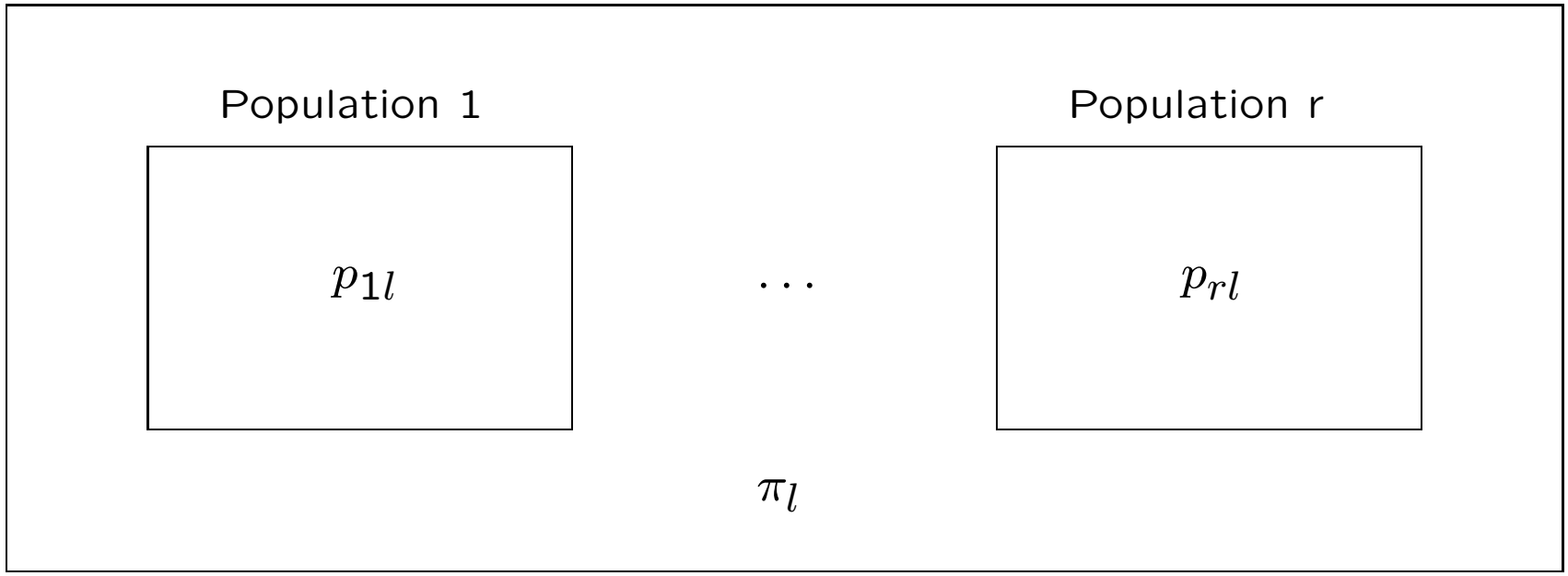
POPULATION STRUCTURE

Genetic Analysis: SNP l Allele Frequencies



Among samples of n_i alleles from population i: counts for the SNP l reference allele follow a binomial distribution with mean p_{il} and variance $n_i p_{il}(1-p_{il})$. Sample allele frequencies $\tilde{p}_{il}$ have expected values p_{il} and (under HWE) variances $p_{il}(1-p_{il})/n_i$.

Among replicates of population i: p_{il} values follow a distribution with mean π_l and variance $\pi_l(1-\pi_l)\theta^i$. Distribution sometimes assumed to be Beta.

What is θ?

Two ways of thinking about θ.

