POPULATION STRUCTURE

Bruce Weir

April 11, 2024

Lincoln (Plant and Food)

Questions of Interest

- How much genetic variation is there? (animal conservation)

- How much migration (gene flow) is there between populations? (molecular ecology)

- How does the genetic structure of populations affect tests for linkage between genetic markers and human disease genes? (human genetics)

- How should the evidence of matching marker profiles be quantified? (forensic science)

- What is the evolutionary history of the populations sampled? (evolutionary genetics)

Statistical Analysis

It is possible to approach these data from purely statistical viewpoint.

It is possible to test for differences in allele frequencies among populations.

It is also possible to use various multivariate techniques to cluster populations.

These statistical analyses may not answer the biological questions, and the alternative is to set up an evolutionary model that takes into account the history of the populations under study. This allows for a broader interpretation of the data.

Multivariate Statistical Approach

Principal Component Analysis (PCA)

"In genetics, by exploiting DNA-based genetic variants, PCA has shown its usefulness to infer shared genetic ancestry from unrelated samples and from related samples, as covariates to correct for confounding due to population structure in genome-wide association and interaction studies to study and understand human population migrations, to reduce the huge genetic variant dimensionality for cluster analysis in clustering of subpopulations, to impute missing genetic variants and to detect outliers for population stratification (PS) in genome-wide association studies."

Patterson et al., 2006

The allele dosage, for a designated allele, at SNP l and individual j is X_{jl} with values 0,1,2.

The allele frequency $\tilde{p}_l$ for a sample of n individuals is $\tilde{p}_l = \sum_{j=1}^{n} X_{jl}/(2n)$. From Section 1 of these notes, over repeated samples from the same population, the expected value of $\tilde{p}_l$ is p_l and its variance is $p_l(1-p_l)(1+f)/(2n)$. Sampled individuals are assumed to be independent. For a single individual, the mean and variance of X_{jl} are $2p_l$ and $2p_l(1-p_l)$ if $f=0$.

Patterson et al. "normalize" the individual allele dosages to $(X_{jl} - 2\tilde{p}_l)/\sqrt{\tilde{p}_l(1-\tilde{p}_l)}$ *"at least if the data is in Hardy-Weinberg equilibrium"* so that "each data column has the same variance." The $n \times L$ matrix of normalized dosages is written as $\mathbf{X}$ where L is the number of SNPs.

Patterson N, et al. 2006. PLoS Genetics 2:e190: in slightly different notation.

Patterson et al., 2006

The $n \times n$ matrix $\mathbf{K} = \mathbf{X}\mathbf{X}'/\mathbf{L}$ has diagonal elements

$$s_j^2 = \frac{1}{L}\sum_{l=1}^{n} \frac{(X_{jl} - 2\tilde{p}_l)^2}{\tilde{p}_l(1-\tilde{p}_l)} \quad \text{for individual } j$$

and off-diagonal elements

$$s_{jj'} = \frac{1}{L}\sum_{l=1}^{n} \frac{(X_{jl} - 2\tilde{p}_l)(X_{j'l} - 2\tilde{p}_l)}{\tilde{p}_l(1-\tilde{p}_l)} \quad \text{for individuals } j, j'$$

These elements are measures of similarity (averaged over SNPs) of pairs of individuals (including self-similarity). There is no implied evolutionary framework.

PCA is a dimension reduction that can help identify ancestry relationships among sampled individuals.

Patterson et al., 2006

Patterson et al. carry out a singular value decomposition of the matrix $\mathbf{X}$.

They call the matrix $\mathbf{K}$ the "sample covariance matrix of the columns of $\mathbf{X}$." They then compute an eigenvector decomposition of $\mathbf{K}$ and say "eigenvectors corresponding to eigenvalues are exposing nonrandom population structure." The matrix will also reflect inbreeding and relatedness among the n individuals.

It would also be possible to use an allele-sharing matrix for all pairs of individuals in place of $\mathbf{K}$.

Abegaz et al.

