatedness. All applications require recognition that quantities with an underlying meaning of allelic identity by descent are not defined in an absolute sense, but instead are made "relative to" some set of alleles other than the target set. The early Weir and Cockerham F_{ST} estimate made explicit that the reference set of alleles was across independent populations. Standard kinship estimates have an implicit assumption that pairs of individuals in a study sample, other than the target pair, are unrelated, whereas other estimates assume alleles within individuals are not identical by descent. However, populations lose independence when there is migration between them, and when individuals in a study are related it is difficult to see how they can also be non-inbred. We have therefore re-cast our treatments of population structure, relatedness and inbreeding to make explicit that the parameters of interest involve differences of probabilities of identity by descent in the target and the reference sets of alleles and so can be negative. We take the reference set to be for the population from which study individuals have been sampled. We provide simple moment estimates of these parameters, phrased in terms of allele matching within and between individuals for relatedness and inbreeding, or within and between populations for population structure. A multi-level hierarchy of alleles within individuals, alleles between individuals within populations, and alleles between populations allows a unified treatment of relatedness and population structure. Our new estimates appear to be sensitive to rare or private variants, to give indications of the effects of natural selection, and to be appropriate for use in association studies.

Keywords Allele matching, Fst, Identity by descent, Rare variants

Introduction

We offer here a unified treatment of relatedness and population structure with an underlying framework of alleles being identical by descent, ibd. We follow Thompson (2013) in regarding ibd for a set of alleles as being relative to some other, reference, set: "There is no absolute measure of ibd: ibd is always relative to some reference population." In other words, ibd implies a reference point and ibd status there is often implicitly assumed to be zero.

A function of ibd of particular interest to us is F_{ST}, which we will show below depends on ibd of pairs of alleles within populations relative to that for pairs of alleles from different populations. The uses of estimates of this quantity are widespread, and here we note a recent discussion by McTavish and Hillis (2015) of the effects of SNP ascertainment, SNP array vs whole-genome sequencing, on inferences about population history. These authors used "pairwise F_{ST} for all pairs of populations using Weir and Cockerham's method." We suggest that a more informative analysis may result from our population-specific F_{ST} estimates (Weir and Hill, 2002; Weir et al., 2005; Browning and Weir, 2010). Several authors (e.g. Balding and Nichols, 1995; Shriver et al., 2004; Beaumont and Balding, 2004; Gaggiotti and Foll, 2010) have discussed the advantages of working with population-specific F_{ST} values instead of single values for a set of populations or of values for each pair of populations. We show below that the usual global F_{ST} measure can be regarded as an unweighted average of population-specific values, and because it is an average it collapses the variation detectable among populations that can indicate the effects of past selection (Weir et al., 2005). The usual measure can otherwise diminish signals of population history and this diminution has become more pronounced as genetic marker data have become richer and real differences among populations have become more evident. As Astle and Balding (2009) noted "population structure and [cryptic] relatedness are different aspects of a single confounder: the unobserved pedigree defining the (often distant) relationships among the study subjects." A similar point was made by Kang et al. (2010): "The presence of related individuals within a study sample results in sample structure, a term that encompasses population stratification and hidden relatedness." Our goal is to provide a unified approach to charactering population structure and individual relatedness and inbreeding, both in terms of the underlying parameters and of the methods of estimation.

A consideration of "genetic sampling" (Weir, 1996) makes it clear that population mean ibd

for alleles in a single population, or ibd for alleles in a single individual, at one point in time cannot be estimated only from data for that population or that individual as there would be no information about variation over replications of the descent paths from past to present. We might regard multiple loci as providing replication of the genetic sampling process or we might collect data from multiple populations. An exception is when allele frequencies and ibd status in the reference population are assumed known, as is implied for standard methods for estimating relatedness and inbreeding (e.g. Ritland, 1996; Wang, 2014; Purcell et al., 2006; Yang et al., 2011). If, instead, these methods make use of frequencies from a sample of individuals they are providing estimates of the inbreeding or coancestry ibd measures relative to those measures for individuals in the whole sample. This point was also made by Yu et al. (2006) who spoke of "adjusting the probability of identity by state between two individuals with the average probability of identity by state between random individuals" in order to address identity by descent. Other relatedness estimation methods that do not use allele frequencies (e.g. KING-robust, Manichaikul et al., 2010) are estimating ibd between individuals (coancestry) relative to that within individuals (inbreeding: assumed zero for KING-robust).

For both population structure and relatedness we propose the use of allelic matching proportions within and between individuals or populations in order to characterize ibd for an individual or a population relative to a reference set of ibd values. We use allele matching rather than heterozygosity (Nei, 1973) or components of variance (Weir and Cockerham, 1984: hereafter WC84) although the distinction is more semantic than real. Our present treatment also differs from that in WC84 by using unweighted averages of statistics over populations instead of the weighted averages that were more appropriate for the WC84 model of independent populations.

The size of current genetic studies requires computationally feasible methods for estimating relatedness between all pairs of individuals, potentially 5 billion pairs for the TOPMed project (http://www.nhlbiwgs.org). The scale of the task may well rule out maximum likelihood approaches (e.g. Thompson, 1975; Ritland, 1996; Milligan, 2003) and Bayesian methods (e.g. Gaggiotti and Foll, 2010). Moment estimates seem still to be relevant and will be presented here.

Materials and Methods

Parameter Values

We write the probability of a pair of alleles being ibd as θ without specifying the reference set of alleles. Subscripts will identify the populations or individuals from which the alleles are drawn. Subscript W will indicate an average over sets of alleles within individuals or populations, and subscript B the average over pairs of distinct individuals or populations.

Pairs of Individuals The coancestry coefficient θ_{XY} for individuals X, Y is the probability an allele taken randomly from X is ibd to one taken randomly from Y. If individual A is ancestral to both X and Y, and if there are n individuals in the pedigree path joining X to Y through A, then $\theta_{XY} = \sum(0.5)^n(1+F_A)$ where F_A is the inbreeding coefficient of A and the sum is over all ancestors A and all paths joining X to Y through A. The coancestry of X with itself is $\theta_{XX} = (1+F_X)/2$.

Pairs of Populations For populations i, i' the quantity $\theta_{ii'}$ is the probability of ibd for an allele from i and one from i'. The two populations may be the same, $i = i'$. The simplest evolutionary scenario is for finite populations of constant size, subject only to genetic drift. For two populations of sizes N_1, N_2 with a common ancestral population t discrete generations in the past, the current within-population values θ_{11}, θ_{22} and the between-population value θ_{12} are

$$\begin{aligned} \theta_{ii}(t) &= 1 - [1 - \theta_{12}(0)] \left(1 - \frac{1}{2N_i}\right)^t, \quad i = 1, 2 \\ \theta_{12}(t) &= \theta_{12}(0) \end{aligned}$$

The between-population ibd probability $\theta_{12}(t)$ at present is the same as it was, $\theta_{12}(0)$, in the common ancestral population. As we do throughout this discussion, we introduce quantities β that measure ibd for a target pair of alleles relative to that in a convenient comparison set, here alleles between the pair of current populations (also, in this case, the common ancestral population):

$$\beta_{ii}(t) = \frac{\theta_{ii}(t) - \theta_{12}(t)}{1 - \theta_{12}(t)} = 1 - \left(1 - \frac{1}{2N_i}\right)^t, \quad i = 1, 2$$

By considering ibd relative to that between populations, we avoid having to know the value in the ancestral population or even having to specify that ancestral population. The two population-specific values $\beta_{ii}(t)$ differ if the two populations have different sizes. We often wish to work with

the averages $\theta_W(t) = [\theta_{11}(t) + \theta_{22}(t)]/2, \theta_B(t) = \theta_{12}(t)$ and we write $\beta_W(t) = [\beta_{11}(t) + \beta_{22}(t)]/2$:

$$\beta_W(t) \quad = \quad \frac{\theta_W(t) - \theta_B(t)}{1 - \theta_B(t)} \approx \frac{t}{2N_h}$$

providing N_1, N_2, t are all large, and N_h is the harmonic mean of the population sizes. From now on we will regard β_W as the parametric form of F_{ST}, and Reynolds et al. (1983) showed that F_{ST} serves as a measure of population distance under the pure drift model. The formulation $F_{ST} = (\theta_W - \theta_B)/(1 - \theta_B)$ makes explicit that F_{ST} is a measure of ibd within populations relative to ibd between *pairs of* populations.

Drift, Mutation and Migration Non-trivial equilibria for two populations drifting apart are obtained when there is mutation, and we illustrate some aspects of our population-specific approach by considering the case of two populations exchanging alleles each generation when there is infinite-alleles mutation. The transition equations for $\theta_1, \theta_2, \theta_{12}$, extending those given by Maruyama (1970), are:

$$\begin{aligned}
\theta_{ii}(t+1) &= (1-\mu)^2 \left[(1-m_i)^2\theta^*_{ii}(t) + 2m_i(1-m_i)\theta_{12}(t) + m_i^2\theta^*_{i'i'}(t)\right], \quad i = 1, 2; i' \neq i \\
\theta_{12}(t+1) &= (1-\mu)^2 \left[(1-m_1)m_2\theta^*_{11}(t) + [(1-m_1)(1-m_2) + m_1m_2]\theta_{12}(t) + m_1(1-m_2)\theta^*_{22}(t)\right]
\end{aligned}$$

where $\theta^*_{ii}(t) = 1/(2N_i) + (2N_i - 1)\theta_{ii}(t)/(2N_i)$, the mutation rate is μ and population $i : i = 1, 2$ receives a fraction m_i of its alleles each generation from population $i' : i' \neq i$. A consequence of these equations is that $\theta_{11}(t) + \theta_{22}(t) \geq 2\theta_{12}(t)$, or that $\theta_W \geq \theta_B$ and so $\beta_W = F_{ST}$ is positive. However, it is not necessary that each of θ_{11}, θ_{22} exceeds θ_{12} and in Figure 1, second row, we show that mutation leads to equilibrium values of θ_{ii} different from 1, and in Figure 1, third row that migration can lead to cases where $\theta_{11} > \theta_{12} > \theta_{22}$. In the absence of migration, mutation drives θ_{12} and β_{12} to zero, so that $\theta_{ii} = \beta_{ii}$ are both positive. In Figure 2 we show the region in the space of N_1, m_1 values where $\beta_{11} \leq 0 \leq \beta_{22}$ for fixed N_2, m_2, μ. Averaging over the two β_{ii}'s to work with F_{ST} hides this potential difference in sign of the β_{ii}'s.