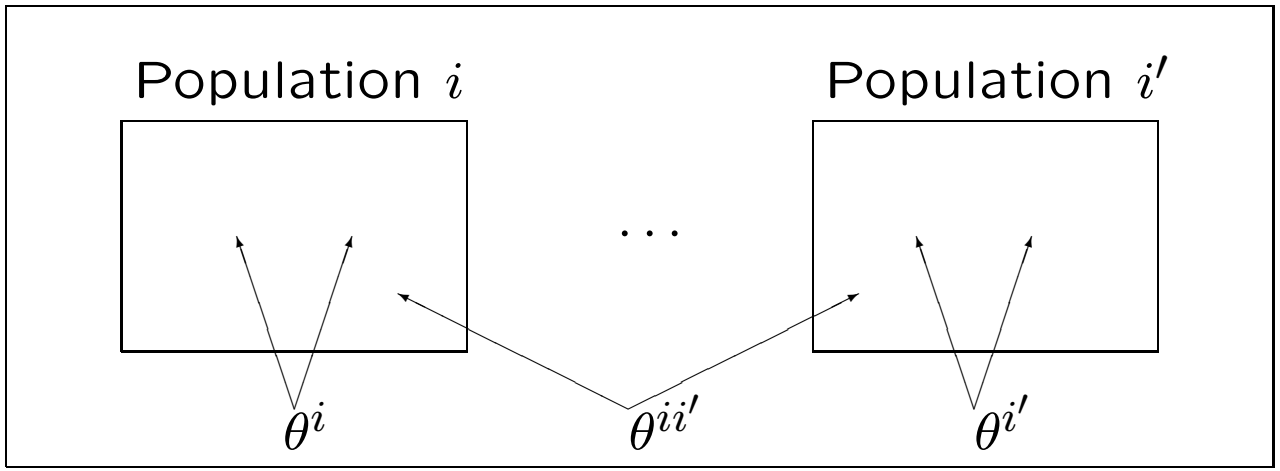
It measures the probability a pair of alleles are identical by descent: and this is with respect to some reference population.

The target alleles may be in specified populations, and this leads to characterization of population structure, of they may be in specified individuals and this leads to characterization of inbreeding and relatedness.

θ also describes the variance of allele frequencies among populations, or among evolutionary replicates of a single population.

Weir BS, Goudet J. 2017. Genetics 206:2085-2103.
Goudet J, Kay T, Weir BS. 2018. Molecular Ecology 27:4121-4135.

Allele-level θ's



θ's are ibd probabilities for pairs of alleles from specified populations.

θ_W^i is average of the within-population probabilities θ^i. Average over populations of θ_W^i is θ_W.

θ_B is average of the between-population-pair probabilities $\theta^{ii'}$.

Allelic Measure Predicted Values

Predicted Values of the θ's: Pure Drift

The estimation procedure for the θ's holds for all evolutionary scenarios, but the theoretical values of the θ's do depend on the history of the sampled populations.

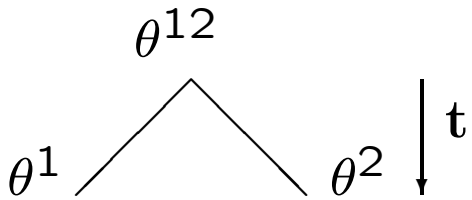
In the case of pure drift, where population i has constant size N_i and there is random mating, t generations after the population began drifting from an ancestral population in which $\theta^i = 0$

$$\theta^i(t) \;=\; 1 - \left(1 - \frac{1}{2N_i}\right)^t$$

If t is small relative to large N_i's, $\theta^i(t) \approx t/(2N_i)$, and $\theta_W(t) \approx t/(2N_h)$ where N_h is the harmonic mean of the N_i.

Drift Model: Two Populations

Now allow ancestral population itself to have ibd alleles with probability θ^{12} (the same value as for one allele from current populations 1 and 2):



$$\theta^i \;=\; 1-(1-\theta^{12})\left(\frac{2N_i-1}{2N_i}\right)^t,\;\; i=1,2$$

It is possible to avoid needing to know the ancestral value θ^{12} by making θ^1, θ^2 *relative to* θ^{12}:

$$\beta^i = \frac{\theta^i-\theta^{12}}{1-\theta^{12}} \;=\; 1-\left(\frac{2N_i-1}{2N_i}\right)^t \approx \frac{t}{2N_i},\;\; i=1,2$$