The eigenvalue decomposition of $\mathbf{K} = \mathbf{X}\mathbf{X}'$ is

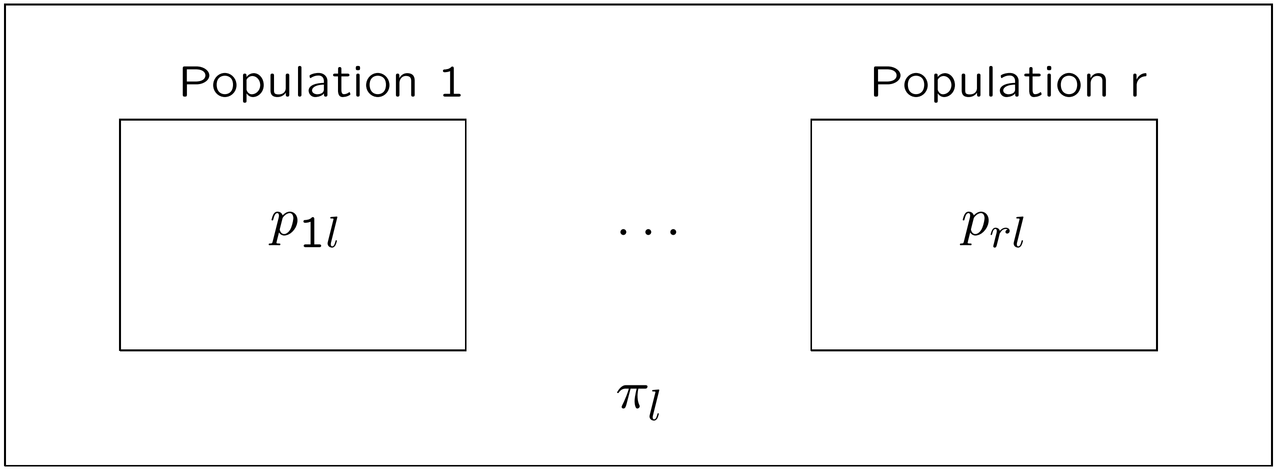
$$\mathbf{X}\mathbf{X}' \;=\; \mathbf{U}\mathbf{D}\mathbf{U}'$$

where the columns of $\mathbf{U}$ are the eigenvectors and $\mathbf{D}$ is the diagonal matrix of positive eigenvalues of $\mathbf{K} = \mathbf{X}\mathbf{X}'$. The principal coordinates for the sample individuals are $\mathbf{U}\mathbf{D}^{\mathbf{1/2}}$.

$$\mathbf{K}\mathbf{U}_j \;=\; d_j\mathbf{U}_j$$

Abegaz F, et al. Briefings in Bioinformatics, 20(6), 2019, 2200-2216.