[Figure 1]

[Figure 2]

Actual vs Predicted θ The probabilities of ibd calculated from path-counting methods for pedigrees of individuals or from transition equations for populations can be regarded as the expected values, over evolutionary replicates, of the actual identity status of a pair of alleles. We have previously discussed the variation of actual identity about the predicted value (Hill and Weir, 2011, 2012). The variance of the actual ibd measure $\ddot{\theta}$ for two alleles is $\Delta - \theta^2$ (Cockerham and Weir, 1983), where Δ is the joint probability of ibd for each of two pairs of alleles. The coefficient of variation of the actual coancestry $\ddot{\theta}$ for two individuals is greater than 1 for individuals with predicted coancestry θ less than 0.125 and it increases as the degree of relationship decreases. The implication of this is that, for a particular pair of populations or individuals, estimated values may not match those expected from pedigrees or transition equations. Evaluation of estimation procedures should, therefore, be performed over many replicate pairs.

Inbred Populations The discussion so far has implicitly assumed Hardy-Weinberg equilibrium within populations and no need to differentiate pairs of alleles within individuals from those between individuals in the same population. We can relax that assumption. Two alleles taken at random from population i may be from the same or different individuals. If the sampling was without replacement, the ibd probability for two alleles from one individual is the inbreeding coefficient F for that individual, whereas sampling with replacement from one individual has ibd probability $(1 + F)/2$. We defer a more extensive discussion to a subsequent publication.

Estimation

Allele Frequencies It is allelic identity in state (ibs) that can be observed, rather than identity by descent (ibd) and we now consider how ibs data can provide ibd estimates. We start by considering the frequencies of the various alleles at a locus.

We can distinguish three classes of allele frequency. The sample frequency $\tilde{p}_{iu}$ of allele u in a sample of alleles taken from population i provides an estimate of the actual allele frequency $\ddot{p}_{iu}$ in that population. These actual frequencies, in turn, vary about the population frequency p_u, where variation refers to the values of the actual frequencies $\ddot{p}_{iu}$ in evolutionary replicates of population i.

For n_i randomly sampled alleles, where n_{iu} are of type u, $n_{iu} = n_i\tilde{p}_{iu}$ has a binomial distribution

$B(n_i, \ddot{p}_{iu})$. The mean of $\tilde{p}_{iu}$ from this statistical sampling process is $\mathcal{E}_S(\tilde{p}_{iu}) = \ddot{p}_{iu}$ and the variance is $\mathrm{Var}_S(\tilde{p}_{iu}) = \ddot{p}_{iu}(1-\ddot{p}_{iu})/n_i$. The distribution of the actual frequency $\ddot{p}_{iu}$ is not known in general, but there is a class of evolutionary models (e.g. Balding and Nichols, 1995) that provide the Beta distribution:

$$\ddot{p}_{iu} \sim \mathrm{Beta}\left(\frac{(1-\theta_i)p_u}{\theta_i}, \frac{(1-\theta_i)(1-p_u)}{\theta_i}\right)$$

The mean for this genetic sampling process is $\mathcal{E}_G(\ddot{p}_{iu}) = p_{iu}$ and the variance is $\mathrm{Var}_G(\tilde{p}_{iu}) = p_u(1-p_u)\theta_i$. We keep these first two genetic-sampling moments although we do not invoke the beta distribution.

The total mean and variance follow from

$$\begin{aligned}
\mathcal{E}_T(\tilde{p}_{iu}) &= \mathcal{E}_G[\mathcal{E}_S(\tilde{p}_{iu})] \\
\mathrm{Var}_T(\tilde{p}_{iu}) &= \mathcal{E}_G[\mathrm{Var}_S(\tilde{p}_{iu})] + \mathrm{Var}_G[\mathcal{E}_S(\tilde{p}_{iu})]
\end{aligned}$$

and the complete set of first and second moments were given by Weir and Hill (2002):

$$\begin{aligned}
\mathcal{E}_T(\tilde{p}_{iu}) &= p_u \\
\mathrm{Var}_T(\tilde{p}_{iu}) &= p_u(1-p_u)\left(\theta_{ii} + \frac{1-\theta_{ii}}{n_i}\right) \\
\mathrm{Cov}_T(\tilde{p}_{iu}, \tilde{p}_{iu'}) &= -p_u p_{u'}\left(\theta_{ii} + \frac{1-\theta_{ii}}{n_i}\right), \quad u' \neq u \\
\mathrm{Cov}_T(\tilde{p}_{iu}, \tilde{p}_{i'u}) &= p_u(1-p_u)\theta_{ii'}, \quad i' \neq i \\
\mathrm{Cov}_T(\tilde{p}_{iu}, \tilde{p}_{i'u'}) &= -p_u p_{u'}\theta_{ii'}, \quad u' \neq u, i' \neq i
\end{aligned} \tag{1}$$

These moments refer to expectations over both repeated samples from the same populations (statistical sampling) and over replications of the populations themselves (genetic sampling). From now on we will drop the T subscript but all expectations are total. WC84 set all within-population θ_{ii} to a common value θ, and all between-population $\theta_{ii'}$ to zero. Note the assumption that all populations have the same expected allele frequencies p_u, although they have different actual frequencies $\ddot{p}_{iu}$.

Allelic Matching We find intuitive appeal in working with proportions of pairs of alleles that are ibs. If the sample of n_i alleles from population i has n_{iu} copies of allele type u, then the matching (allele sharing) proportion for pairs of alleles drawn without replacement from population i is $\tilde{M}_{ii}$

where

$$\tilde{M}_{ii} = \frac{1}{n_i(n_i-1)}\sum_u n_{iu}(n_{iu}-1) = \frac{n_i}{n_i-1}\sum_u \tilde{p}_{iu}^2 - \frac{1}{n_i-1}$$

For sampling with replacement the sample-size corrections are not necessary: $\tilde{M}_{ii} = \sum_u \tilde{p}_{iu}^2$. The average over samples from r populations is $\tilde{M}_W = \sum_i \tilde{M}_{ii}/r$. The allele-pair matching proportion between populations i and i' is

$$\tilde{M}_{ii'} = \frac{1}{n_i n_{i'}}\sum_u n_{iu}n_{i'u} = \sum_u \tilde{p}_{iu}\tilde{p}_{i'u}$$

and these have an average over pairs of samples from r populations of $\tilde{M}_B = \sum_{i\neq i'} \tilde{M}_{ii'}/[r(r-1)]$.

Population Structure Estimates From Equations 1, the matching proportions have expectations

$$\mathcal{E}(\tilde{M}_{ii}) = 1 - H(1-\theta_i) \quad , \quad \mathcal{E}(\tilde{M}_{ii'}) = 1 - H(1-\theta_{ii'})$$

where $H = 1 - \sum_u p_u^2$. Averaging over populations or pairs of populations:

$$\mathcal{E}(\tilde{M}_W) = 1 - H(1-\theta_W) \quad , \quad \mathcal{E}(\tilde{M}_B) = 1 - H(1-\theta_B)$$

The expectations lead immediately to simple method-of-moment estimates for any number of sampled populations, any number of alleles sampled per population, and any numbers of alleles per locus:

$$\hat{\beta}_{ii} = \frac{\tilde{M}_{ii} - \tilde{M}_B}{1-\tilde{M}_B}, \quad \hat{\beta}_W = \frac{\tilde{M}_W - \tilde{M}_B}{1-\tilde{M}_B}, \quad \hat{\beta}_{ii'} = \frac{\tilde{M}_{ii'} - \tilde{M}_B}{1-\tilde{M}_B} \tag{2}$$

To the extent that the expectation of a ratio is the expectation of ratios, Equations 1,2 show that each $\hat{\beta}$ is unbiased for the corresponding β:

$$\mathcal{E}(\hat{\beta}_{ii}) = \frac{\theta_{ii} - \theta_B}{1-\theta_B}, \quad \mathcal{E}(\hat{\beta}_W) = \frac{\theta_W - \theta_B}{1-\theta_B}, \quad \mathcal{E}(\hat{\beta}_{ii'}) = \frac{\theta_{ii'} - \theta_B}{1-\theta_B}$$

Note that the pairwise estimates $\hat{\beta}_{ii'}, i' \neq i$ sum to zero by construction. Although it is not possible to find estimates for each θ when the sampled populations have correlated sample allele frequencies, it is possible to rank the $\hat{\beta}$'s, and these are likely to have the same ranking as the θ's. We now show how this approach also gives estimates of individual inbreeding coefficients and individual-pair coancestry coefficients.