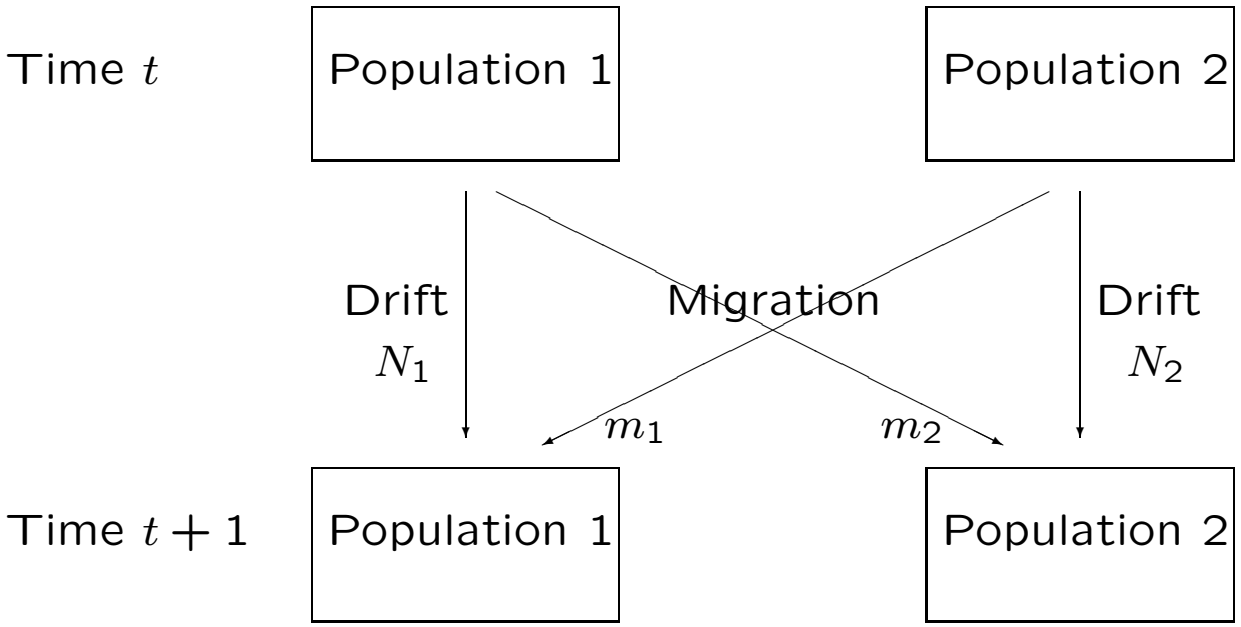
Drift and Mutation

If there is no migration, the θ's tend to equilibrium values of

$$\begin{aligned} \widehat{\theta}^1 &\approx \frac{1}{1+4N_1\mu} \\ \widehat{\theta}^2 &\approx \frac{1}{1+4N_2\mu} \\ \widehat{\theta}^{12} &= 0 \end{aligned}$$

so $\beta^i = \theta^i,\ i = 1, 2.$

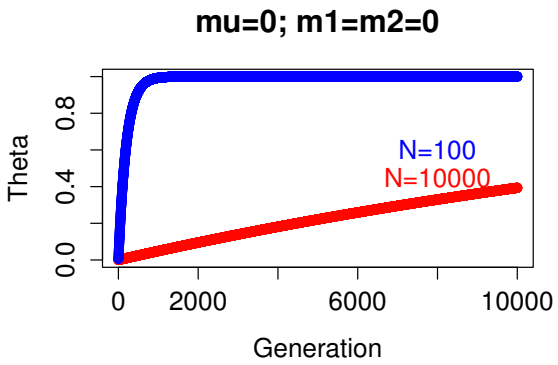
Two populations: drift, migration, mutation



There is also a probability μ that an allele mutates to a new type.

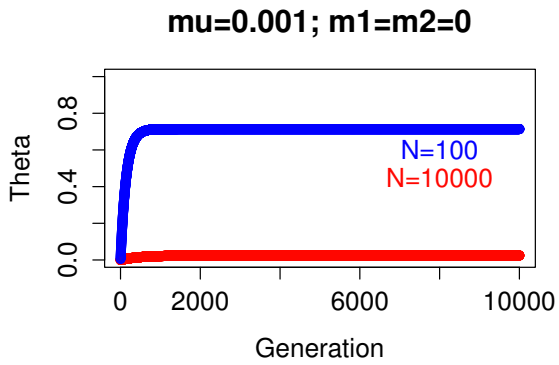
Drift, Mutation and Migration

The θ's are non-negative, but one of the β's may be negative.