Genetic Analysis: SNP l Allele Frequencies



Population 1

Population r

p_{1l} $\cdots$ p_{rl}

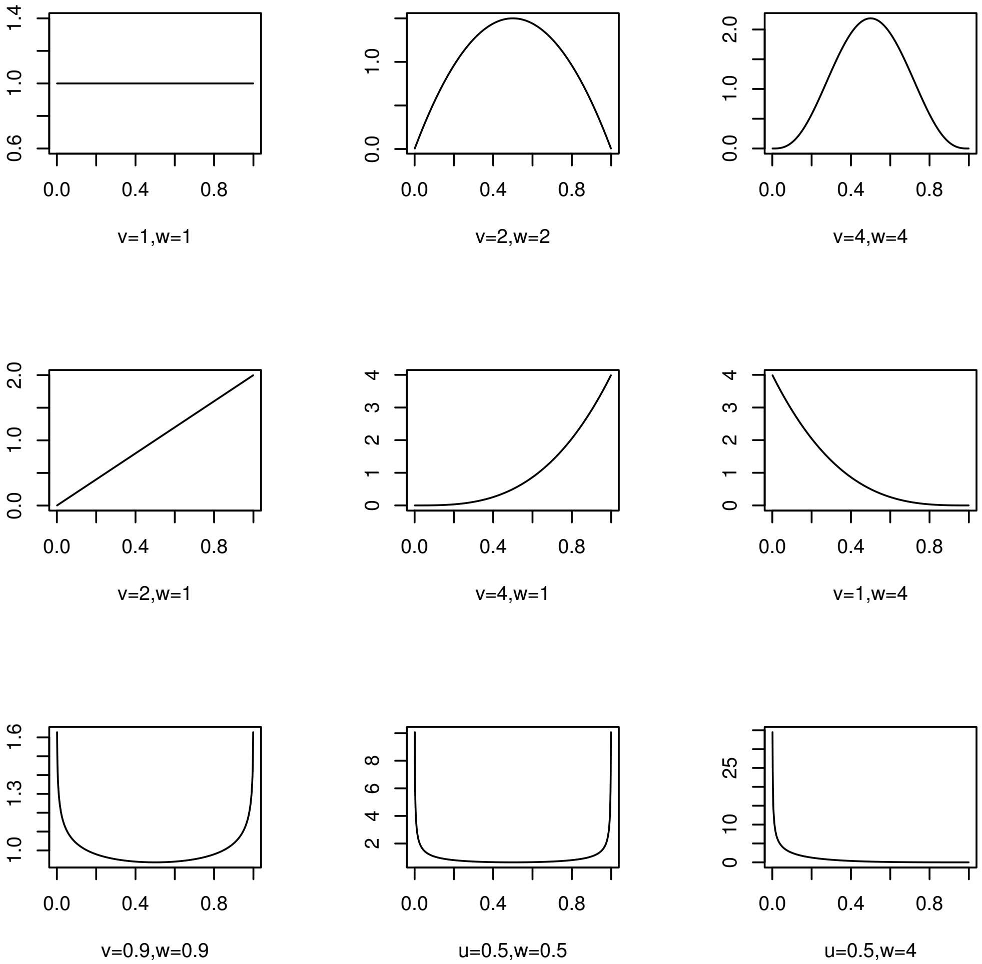
π_l

Among samples of $2n_i$ independent alleles from population i: counts for the designated allele for SNP l follow a binomial distribution with mean p_{il} and variance $2n_i p_{il}(1-p_{il})$. Sample allele frequencies $\tilde{p}_{il}$ have expected values p_{il} and (under HWE) variances $p_{il}(1-p_{il})/(2n_i)$.

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.

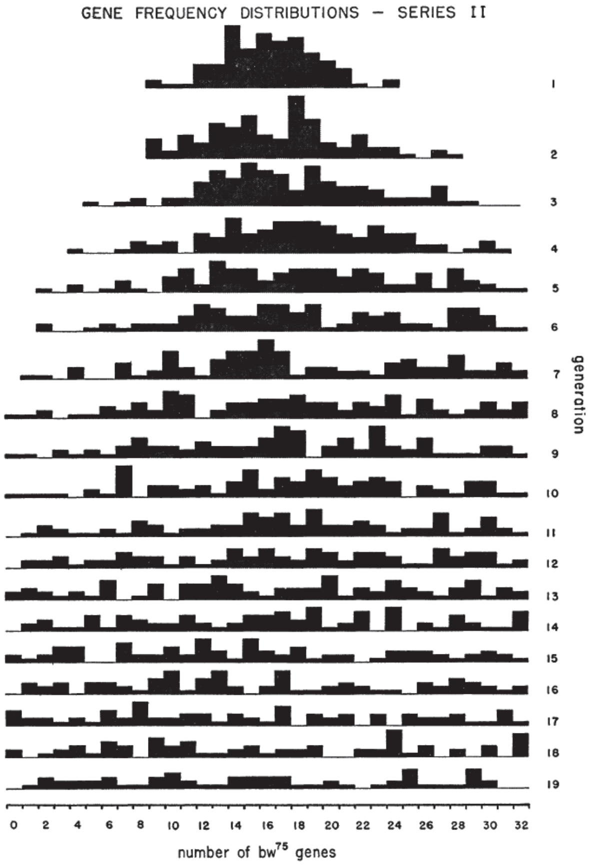
Beta distribution: Theoretical

The beta probability density is proportional to $p^{v-1}(1-p)^{w-1}$ and can take a variety of shapes.