Relatedness Estimates Suppose now we have a series of individuals $i; i = 1, 2, \ldots r$ and we sample two alleles with replacement from an individual or one allele randomly from each of two individuals. Each individual is regarded as a population with its own actual allele frequencies. Setting the sample sizes to 2 in the matching proportions for one individual and to 1 each for pairs of individuals in the previous section leads to estimates of $\theta_{ii} = (1 + F_i)/2$ or $\theta_{ii'}$ for individual i or individuals i, i'. The expectations of these estimates are $[(1 + F_i)/2 - \theta_B]/(1 - \theta_B)$ and $(\theta_{ii'} - \theta_B)/(1 - \theta_B)$, respectively, where θ_B is the average of all $\theta_{ii'}, i \neq i'$. Inbreeding and coancestry are estimated relative to the average coancestry of all pairs of individuals in the study. Yang at al. (2010) also discuss estimates relative to the study population, and say "Estimates of relationships are always relative to an arbitrary base population in which the average relationship is zero. We use the individuals in the sample as the base so that the average relationship between all pairs of individuals is 0 and the average relationship of an individual with him- or herself is 1." Although our estimates of pairwise relationship sum to zero, we retain the unknown value θ_B in their expectations. We cannot estimate θ_B and we may prefer to report estimates relative to those for the least related pairs as described below in Equation 6.

It is customary (e.g. Yang et al., 2011) use allelic dosage to express relatedness estimates or other analyses (Patterson et al., 2006). Writing the number of copies of allele u carried by individual i as x_{iu}, the U-allele versions of these standard estimates are

$$\hat{\theta}_{ii'} = \frac{\sum_{u=1}^{U}(x_{iu} - 2p_u)(x_{i'u} - 2p_u)}{4\sum_{u=1}^{U} p_u(1 - p_u)} \quad \text{or} \quad \hat{\theta}_{ii'} = \frac{1}{U}\sum_{u=1}^{U} \frac{(x_{iu} - 2p_u)(x_{i'u} - 2p_u)}{4p_u(1 - p_u)} \tag{3}$$

where i, i' may be the same or different. If the allele frequencies p_u are known, these estimates are unbiased for $\theta_{ii'}$. For biallelic SNPs, there is no need to sum over alleles, and the u subscripts can be dropped.

Our coancestry estimates have the same functional form as those for population structure, but they may best be compared to the standard estimates by expressing allelic matching proportions in terms of allelic dosages. Noting that individual matching proportions are 1 and 0 for homozygotes and heterozygotes, respectively, and that matching proportions for pairs of individuals are 1 when they are the same homozygote, 0.5 when they are the same heterozygote or one is homozygous and the other heterozygous with one allele shared with the first, and 0 when they have no shared alleles:

$$\tilde{M}_{ii} = \frac{1}{2U}\sum_{u=1}^{U}[1 + (x_{iu} - 1)^2] \quad , \quad \tilde{M}_{ii'} = \frac{1}{2U}\sum_{u=1}^{U}[1 + (x_{iu} - 1)(x_{i'u} - 1)]$$

In particular, for SNPs,

$$\tilde{M}_{ii} = \frac{1}{2}[1 + (x_i - 1)^2] \quad , \quad \tilde{M}_{ii'} = \frac{1}{2}[1 + (x_i - 1)(x_{i'} - 1)]$$

where x_i are the dosages for, say, the reference allele. Our relatedness and inbreeding estimates are

$$\hat{\beta}_{ii} = \frac{\tilde{M}_{ii} - \tilde{M}_B}{1 - \tilde{M}_B} \quad , \quad \hat{\beta}_{ii'} = \frac{\tilde{M}_{ii'} - \tilde{M}_B}{1 - \tilde{M}_B} \tag{4}$$

where $\tilde{M}_B = \sum_{i=1}^{r} \sum_{i'=1, i' \neq i}^{r} \tilde{M}_{ii'}/r(r-1)$.

Storey and Ochoa (accompanying papers) have equivalent estimates. Their expressions are a little different because their reference is for all pairs of alleles in a sample, including those within individuals, whereas ours is for pairs of alleles in different individuals. Astle and Balding (2009, equation 2.3) gave similar estimates although, in effect, they set θ_B, the average coancestry of all pairs of individuals in a sample, to zero.

Combining Over Loci Single-locus analyses do not provide meaningful results, and combining estimates over loci l has often been considered in the literature. In a parallel discussion of weighting over alleles u at a single locus, Ritland (1996) considered weights w_u chosen to minimize variance.

Two extreme weights are $w_l = 1$ and $w_l = (1 - \tilde{M}_{B_l})$. The first may be called "unweighted" and the second "weighted". In an obvious notation

$$\hat{\beta}_i^{\,u} = \frac{1}{L} \sum_{l=1}^{L} \frac{\tilde{M}_{i_l} - \tilde{M}_{B_l}}{1 - \tilde{M}_{B_l}} \quad , \quad \hat{\beta}_i^{\,w} = \frac{\sum_{l=1}^{L} (\tilde{M}_{ii'_l} - \tilde{M}_{B_l})}{\sum_{l=1}^{L} (1 - \tilde{M}_{B_l})} \tag{5}$$

Note the parallel to averaging over alleles in Equations 3. Bhatia et al. (2013) refer to the first estimate as the "average of ratios" and the second as the "ratio of averages." WC84 advocated the second, with justification given in the Appendix to that paper, as did Bhatia et al.

The unweighted estimate $\hat{\beta}_i^{\,u}$ is unbiased for all allele frequencies but is susceptible to the effects of rare variants, when $(1 - \tilde{M}_{B_l})$ can be very small. Rare variants may have little effect on the weighted average $\hat{\beta}_i^{\,w}$, and the variance of the estimate is seen in simulations to be less than for the unweighted average, but it is unbiased only if every locus has the same value of the ibd probabilities. A more extensive discussion was given in the Appendix of WC84 for population structure, and by Ritland (1996) for inbreeding and relatedness.

Private Alleles Current sequence-based studies are revealing large numbers of low-frequency variants, including those found in only one population. These private alleles were identified by Slatkin (1985) and Mathieson and McVean (2012) as being of particular interest. They are very frequent in the 1000 genomes project data (The 1000 Genomes Project Consortium, 2010). If x_1 is the sample count of an allele observed only in population 1 of r populations ($\tilde{p}_1 = x_1/n_1; \tilde{p}_i = 0, i \neq 1$) the sample matching proportions are

$$\begin{aligned}
\tilde{M}_{ii} &= \begin{cases} 1 - 2\tilde{p}_1(1-\tilde{p}_1)\frac{n_1}{n_1-1} & i = 1 \\ 1 & i \neq 1 \end{cases} \\
\tilde{M}_{ii'} &= \begin{cases} 1 - \tilde{p}_1 & i = 1, i' \neq 1 \\ 1 & i, i' \neq 1, i \neq i' \end{cases} \\
\tilde{M}_B &= 1 - \frac{2\tilde{p}_1}{r}
\end{aligned}$$

so the β estimates are

$$\hat{\beta}_{11} = 1 - r(1-\tilde{p}_1)\frac{n_1}{n_1-1};\ \hat{\beta}_{1i} = \frac{1}{2}\tilde{p}_1(r-2), i \neq 1;\ \hat{\beta}_{ii'} = 1, i, i' \neq 1;\ \hat{\beta}_W = \tilde{p}_1$$

The estimate of F_{ST} for a private allele is its own-population sample frequency, but the population-specific value for its own population ranges from approximately $-r+1$ when $x_1 = 1$ to 1 when $x_1 = n_1$. This amplifies the comment "populations can display spatial structure in rare variants, even when Wright's fixation index F_{ST} is low" of Mathieson and McVean (2012). A population with many private alleles at low to intermediate frequencies will thus likely have a negative $\hat{\beta}$, and how negative will depend on how many populations have been sampled. Note that this implies $\hat{\beta}_{ii}$ must be allowed to go negative, whereas Bayesian estimators of population specific F_{ST} are forced to belong to $[0, 1]$, although this assumption can be relaxed (Ritland, 1996).

RESULTS

Population Structure

We have conducted a series of simulations to evaluate the performance of our F_{ST} estimates, and we have looked at 1000 Genomes SNP data to explore the role of rare variants on the estimates.

Some of the simulations were conducted with *sim.genot.metapop.t* available in the *hierfstat* package (Goudet, 2005). The migration model we have used allows for a matrix of migration rates between each pair of populations, and the mutation model allows for multiple alleles at a locus. The notation for a two-population model was given above.

Model 1. Same Migration Rates, Different Population Sizes. We considered two populations, with sizes $N_1 = 100, N_2 = 1,000$ and migration rates $m_1 = m_2 = 0.01$. The mutation rate was $\mu = 10^{-6}$. After 400 generations, the β's have values $\beta_1 = 0.156, \beta_2 = -0.037$ and $\beta_{12} = 0.059$. We simulated 50 individuals from each population under this scenario, with 1,000 loci and up to 20 alleles per locus. From the resulting allelic data we obtained estimates, and 95% confidence intervals by bootstrapping over loci. The results are shown in Table 1. The predicted values are contained in the confidence intervals, and there are negative values for both the parametric and the estimated value of β_2. Note that we cannot estimate β_{12} with data from two populations.

[Table 1]

Model 2. Continent-Island Model. In this scenario we have an infinite continent supplying a proportion $m = 0.01$ of the alleles independently to populations 1 and 2, still with sizes $N_1 = 100, N_2 = 1,000$. There is no migration between the two populations, so $\theta_{12} = 0$. The predicted values and estimated values after 400 generations are shown in Table 1.

Model 3. Migrant-pool Island Model. In this scenario, each population contributes to a migrant pool, from which migrant alleles are drawn. Among the migrant alleles in the case of two populations, half of the "migrant alleles" will in fact be resident alleles if the gametic pool is composed of the same proportion of alleles from each island, independent of its size. With otherwise the same parameter values, the predicted values and our estimates after 400 generations are shown in Table 1.

Model 4. Different Population Sizes, Different Migration Rates. We return to the two-populations model described above, but now with $N_1 = 10,000, N_2 = 100$ and different migration rates $m_1 = 0.01$, $m_2 = 0$. Predicted values after 1,000 generations are shown in Figure 3, and our estimates in Table 1.