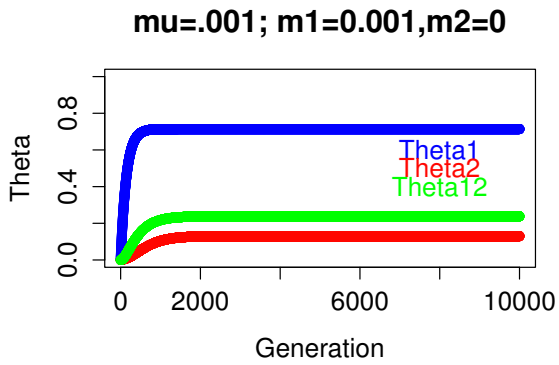
Drift Only

$\beta^1, \beta^2 > 0$



Drift and Mutation

$\beta^1, \beta^2 > 0$



Drift, Mutation and Migration

$\beta^1 > 0, \beta^2 < 0$

Multiple Populations

For random union of gametes, when pairing of alleles into individuals is not needed, the ibd probability θ_W^i for any distinct pair of alleles within population i *relative to* the ibd probability between populations is

$$\beta_{WT}^i = \frac{\theta_W^i - \theta_B}{1 - \theta_B}$$

This is the population-specific F_{WT}^i for alleles.

Averaging over populations:

$$\beta_{WT} = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

and this is the global F_{WT} for alleles. This is the quantity often referred to as "F_{ST}", but see later discussion in Kinship section.

Genotypes vs Alleles

So far, this treatment has ignored individual genotypic structure, leading to an analysis of population allele frequencies as opposed to genotypic frequencies.

θ^i is the probability two alleles drawn randomly from population i are ibd, and $\theta^{ii'}$ is the probability an allele drawn randomly from population i is ibd to an allele drawn from population i'.

Within population i, define θ^i_{jj} as the probability that two alleles drawn randomly from individual j are ibd, and $\theta^i_{jj'}$ as the probability that allele drawn randomly from individual j is ibd to an allele from individual j'.

Allelic Matching Proportions Within Populations

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

What is the proportion $\tilde{M}^i_{Wl}$ of pairs of distinct 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:

$$\tilde{M}^i_{Wl} \approx \tilde{p}^2_{il} + (1 - \tilde{p}_{il})^2$$

The expected value of this over replicates of the population is

$$\mathcal{E}(\tilde{M}^i_{Wl}) = M_l + (1 - M_l)\theta^i_W$$

where $M_l = \pi^2_l + (1 - \pi_l)^2$. 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 M_l of allele probabilities.

Matching Proportions between Populations

The observed proportion of matching allele pairs between populations i and i' is

$$\tilde{M}_{Bl}^{ii'} = \tilde{p}_{il}\tilde{p}_{i'l} + (1 - \tilde{p}_{il})(1 - \tilde{p}_{i'l})$$

The expected value of this over replicates of the population is

$$\mathcal{E}(\tilde{M}_{Bl}^{ii'}) = M_l + (1 - M_l)\theta_B^{ii'}$$

and, averaging over all pairs of populations

$$\mathcal{E}(\tilde{M}_{Bl}) = M_l + (1 - M_l)\theta_B$$

Allele-based Estimate of F_{ST}

The need to know M_l is avoided by considering allele-pair matching within a population *relative to* the allele-pair matching between pairs of populations:

$$\widehat{\beta}^i_{WT} = \widehat{F}^i_{WT} \;=\; \frac{(\tilde{M}^i_{Wl} - \tilde{M}_{Bl})}{(1 - \tilde{M}_{Bl})}$$

and this has expected value $F^i_{WT} = (\theta^i_W - \theta_B)/(1 - \theta_B)$ which is the population-specific value.

Average over populations:

$$\widehat{F}_{WT} = \widehat{\beta}_{WT} \;=\; \frac{\tilde{M}_{W_l} - \tilde{M}_{B_l}}{1 - \tilde{M}_{B_l}}$$

and the parametric global value $F_{WT} = (\theta_W - \theta_B)/(1 - \theta_B)$.

Combining information from multiple SNPs

If the θ parameters are the same for all SNPs, then information can be combined over SNPs. The "ratio of averages" method is

$$\widehat{\beta}^i_{WT} = \widehat{F}^i_{WT} = \frac{\sum_l (\tilde{M}^i_{Wl} - \tilde{M}_{Bl})}{\sum_l (1 - \tilde{M}_{Bl})}$$

and this has expected value $F^i_{WT} = (\theta^i_W - \theta_B)/(1 - \theta_B)$ which is the population-specific value. This is better than the "average of ratios" method of simply averaging the single-SNP estimates.