Beta distribution: Experimental

The beta distribution is suggested by a *Drosophila* experiment with 107 replicate populations of size 16, starting with all heterozygotes:



Buri P. 1956. Evolution 10:367

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.

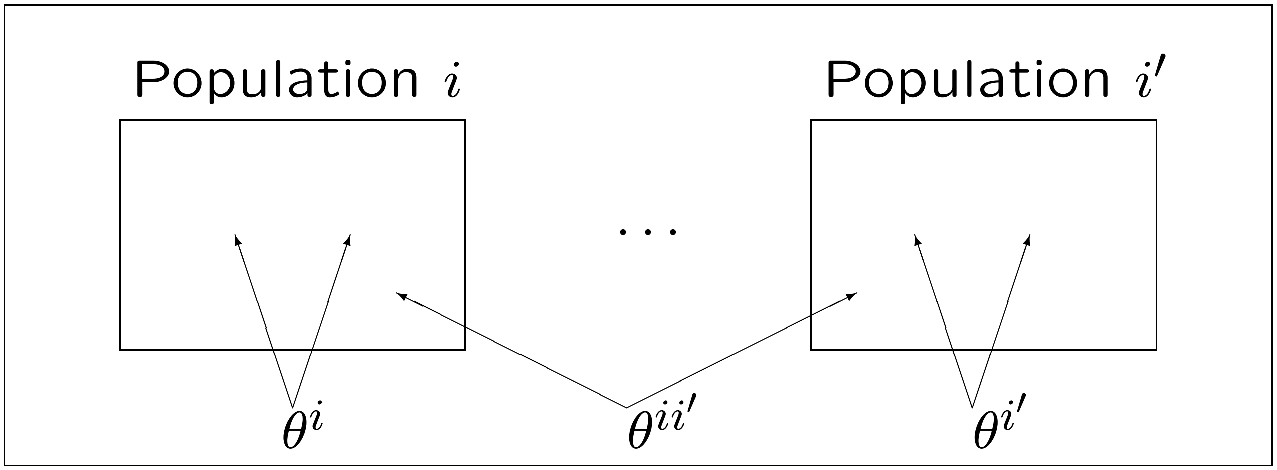
The target alleles may be in specified populations, and this leads to characterization of population structure, or they may be in specified individuals and this leads to characterization of inbreeding and relatedness.

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 distinct alleles from specified populations.

θ_D^i is average over distinct allele pairs of the within-population allele-pair probabilities. Average over populations of θ_D^i is θ_D.(For pairs of alleles, not pairs of genotypes.

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 $\theta^i = 0$

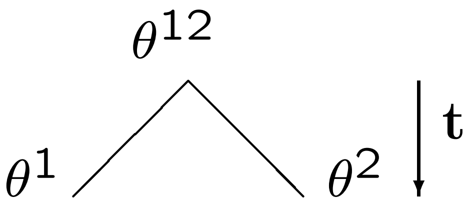
$$\theta^i(t) \ = \ 1 - \left(1 - \frac{1}{2N_i}\right)^t$$

The θ's are for distinct alleles in a population, so should be written as θ_D.

If t is small relative to large N_i's, $\theta^i(t) \approx t/(2N_i)$, and $\theta_D(t) \approx t/(2N_h)$ 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}:

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

Aside: Derivation of Equation for θ^i

If there is no mutation, and no migration between the populations:

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

This first-order difference equation has solution

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

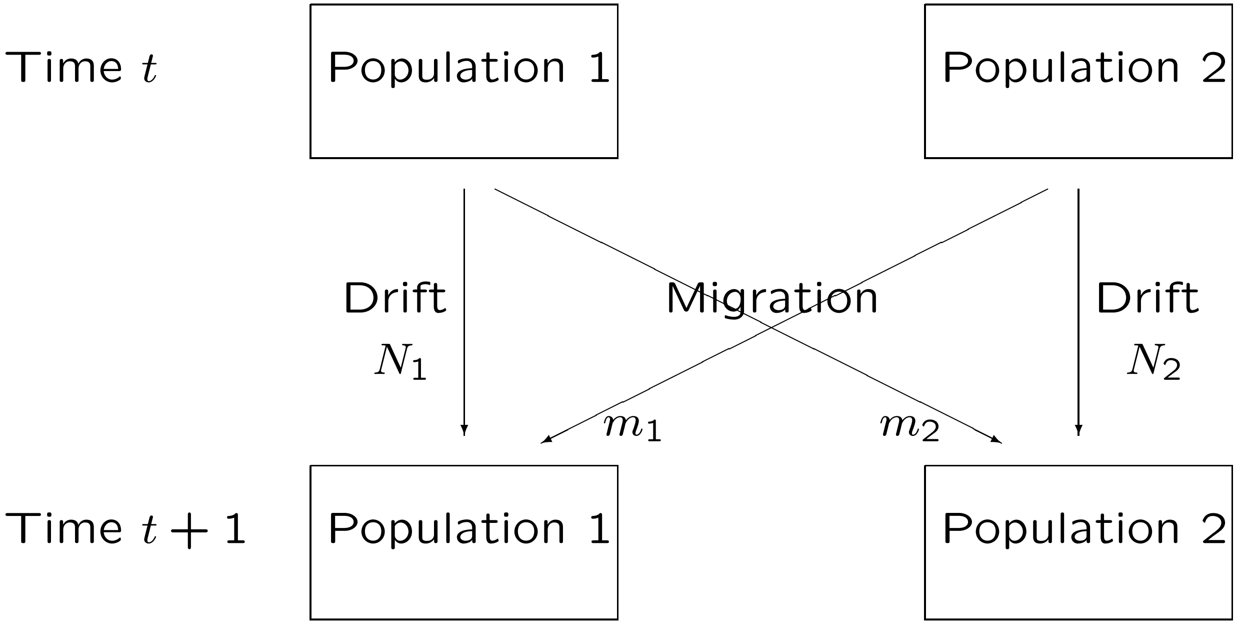
At $t = 0$, $\theta^i(0) = A + B = \theta^{12}$ and at $t = 1$, $\theta^i(1) = A + B[1 - 1/(2N_i)]$ so $B = (1 - \theta^{12}), A = 1$.

Genetic Distance between Two Populations

It is the average relative kinship $(\psi^i + \psi^{i'})/2$ rather than the average ibd probability $(\theta^i + \theta^{i'})/2$ that serves as a measure of distance between pairs of populations.

The ψ's have the same ranking as the θ's.

Two populations: drift, migration, mutation



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

Aside: Drift, Mutation and Migration

It is possible to predict the values of $\theta^i, \theta^{ii'}$ and, therefore, the values of $\psi^i = (\theta^i - \theta_B)/(1 - \theta_B)$.

For two populations, $\theta_B = \theta^{12}$. Although $\theta^1, \theta^2, \theta^{12}$ are all non-negative probabilities, it is possible that both of $\psi^1 = (\theta^1 - \theta^{12})/(1 - \theta^{12})$ and $\psi^2 = (\theta^2 - \theta^{12})/(1 - \theta^{12})$ are positive, or that one of them is negative and the other one positive. The average $(\psi^1 + \psi^2)/2$ is non-negative.

Aside: Drift, Mutation and Migration

For populations 1 or 2 with sizes N_1 or N_2, if m_i are the proportions of alleles in population i that have migated from the other population, the changes in the θ's from generation t to $t+1$ are

$$\begin{aligned}
\theta^1(t+1) &= (1-\mu)^2\Big[(1-m_1)^2\phi^1(t)+2m_1(1-m_1)\theta^{12}(t) \\
&\quad +m_1^2\phi^2(t)\Big] \\
\theta^2(t+1) &= (1-\mu)^2\Big[m_2^2\phi^1(t)+2m_2(1-m_2)\theta^{12}(t) \\
&\quad +(1-m_2)^2\phi^2(t)\Big] \\
\theta^{12}(t+1) &= (1-\mu)^2\Big[(1-m_1)m_2\phi^1(t)+[(1-m_1)(1-m_2) \\
&\quad +m_1m_2]\theta^{12}(t)+m_1(1-m_2)\phi^2(t)\Big]
\end{aligned}$$

where $\phi^i(t)=1/(2N_i)+(2N_i-1)\theta^i(t)/(2N_i)$ and μ is the infinite-allele mutation rate.