The results in Table 1 show generally good behavior of our β estimates. In Figure 3 we show the estimates for 10 different time points (independent replicates) for Model 4. As time increased, the number of polymorphic loci decreased. In generations 600, 800, 1000 the numbers of polymorphic loci had dropped from 1,000 to 712, 349 and 151 respectively and the quality of the estimates decreased: higher bias and higher variance.

[Figure 3]

Rare Alleles. For r populations with total sample size n_T, and with x_1 copies of an allele private to population 1, the total count for this alleles is $x_T = x_1$ and $\tilde{p}_T = n_1\tilde{p}_1/n_T$ so $\hat{\beta}_W = n_T\tilde{p}_T/n_1 \approx r\tilde{p}_T$ assuming similar sample sizes for each sample. In Figure 4 we display $\hat{\beta}_W$ as a function of allele frequencies for SNPs located on chromosome 2 in the 1000 Genomes project. Individuals were grouped by regions (Africa, Europe, South Asia, East Asia and the Americas). The drawn line corresponds to $\beta_W = 5p_T$. The initial linear segment corresponds to alleles that are present in one continent only. β_W's start departing from this line for allele counts larger than 80, or equivalently, for worldwide frequencies larger than ≈ 0.01, given the sampled chromosome number of $2,426$.

When a new allele appears, it will be present in one population only. We expect most if not all rare alleles to be private alleles, and thus the expected values for F_{ST} (β_W) for these rare alleles are their (sub-population) frequencies. When β_W starts departing from the allele frequency, it implies that some scattering has been happening. In species with a lot of migration, this will happen at low frequencies, whereas the species that are more sedentary should show a one to one relation between sub-population allele frequencies and β_W for a larger range of their site frequency spectrum.

[Figure 4]

In Buckleton et al. (2016) we gave population-specific F_{ST} estimates for a set of 446 populations, using published data for 24 microsatellite loci collected for forensic purposes. We showed in that paper how the choice of a reference set of populations can affect results. For a set of African populations, the average within-population matching proportion was $\tilde{M}_W = 0.1884$ and the average between-population-pair averages were $\tilde{M}_B = 0.1691$ within the African region and $\tilde{M}_B = 0.1726$ for all pairs of populations. There is a larger F_{ST} for the set of African populations ($\hat{\beta}_W = 0.0082$) with Africa as a reference set than there is ($\hat{\beta}_W = 0.0020$) with the world as a reference set. The

opposite was found for a collection of Inuit populations: the average within-population matching proportion was $\tilde{M}_W = 0.4379$ whereas the average between-population-pair matching proportions were $\tilde{M}_B = 0.1726$ for pairs within the Inuit group and $\tilde{M}_B = 0.0090$ for all pairs in the study: so F_{ST} is less with Inuit as a reference set ($\hat{\beta}_W = 0.0205$) than with the world as a reference set ($\hat{\beta}_W = 0.1057$).

Inbreeding and Relatedness

To check on the validity of our estimators of individual inbreeding and coancestry coefficients, we simulated data for a range of 11 coancestries: $(i/32 : i = 0, 1, 2, \ldots, 10)$. Using the *ms* software (Hudson, 2002), we generated data from an island model with two populations exchanging $Nm = 1$ migrant per generation. We simulated 5,000 independent loci, read either as haplotypes (5,000) or as SNPs (approximately 80,000 polymorphic sites for the founders). We then chose 20 individuals from one of these populations and let them mate at random, without selfing. We did not assign or consider sex for these 20 founders. The number of offspring per mating was Poisson with mean of five. These offspring were than allowed to mate at random, without selfing, to produce families of size Poisson with mean three. By keeping records of all matings we could generate the pedigree-based inbreeding and coancestry values for all 135 individuals: founder, their offspring and their grand-offspring. The pedigree-based coancestries for all 9,045 pairs of individuals are shown in Figure 5, although we note (Hill and Weir, 2011) that the actual values have variation about expected or pedigree values. We used the same pedigree to simulate another data set, where this time 10 founders were coming from the first population and the other 10 from the second population, thus creating admixture among the children and grand-children.

[Figure 5]

The left hand plot of Figure 6 reflects the summing to zero by construction of the $\hat{\beta}_{ii'}, i \neq i'$ coancestries, whereas the pedigree coancestries are necessarily non-negative. The right hand plot shows a "correction" of the estimates: we took the set of smallest $\hat{\beta}_{ii'}$ values in the left hand plot to represent the unrelated (relative to the assumed-unrelated) founders. If we write $\hat{\beta}^0$ as the average

value in this distribution then our corrected values $\hat{\beta}^c_{ii'}$ are

$$\hat{\beta}^c_{ii'} = \frac{\hat{\beta}_{ii'} - \hat{\beta}^0}{1 - \hat{\beta}^0} \quad (6)$$

The corrected estimates are clearly close to the pedigree values. However, we are not sure if it is necessary, in general, to undertake this correction process. Whether or not it is applied, the $\hat{\beta}$ values are still relative to those among all pairs of individuals in a study sample. In general, we will not have any individuals identified for which it is justified to assume zero relatedness or zero inbreeding, and we note the comment by Thompson (2013) "in most populations IBD within individuals is at least as great as IBD between."

[Figure 6]

The distributions of estimates in Figure 7 are tightly clustered around 11 values, corresponding to the 11 distinct pedigree values $i/16, i = 0, 1, 2 \ldots 11$. A contrasting result is shown in Figure 8, for the CGTA estimates, calculated as weighted averages over loci (in the sense of equation 3 by taking the ratio of two sums over loci).

[Figure 7]

[Figure 8]

There is a current tendency in genome wide association studies (GWAS) to restrict the SNPs used in relatedness estimation to having a minor allele frequency (MAF) above some threshold. For example, the *KING* manual (http://people.virginia.edu/∼wc9c/KING/manual.html) lists a parameter **minMAF** to specify the minimum minor allele frequency to select SNPs for relationship inference in homogeneous populations. The thought is that lesser frequencies give rise to biased values, but that is not likely the case if "ratio of averages" estimates are used. To illustrate the effect of MAF filtering, we applied four different thresholds for our simulated data and we show the means and standard deviations for estimates for each of nine pedigree values in Table 2. The estimates are the corrected values – i.e. relative to an assigned value of zero for the least-related class. There is clear evidence for the merits of retaining all SNPs, both in terms of bias and variance.

[Table 2]

We continued a comparison of $\hat{\beta}$ values by applying the estimates described by Wang (2014) and computed using the *related* R package (Pew et al., 2015), listed in Table 3. Additionally, *related* offers maximum likelihood estimators, derived by Milligan (2003) and Wang and Santure (2007). They are not computed here, because they require substantial computing time, which rules them out for genomic data.

[Table 3]

In Figure 9 we display box plots of coancestry estimates for eight alternative estimates, displayed according to 9 pedigree values. The β estimates are not corrected, yet have good bias and variance properties compared to other estimates.

[Figure 9]

In Figure 10 we compare our β estimates with those from GCTA for admixed individuals with two ancestral populations. We used the same pedigree as in the section above, but took as founders 10 individuals from each of the two populations. Coancestries were calculated for all pairs of individuals in the pedigree. Figure 10 illustrates the accuracy of our β estimate compared with GCTA using the coancestries among founders. The $\hat{\beta}$'s for pairs of founders from the same population are tightly distributed around 0.015, while $\hat{\beta}$'s for pairs of individuals one from each population are tightly distributed around -0.11. The distribution for the same two categories for the GCTA estimators is wider, in particular for pairs of individuals originating from the same population.

[Figure 10]

DISCUSSION

A Unified Approach

Although there has been general recognition that family and evolutionary relatedness are just two ends of a continuum, we are not aware of previous estimates of population structure quantities such as F_{ST} or individual-pair coancestries that rest on this common framework. We have presented estimates that apply equally well to populations and individuals. While their statistical properties

remain to be fully explored, it is reassuring to see how well they performed in the few simulations presented here.

Although individual-specific inbreeding coefficients and individual-pair-specific coancestry coefficients are used routinely in association studies, we have not seen widespread adoption of population-specific F_{ST} values in evolutionary studies. We have shown here, theoretically and empirically, that these values can differ substantially among populations. This may simply reflect population size and migration rate differences, but different values may also provide signatures of natural selection.

There is also general understanding that identity by descent is a relative concept, rather than an absolute concept. This understanding has not led to an apparent recognition that the usual estimates of inbreeding and kinship are not unbiased for expected or pedigree values. Replacing population allele frequencies by sample values leads to bias in the usual estimates, *regardless of sample size*. As the allele frequencies enter GCTA estimates, for example, as squares the expected values of the estimates depend on the variances these frequencies. These, in turn, depend on the parameters being estimated.

We also stress that all allelic variants, whatever their frequencies, need to be included in the estimation of population structure and inbreeding or relatedness. The estimates certainly depend on the allele frequencies, and restricting the range of frequencies used may reveal features of interest, but the underlying ibd parameters do not depend on the frequencies. Exclusion of some alleles based on their frequencies will lead to biased estimates of the parameters.

Previous Estimates

Weir and Cockerham Estimates of F_{ST}. The F_{ST} estimate of WC84 has been widely adopted and it performs well for the model stated in that paper: data from a series of independent populations with equivalent histories. In the present notation, WC84 assumed $\theta_{ii} = \theta, \theta_{ii'} = 0$ for all populations i and all $i' \neq i$. The estimate was designed to be unbiased for any number of populations, any sample sizes and any number of alleles per locus. The analysis was a weighted one over population: the average allele frequencies $\bar{p}_u$ for a study had sample size weights, $\bar{p}_u = \sum_i n_i \tilde{p}_{iu} / \sum_i n_i$. Although our β estimates do not make explicit mention of allele frequencies, there is implicit use of sample frequencies that are unweighted averages over individuals or populations.