Ochoa and Storey showed that, as the number of SNPs increases, the ratio of averages estimate converges to the parametric value F^i_{ST}.

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

Alternative Computing Equations for F_{WT}

For large sample sizes and r populations:

$$\begin{aligned}\tilde{M}_W^i &\approx \sum_l [\tilde{p}_{il}^2 + (1-\tilde{p}_{il})^2] \\ \tilde{M}_W &= \frac{1}{r}\sum_{i=1}^{r} \tilde{M}_{Wl}^i = \sum_l [\bar{p}_l^2 + (1-\bar{p}_l)^2 + 2\frac{r-1}{r}s_l^2]\end{aligned}$$

where $\bar{p}_l = \sum_{i=1}^r \tilde{p}_{il}/r$ is the average sample allele frequency over populations, and $s_l^2 = \sum_{i=1}^r (\tilde{p}_{il} - \bar{p}_l)^2/(r-1)$ is the variance of sample allele frequencies over populations.

For all sample sizes:

$$\begin{aligned}\tilde{M}_B^{ii'} &= \sum_l [\tilde{p}_{il}\tilde{p}_{i'l} + (1-\tilde{p}_{il})(1-\tilde{p}_{i'l})] \\ \tilde{M}_B &= \frac{1}{r(r-1)} \sum_{\substack{i=1 \\ i\neq i'}}^{r} \sum_{i'=i}^{r} \sum_l \tilde{M}_{Bl}^{ii'} \\ &= \sum_l [\bar{p}_l^2 + (1-\tilde{p}_l)^2 - 2\frac{1}{r}s_l^2]\end{aligned}$$

Alternative Estimates for F_{WT}

The population-specific estimates are

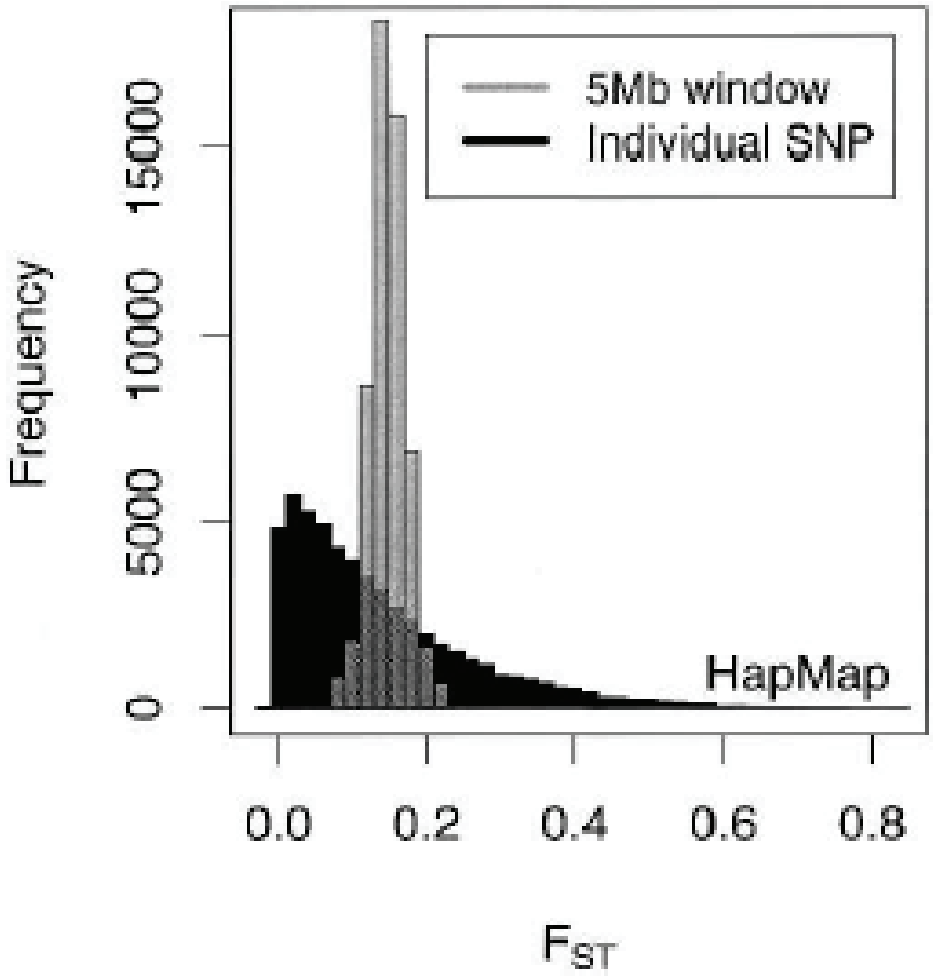
$$\widehat{F}^i_{WT} \;=\; 1 - \frac{\sum_l \tilde{p}_{il}(1-\tilde{p}_{il})}{\sum_l [\bar{p}_l(1-\bar{p}_l) + \frac{1}{r}s_l^2]}$$