Drift and Mutation

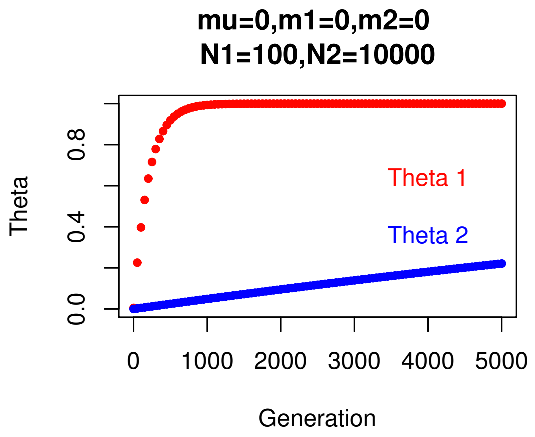
If there is no migration, the θ's tend to equilibrium values of

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

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

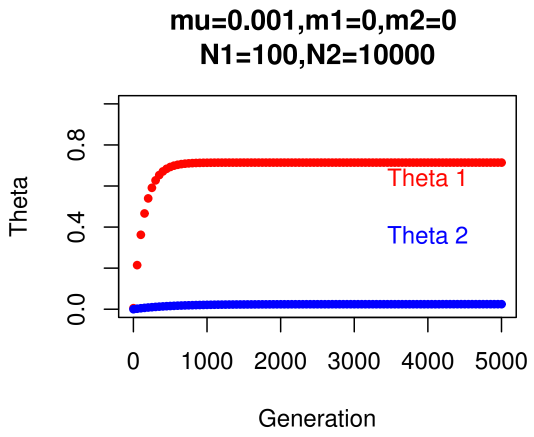
Drift, Mutation and Migration

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



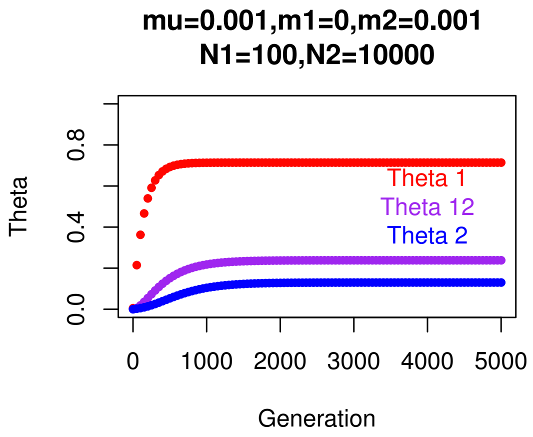
Drift Only

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



Drift and Mutation

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



Drift, Mutation and Migration

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

Multiple Populations

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

$$\psi_{DT}^i = \frac{\theta_D^i - \theta_B}{1 - \theta_B}$$

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

Averaging over populations:

$$\psi_{DT} = \frac{\theta_D - \theta_B}{1 - \theta_B}$$

and this is the global "F_{ST}" for alleles. Under HWE, $\theta_D = \theta_S$, where θ_S refers to pairs of alleles from distinct individuals, and we have $F_{ST} = (\theta_S - \theta_B)/(1 - \theta_B)$.

Estimators for Populations

Allelic Matching Proportions for Individuals

The previous discussion on inbreeding and relatedness was phrased in terms of allele-matching and heterozygosity.

We showed the allele-matching quantities $\tilde{A}_j$ within individuals and $\tilde{A}_{jj'}$ between individuals.

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

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

Note value of 0.5 for two heterozygotes. Different from value of 1 for "number of ibs pairs of alleles."