Weighting over populations has been discussed by Tukey (1957) and Robertson (1962). Those

authors were concerned with bias and variance and they used the language of variance components, within and between populations. For allele u these components were given as $(1-\theta)p_u(1-p_u)$ and $\theta p_u(1-p_u)$, respectively, by WC84. Tukey said "In practice, we select two quadratic functions by some scheme involving intuition, find how their average values are expressed linearly in terms of the variance components, and then form two linear combinations of the original quadratics whose average values are the variance components. These linear combinations are then our estimates. Much flexibility is possible." The estimates of WC84, Weir and Hill (2002) and Bhatia et al. (2013) all have this structure although ratios of linear combinations are taken to remove the allele frequency parameters. Tukey went on to say that the weights $w_i = n_i$ (in the present notation) "gives the customary analyses, which treat observations as important and columns [i.e. populations] as unimportant." Further, "the choice $w_i = 1$ ... treat the columns as important. This [unweighted] approach is appropriate when the column variance component is large compared with the within variance component." Robertson (1962) also pointed to sample-size weights for small between-population variance components and equal weights for large values. Bhatia et al. (2013) were concerned with unequal F_{ST} values so their use of equal weights is consistent with Turkey's statements. Their work provides simple averages of the different F_{ST}'s as opposed to averages weighted by sample sizes. For unequal F_{ST}'s and unequal sample sizes, Weir and Hill (2002) said "the usual moment estimate [with sample-size weights] is of a complex function [of the F_{ST}'s]." In our current model of unequal θ_i's and non-zero $\theta_{ii'}$'s we agree that unweighted analyses (population weights of 1) are appropriate, and that is what we have used in this paper. We note that Tukey's "flexibility" in the choice of moment estimators, phrased in terms of weights, does not arise with maximum likelihood approaches. If sample allele frequencies are taken to be approximately normally distributed then REML methods give appropriate and unique estimates.

What are the consequences of using the WC84 estimates when the current model is more appropriate? We can show that the expected value of the Weir and Cockerham estimate $\hat{\theta}_{WC}$ is

$$\mathcal{E}(\hat{\theta}_{\rm WC}) \;=\; \frac{\theta_W^c - \theta_B^c + Q}{1 - \theta_B^c + Q}$$

This expression uses three functions of sample sizes: $\bar{n} = \sum_{i=1}^{r} n_i/r$, $n_i^c = n_i - n_i^2/\sum_i n_i$ and $n_c = \sum_i n_i^c/(r-1)$. The two weighted averages are $\theta_W^c = \sum_i n_i^c\theta_{ii}/\sum_i n_i^c$ and $\theta_B^c = \sum_i \sum_{i'\neq i} n_i n_{i'}\theta_{ii'}/\sum_i \sum_{i'\neq i} n_i n_{i'}$. The quantity Q is $Q = [\sum_i (n_i/\bar{n} - 1)\theta_{ii}]/[n_c(r-1)]$. For equal sample sizes, $n_i = n$, or for equal

values of F_{ST}, $\theta_{ii} = \theta_W = \theta_W^c = \theta$, $Q = 0$. Under these circumstances $\mathcal{E}(\hat{\beta}_{WC}) = (\theta_W - \theta_B)/(1 - \theta_B)$ and we find the WC84 estimator performs well unless the θ_{ii}'s and/or the n_i's are quite different. We stress though that it is $(\theta_W - \theta_B)/(1 - \theta_B)$ being estimated.

Nei Estimates of F_{ST} Although we have phrased estimates in terms of matching proportions, we note that they are the complements of "heterozygosities" $\tilde{M} = 1 - \tilde{H}$. Our approach uses $\tilde{M}_B$, the average population-pair allele matching, whereas most previous treatments, from Nei (1973) onwards, use total heterozygosities $\tilde{H}_T = 1 - \sum_u \bar{p}_u^2$ where $\bar{p}_u$ is the average sample allele frequency over populations: $\bar{p}_u = \sum_{i=1}^r \tilde{p}_{iu}/r$. From Equations 1, the variance of $\bar{p}_u$ is

$$\mathrm{Var}(\bar{p}_u) \quad = \quad p_u(1 - p_u)\left(\theta_B + \frac{\theta_W - \theta_B}{r} + \frac{1}{r^2}\sum_{i=1}^{r}\frac{1 - \theta_i}{n_i}\right) \tag{7}$$

For large sample sizes $\tilde{H}_T = (r-1)\tilde{H}_B/r + \tilde{H}_W/r$ and Nei's G_{ST} quantity and its expectation, in our notation, are

$$G_{ST} = 1 - \frac{\tilde{H}_W}{\tilde{H}_B - \frac{1}{r}(\tilde{H}_B - \tilde{H}_W)} \quad , \quad \mathcal{E}(G_{ST}) = \frac{\theta_W - \theta_B}{1 - \theta_B + \frac{1}{r-1}(1 - \theta_W)} \tag{8}$$

which reduce to $\hat{\beta}_W$ and $\mathcal{E}(\hat{\beta}_W)$ as r becomes large. Otherwise, the expectation of G_{ST} depends on the number r of populations. This expectation is bounded above by one, contrary to the claim of Bhatia et al. (2013). Bounds on F_{ST}, when that is defined as $(1 - \tilde{H}_W/\tilde{H}_T)$, were given by Jakobsson et al. (2013).

Nei and Chesser (1983) and Nei (1987) modified Nei's earlier approach to remove the effects of the number of populations. Jost (2008) pointed out that G_{ST} does not provide a good measure of differentiation among populations, where differentiation reflects the collection of allele frequencies $\ddot{p}_{iu}$, or their sample values $\tilde{p}_{iu}$. We regard θ's as indicators of evolutionary history, rather than of allele frequencies, and we interpret them as probabilities of pairs of alleles being identical by descent. Jost introduced $D = (H_B - H_W)/(1 - H_W)$ or $D = (\theta_W - \theta_B)/\theta_W$ as a measure of differentiation among populations. For the two-population drift scenario without mutation D, unlike β_W, does not have a simple dependence on time and so does not serve as a measure of evolutionary distance.

CGTA Estimates of Relatedness The expressions in Equation 3 provide unbiased estimates of $\theta_{ii} = (1 + F_i)/2$ and $\theta_{ii'}, i \neq i'$ when the allele frequencies are known. When study sample allele

frequencies are used the expectations of these expressions, for one locus and large samples, are

$$\begin{aligned} \mathcal{E}(\hat{\theta}_{ii}) &= \frac{\theta_{ii} - 2\psi_i + \theta_B}{1 - \theta_B} \\ \mathcal{E}(\hat{\theta}_{ii'}) &= \frac{\theta_{ii'} - \psi_i - \psi_{i'} + \theta_B}{1 - \theta_B} \end{aligned}$$

where $\psi_i = \sum_{i'=1,i'\neq i}^{r} \theta_{ii'}/(r-1)$. The extent of bias depends on how different the average coancestry of a target individual with all other study individuals is from the average coancestry of all pairs of study individuals. We stress that these estimates are not unbiased for $\theta_{ii'}$.

Association Mapping

One of our motivations for seeking a unified characterization of population structure and relatedness is that both phenomena affect samples used in association mapping. Many analyses, such as those in GCTA, use mixed linear models with an estimated Genetic Relatedness Matrix $\mathbf{A}$ being used in the formulation of the variance-covariance matrix for trait values of the study individuals. For a trait with additive genetic variance σ_A^2 and no other genetic variance components, the variance matrix for individuals includes the term $\mathbf{A}2\sigma_A^2$ and $\mathbf{A}$ has diagonal elements $(1 + F_i)/2$ and off-diagonal elements $\theta_{ii'}$. We suggest that these be estimated by $\hat{\beta}_{ii}$ and $\hat{\beta}_{ii'}$ to accommodate any (hidden) relatedness and inbreeding among study subjects. We are less sure about the common practice of also using principal components of $\mathbf{A}$ as fixed effects to accommodate population structure, especially when $\mathbf{A}$ uses $\hat{\beta}$'s, and we see the need for further investigation.

Population History

We also see the need for further exploration of the role of population-specific F_{ST} estimates in evolutionary genetic studies, given the generally unrecognized prevalence of negative expected values for populations with correlated allele frequencies shown in Figure 1 and the relationship of estimates with the site-frequency spectrum suggested in Figure 4.

Conclusion

We have presented moment estimators for the probabilities that pairs of alleles, taken from individuals or from populations, are identical by descent relative to the ibd probabilities for alleles from all pairs of individuals or populations in a study. By identifying the reference set of alleles as those in the current study we allow for negative values for population structure or relatedness parameters and their estimates. Alleles may have smaller ibd probabilities within some populations than between all pairs of populations in a study, for example. Some pairs of individuals in a study may be less related than the average for all pairs. Our estimates are phrased in terms of the proportions of pairs of alleles, within and between populations or individuals, that are of the same type (ibs).

For sets of populations, we advocate the use of population-specific F_{ST} values as these more accurately reflect population history. For sets of individuals, our estimates seem to behave at least as well as those given previously. We note that our estimates have the same logical basis, and algebraic expressions, for populations and for individuals. The chief novelty of our approach is in allowing for allele frequencies to be correlated among populations when characterizing population structure, and correlated among all individuals when characterizing individual-pair relatedness.

Acknowledgments

This work was supported in part by grants GM 075091 and GM 099568 from the US National Institutes of Health and by grant IZK0Z3_157867 from the Swiss National Science Foundation.

Literature Cited

Astle W, Balding DJ. 2009. Population structure and cryptic relatedness in genetic association studies. Statistical Science 24:451-471.

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3–12.

Beaumont MA, Balding DJ. 2004. Identifying adaptive genetic divergence among populations from genome scans. Molecular Ecology 13:969-980.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting F_{ST}: The impact of rare variants. Genome Research 23:1514-1521.

Browning SR, Weir BS. 2010. Population structure with localized haplotype clusters. Genetics 185:1337-1344.