The global estimates are

$$\widehat{F}_{WT} \;=\; \frac{\sum_l (s_l^2)}{\sum_l [\bar{p}_l(1-\bar{p}_l) + \frac{1}{r}s_l^2]}$$

The classical expression $s^2/\bar{p}(1-\bar{p})$ is fine if there is a large number of populations, but not for $r = 2$.

Effect of Number of Loci



Weir BS, et al. 2005. Genome Research 15:1468-1476.

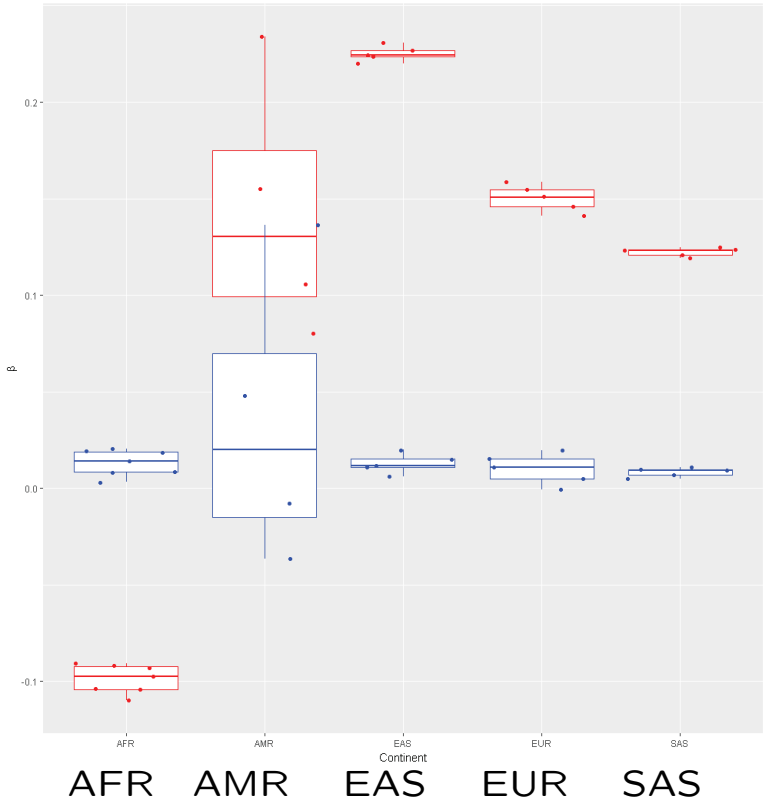
F_{WT} is relative, not absolute

Using data from the 1000 genomes, using 1,097,199 SNPs on chromosome 22.

For the samples originating from Africa, there is a larger F_{WT}, $\widehat{\beta}_{WT} = 0.013$, with Africa as a reference set than there is, $\widehat{\beta}_{WT} = -0.099$, with the world as a reference set. African populations tend to be more different from each other on average than do any two populations in the world on average.

The opposite was found for East Asian populations: there is a smaller F_{WT}, $\widehat{\beta}_{WT} = 0.013$ with East Asia as a reference set than there is, $\widehat{\beta}_{WT} = 0.225$ with the world as a reference set. East Asian populations are more similar to each other than are any pair of populations in the world.

SNP F_{ST}'s are relative, not absolute



Blue box: Population relative to pairs of populations in same continent.

Red box: Population relative to pairs of populations in whole world.

Evolutionary Inferences

$\widehat{\beta}_{WT}$ in LCT Region: 3 Populations