Buckleton JS, Curran JM, Goudet J, Taylor D, Thiery A, Weir BS. 2016. Population-specific F_{ST} values: A worldwide survey. Forensic Science International: Genetics 23:91-100.

Cockerham CC, Weir BS. 1983. Variance of actual inbreeding. Theoretical Population Biology 23:85-109.

Foll M, Gaggiotti O. 2006. Identifying the environmental factors that determine the genetic structure of populations. Genetics 174:875–891.

Gaggiotti OE, Foll M. 2010. Quantifying population structure using the F-model. Molecular Ecology Resources 10:821–830.

Goudet J. 2005. *hierfstat*, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes 5:184-186.

Hill WG, Weir BS. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics Research 93:47-74.

Hill WG, Weir BS. 2012. Variation in actual relationship among descendants of inbred individuals.

Genetics Research 94:267-274.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model. Bioinformatics 18:337-8

Jakobsson M, Edge MD, Rosenberg NA. 2013. The relationship between F_{ST} and the frequency of the most frequent allele. Genetics 193:515–528.

Jost L. 2008. G(ST) and its relatives do not measure differentiation. Molecular Ecology 17:4015–4026.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42:348-354.

Li CC, Weeks DE, Chakravarti, A. 1993. Similarity of DNA fingerprints due to chance and relatedness. Human Heredity 43: 45-52.

Lynch M. 1988. Estimation of relatedness by DNA fingerprinting. Molecular Biology and Evolution 5: 584-599.

Lynch M, Ritland K. 1999. Estimation of pairwise relatedness with molecular markers. Genetics 152: 1753-1766.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sal M, Chen W-M. 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26:2867-2873.

Mathieson I, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. Nature Genetics 44:243-248.

Maruyama T. 1970. Effective number of alleles in a subdivided population. Theoretical Population Biology 1:27-306.

McTavish EJ, Hillis DM. 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics 16:266–278.

Milligan BG. 2003. Maximum-likelihood estimation of relatedness. Genetics 163:1153-1167.

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proceedings of the National

Academy of Sciences, USA 70:3321–3323.

Nei M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei M, Chesser RK. 1983. Estimation of fixation indices and gene diversities. Annals of Human Genetics 47:253-259.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genetics 2:2074–2093.

Pew J, Muir PH, Wang J, Frasier TR. 2015. related: an R package for analysing pairwise relatedness from codominant molecular markers. Mol Ecol Resour 15: 557-561

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. 2007. PLINK: A toll set for while-genome association and population-based linkage analysis. American Journal of Human Genetics 81:559–575.

Queller DC, Goodnight KF. 1989. Estimating relatedness using molecular markers. Evolution 43: 258-275.

Reynolds J, Weir BS, Cockerham CC. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105:767–779.

Ritland K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical Research 67:175–185.

Robertson A. 1962. Weighting in the estimation of variance components in the unbalanced single classification. Biometrics 18:3-17.

Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpa V, Huang J, Akey JM, Jones KW. 2004. The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Human Genomics 41:274-286.

Slatkin M. 1985. Rare alleles as indicators or gene flow. Evolution 39:53-65.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Thompson EA. 1975. Estimation of pairwise relationships. Annals of Human Genetics 39:173-188.

Thompson EA. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194:301–326.

Tukey JW. 1957. Variances of variance components: II. The unbalanced single classification. Annals of Mathematical Statistics 28: 43-56.

Wang J. 2002. An estimator for pairwise relatedness using molecular markers. Genetics 160: 1203-1215.

Wang J. 2014. marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. Journal of Evolutionary Biology 27:518–530.

Wang J, Santure AW. 2009. Parentage and sibship inference from multilocus genotype data under polygamy. Genetics 181:1579-1594

Weir BS. 1996. Genetic Data Analysis II. Sinauer, Sunderland, MA.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Research 15:1468-1476.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Weir BS, Hill WG. 2002. Estimating F-statistics. Annual Review of Genetics 36:721-750.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42:565-569.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis. American Journal of Human Genetics 88:76–82.

Yu J, Pressoir G, Briggs WR, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203-208.

Table 1. Predicted and estimated β's for two populations.

Model	N_1	N_2	m_1	m_2	β_1	$\hat{\beta}_1$	β_2	$\hat{\beta}_2$	β_{12}
1	100	1,000	0.01	0.01	0.156	0.163 (0.152, 0.173)	-0.037	-0.039 (-0.047,-0.032)	0.059
2	100	1,000	0.01	0.01	0.198	0.199 (0.192, 0.206)	0.024	0.026 (0.023, 0.029)	0
3	100	1,000	0.01	0.01	0.278	0.283 (0.268, 0.296)	-0.061	-0.059 (-0.067,-0.050)	0.112
4	10,000	100	0.01	0	-0.319	-0.302 (-0.409, -0.219)	0.489	0.468 (0.372, 0.599)	0.085

Mutation rate $\mu = 10^{-6}$

95% confidence intervals from bootstrapping over loci.

Table 2. Effects of filtering to L SNPs on coancestry estimate means (and standard deviations $\times 100$).

	$L = 79,069$	$L = 72,012$	$L = 56,979$	$L = 44,061$
Pedigree value	All SNPs	MAF≥ 0.01	MAF≥ 0.05	MAF≥ 0.10
0	0.000 (0.50)	0.000 (1.00)	0.000 (1.99)	0.000 (2.43)
0.03125	0.031 (0.30)	0.026 (0.30)	0.010 (0.89)	0.003 (1.45)
0.06750	0.061 (0.34)	0.056 (0.35)	0.041 (1.13)	0.036 (1.79)
0.09375	0.092 (0.27)	0.087 (0.27)	0.069 (0.72)	0.061 (1.13)
0.12500	0.124 (0.41)	0.120 (0.46)	0.112 (1.90)	0.109 (2.69)
0.15625	0.156 (0.29)	0.151 (0.29)	0.133 (0.65)	0.122 (1.15)
0.18750	0.184 (0.26)	0.179 (0.27)	0.157 (1.07)	0.144 (1.64)
0.25000	0.249 (0.42)	0.245 (0.45)	0.241 (1.87)	0.239 (2.62)
0.31250	0.311 (0.20)	0.307 (0.20)	0.285 (0.77)	0.271 (1.23)

Table 3. Other estimates of relatedness.

Method	Description
`ped`	The pedigree based relatedness. It can be considered as the true relatedness value, although it wil depend on the depth of the pedigree.
`bij`	β_{ij}, developed here. These values are relative to the mean of the population, and hence, the mean of these relatedness must be 0.
`wang`	The estimator developed by Wang (2002).
`lynchli`	The estimator derived by Lynch (1988) and improved by Li et al. (1993), equation (7) in Wang (2
`lynchrd`	The estimator derived by Lynch and Ritland (1999) [eq (5,6) in Wang, 2014].
`GCTA`	The estimator used in GCTA (Yang et al., 2011). For multi-allelic markers, alleles are weighted according to their variance.
`ritland`	The estimator derived by Ritland (1996) [eq (4) in Wang (2014)]. For SNPs, it is the same as `GCTA`, but for multi-allelic markers, each allele is given the same weight.
`quellergt`	The estimator derived by Queller and Goodnight (1988) [eq (2,3) in Wang (2014)].

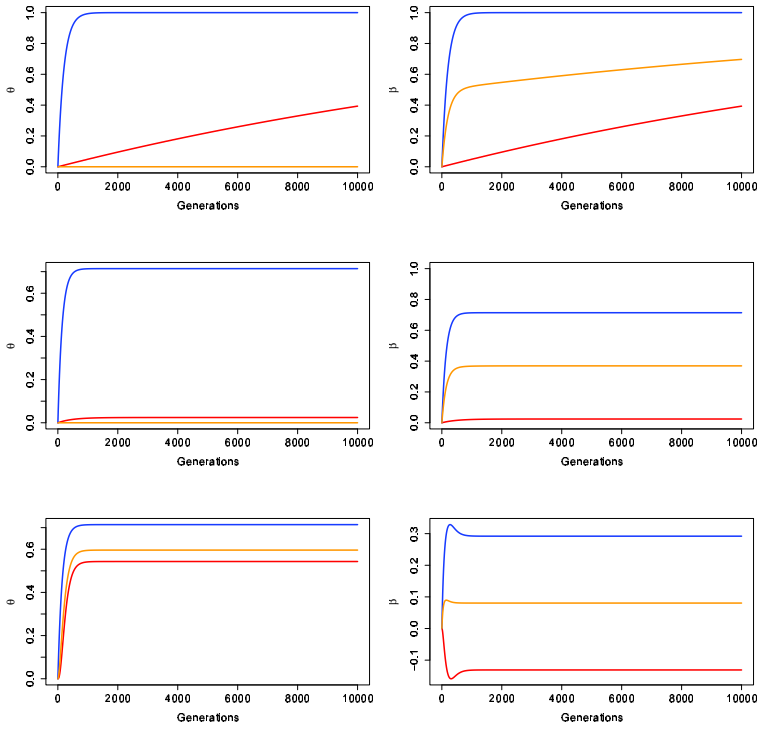

Figure 1. Effects of Drift, Mutation and Migration. For all panels, $N_1 = 10000, N_2 = 100$ First row: drift only (no mutation nor migration). θ_1, θ_2's and β's tend to 1,$\theta_{12} = 0.000$. Second row: Drift and Mutation $\mu = 10^{-3}, m_1 = m_2 = 0$. θ's and β's have positive limits less than 1. At equilibrium, $\theta_1 = 0.024, \theta_2 = 0.714, \theta_{12} = 0.000, \beta_1 = 0.024, \beta_2 = 0.714, \beta_w = 0.369$. Third Row: Drift. Mutation and Migration. $\mu = 10^{-3}, m_1 = 10^{-2}, m_2 = 0$. θ's positive and less than 1, β_W is positive but β_{ii}'s may be negative. At equilibrium, $\theta_1 = 0.543, \theta_2 = 0.714, \theta_{12} = 0.596, \beta_1 = -0.131, \beta_2 = 0.292, \beta_w = 0.080$.

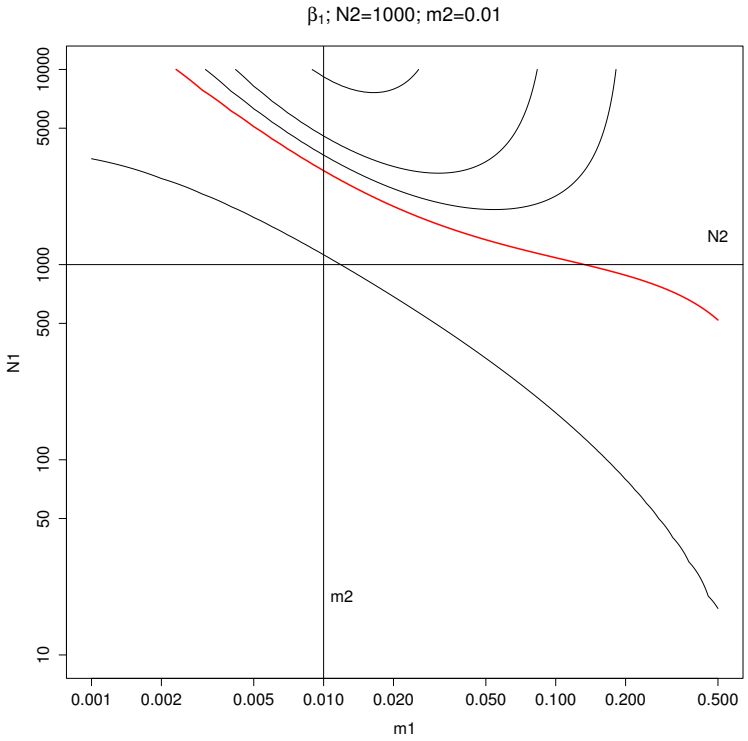

Figure 2. The region above and to the right of the red line has equilibrium values of $\theta_1 \leq \theta_{12} \leq \theta_2$, i.e. $\beta_1 \leq 0 \leq \beta_2$. In that region a pair of alleles within population 1 has a smaller probability of ibd than does an allele from population 1 with an allele from population 2.

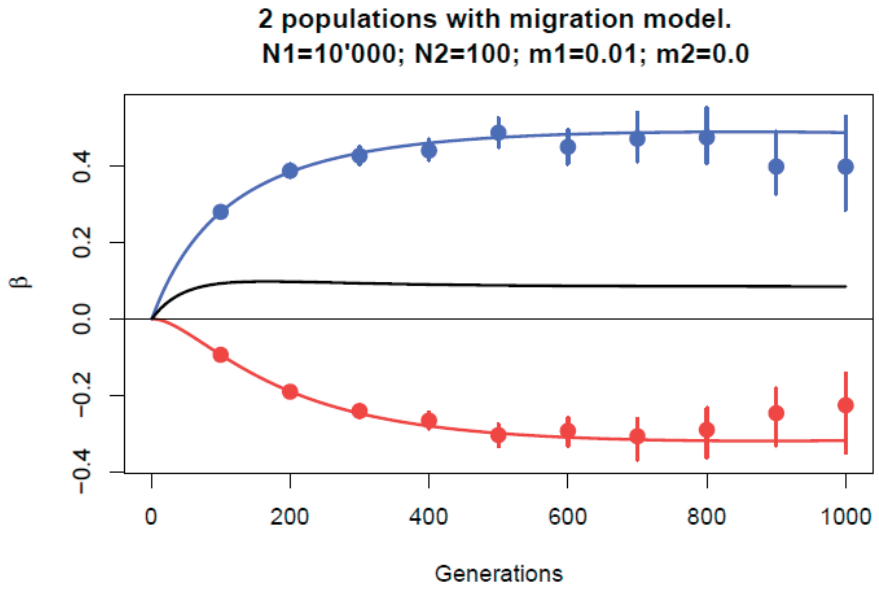

Figure 3. Estimated β's for independent simulations at different times, showing increase in bias and variances as the number of polymorphic loci decreases.

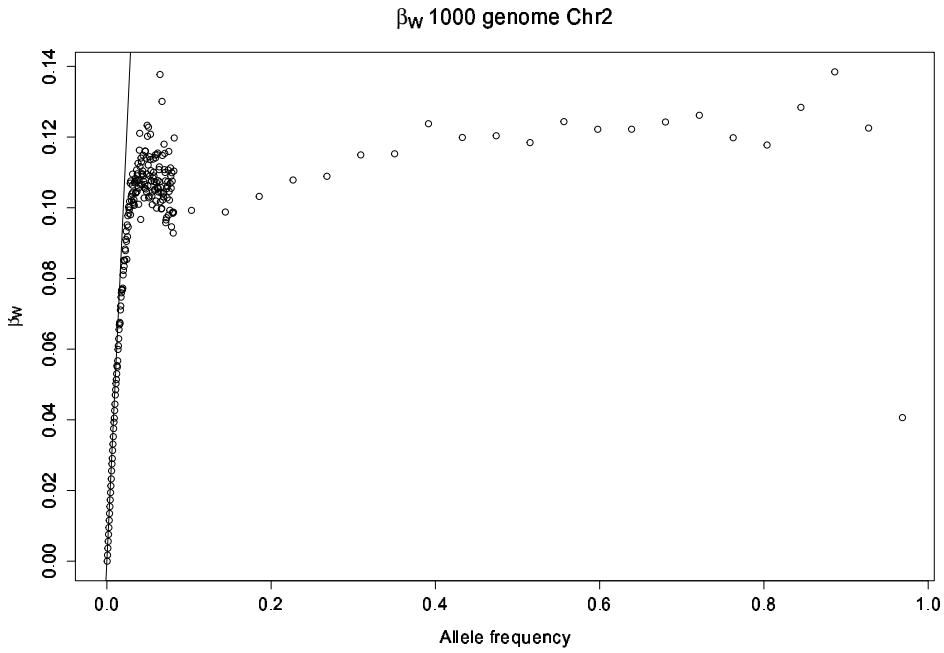

Figure 4. β_W as a function of allele frequencies (n_u/n_T) for SNPs located on chromosome 2. Data from the 1000 genomes project, individuals were grouped by regions (Africa, Europe, South Asia, East Asia and Americas). The drawn line corresponds to $5n_u/n_T$. The initial linear segment corresponds to alleles that are present in one continent only. β_Ws start departing from this line for allele counts larger than 80, or equivalently, for worldwide frequencies larger than ≈ 0.01, given the sampled chromosome number of 2426.

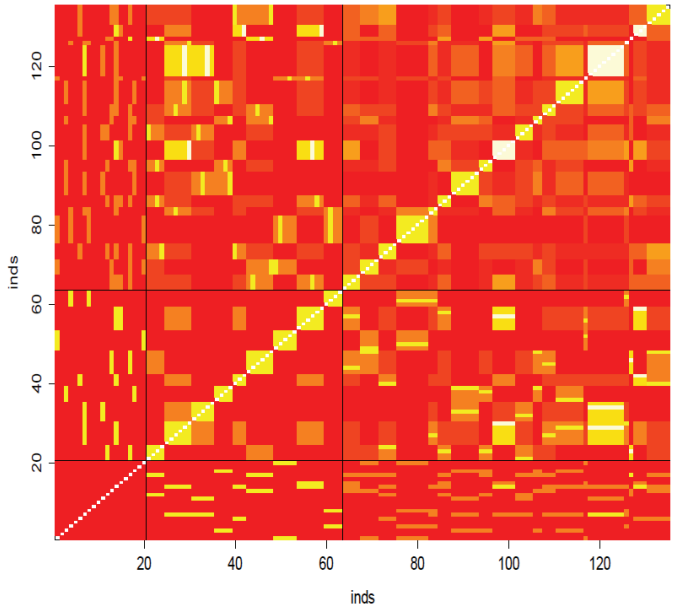

Figure 5. Pedigree-based inbreeding and coancestry coefficients for simulated data.

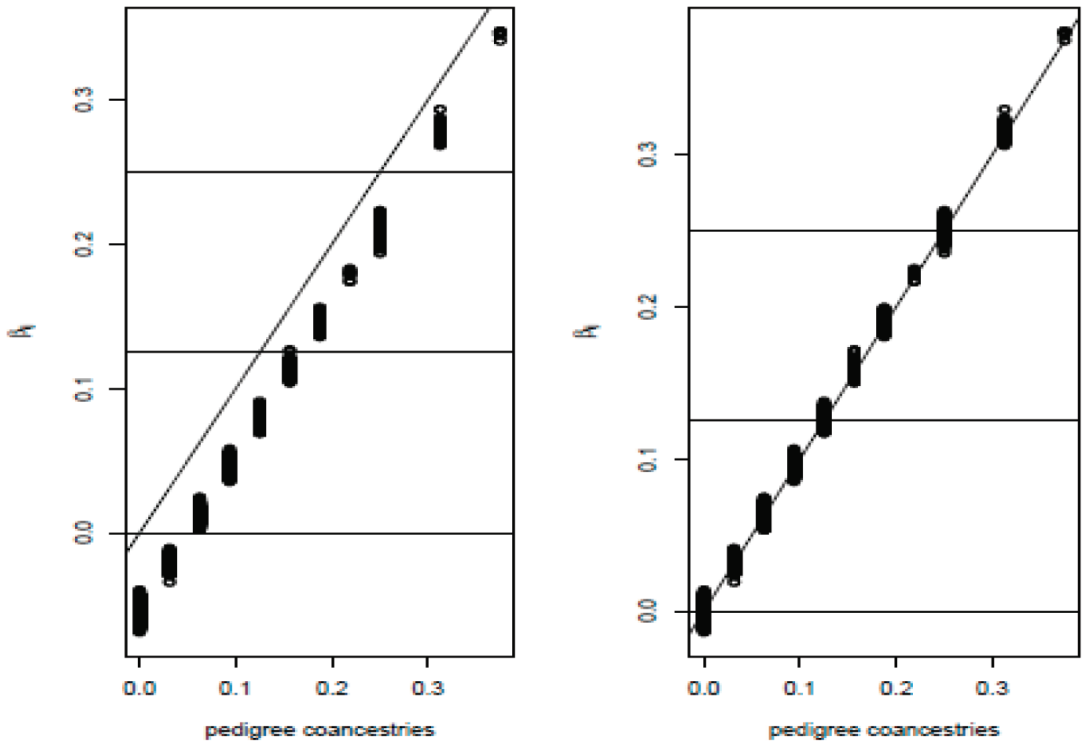

Figure 6. Comparison of estimated and pedigree coancestries. Uncorrected estimates on left, corrected estimates on right. Correction procedure described in text.

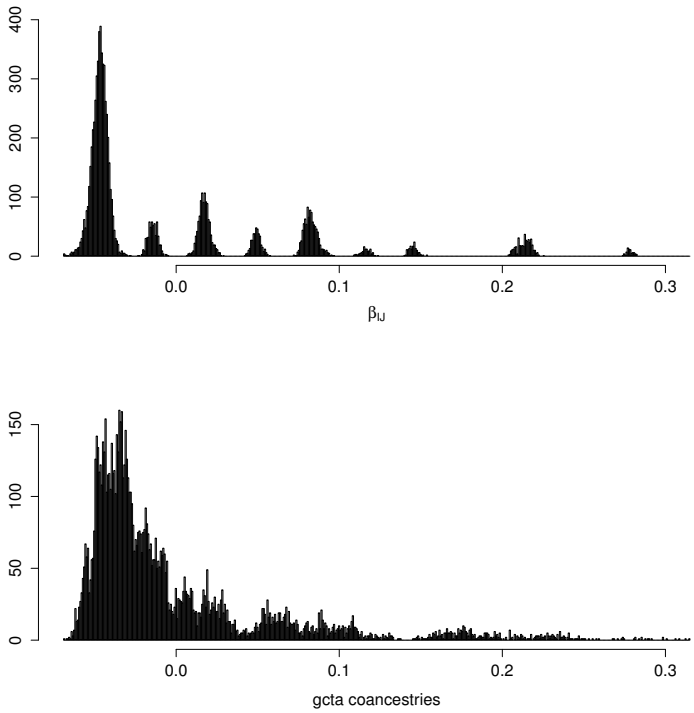

Figure 7. Comparison of β and CGTA coancestry estimates, when founders are drawn from a single population.

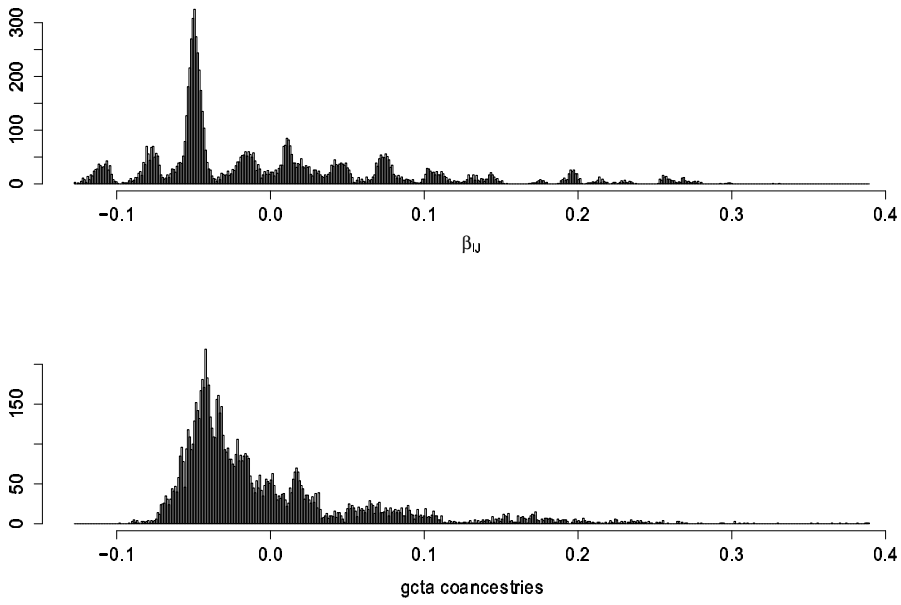

Figure 8. Comparison of β and CGTA coancestry estimates, when founders are drawn from two populations.

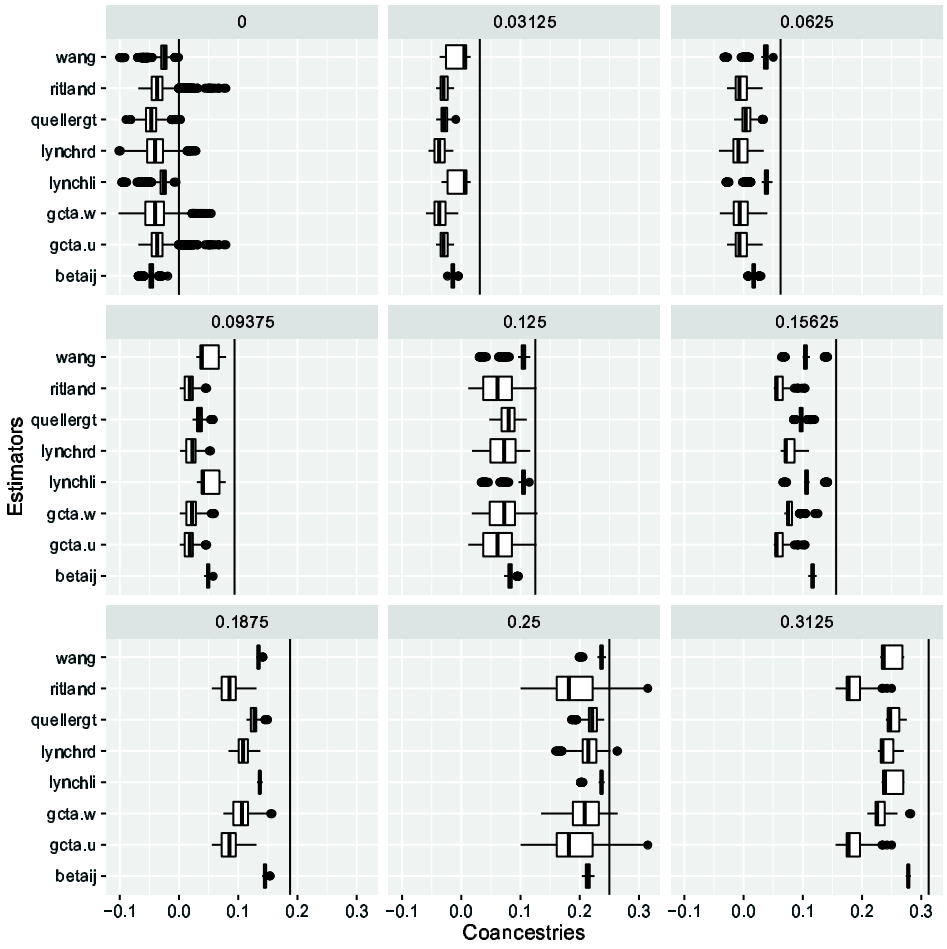

Figure 9. Boxplots of coancestry estimates for eight alternative estimates, displayed according to nine pedigree values. The β estimates are not corrected, yet have good bias and variance properties.

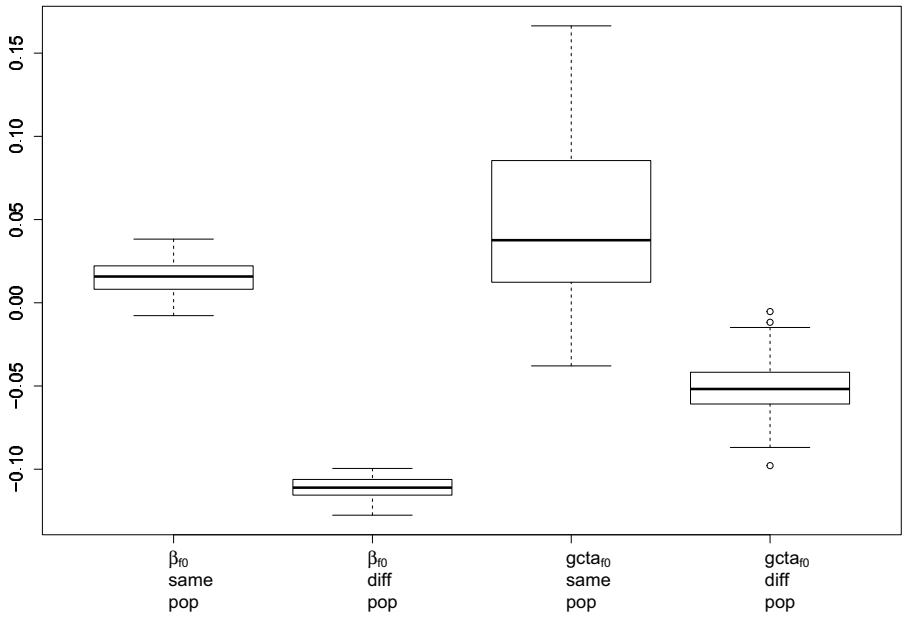

Fig 10. Boxplots of coancestry estimates β and GCTA when the founders come from two populations. Coancestries were estimated for all the individuals in the pedigree, but only those between founders are shown. For each panel, the left boxplot is for pairs of founders from the same population, while the right boxplot is for pairs when the two members come from different populations.