In terms of allele dosages:

$$\begin{aligned}
\tilde{A}_j = (X_j - 1)^2 \quad &, \quad \mathcal{E}(\tilde{A}_j) = A + (1-A)F_j \\
\tilde{A}_{jj'} = \frac{1}{2}[1 + (X_j - 1)(X_{j'} - 1)] \quad &, \quad \mathcal{E}(\tilde{A}_{jj'}) = A + (1-A)\theta_{jj'} \\
A \;\; &= \;\; 1 - 2\pi(1-\pi)
\end{aligned}$$

Allelic Matching Proportions for Pairs of Individuals

Averaging over pairs of distinct individuals:

$$\begin{aligned}\tilde{A}_S &= \frac{1}{n(n-1)} \sum_{\substack{j=1 \\ j\neq j'}}^{n} \sum_{j'=1}^{n} \tilde{A}_{jj'} \approx 1 - 2\tilde{p}(1-\tilde{p}) \\ \mathcal{E}(\tilde{A}_S) &= A + (1-A)\theta_S\end{aligned}$$

$\tilde{A}_S$ is for genotypic data.

ISISIThe allele sharing kinship estimators and their expected values are

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

Note the the standard kinship estimators and their expected values are

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

Allelic Matching Proportions for Pairs of Populations

The average of $\widehat{\psi}_{jj'}$ values for a sample from one population is zero by the way they were constructed. To compare the amounts $\tilde{A}_S^i$ of within-population allele sharing among populations i we use the average between-population sharing $\tilde{A}_B$ as a reference. Between populations i, i' the allele-sharing proportion is

$$\tilde{A}_B^{ii'} \;=\; \tilde{p}_i\tilde{p}_{i'} + (1-\tilde{p}_i)(1-\tilde{p}_{i'})$$

and averaging over pairs of populations

$$\begin{aligned}
\tilde{A}_B &= \frac{1}{r(r-1)}\sum_{\substack{i=1 \\ i\neq i'}}^{r}\sum_{i'=1}^{r}\tilde{A}_B^{ii'} \\
&= 1-\frac{r}{r-1}2\bar{p}(1-\bar{p})+\frac{1}{r(r-1)}\sum_{i=1}^{r}2\tilde{p}_i(1-\tilde{p}_i)
\end{aligned}$$

F_{ST}

The population-specific F_{ST} values are

$$\begin{aligned} \widehat{F}_{ST}^i &= \frac{\tilde{A}_S^i - \tilde{A}_B}{1 - \tilde{A}_B} \\ \mathcal{E}(\widehat{F}_{ST}^i) &= \frac{\theta_S^i - \theta_B}{1 - \theta_B} \end{aligned}$$

and the global values (for genotypic data) are

$$\begin{aligned} \widehat{F}_{ST} &= \frac{1}{r}\sum_{i=1}^{r} \widehat{F}_{ST}^i \\ \mathcal{E}(\widehat{F}_{ST}) &= \frac{\theta_S - \theta_B}{1 - \theta_B} \end{aligned}$$

F_{IS} and F_{IT}

With genotypic data, there are two other F-statistics. The allele-sharing estimators for one population are

$$\hat{F}^i_{IS} = \frac{\tilde{A}^i_I - \tilde{A}_S}{1 - \tilde{A}_S} \quad , \quad \mathcal{E}(\hat{F}^i_{IS}) = \frac{F^i_I - \theta_S}{1 - \theta_S}$$

$$\hat{F}^i_{IT} = \frac{\tilde{A}^i_I - \tilde{A}_B}{1 - \tilde{A}_B} \quad , \quad \mathcal{E}(\hat{F}^i_{IT}) = \frac{F^i_I - \theta_B}{1 - \theta_B}$$

where $\tilde{A}^i_I = \sum_{j=1}^{n_i}(X^i_j - 1)^2/n_i$.

Taking averages over populations gives global values

$$\hat{F}_{IS} = \frac{\tilde{A}_I - \tilde{A}_S}{1 - \tilde{A}_S} \quad , \quad \mathcal{E}(\hat{F}_{IS}) = \frac{F_I - \theta_S}{1 - \theta_S}$$

$$\hat{F}_{IT} = \frac{\tilde{A}_I - \tilde{A}_B}{1 - \tilde{A}_B} \quad , \quad \mathcal{E}(\hat{F}_{IT}) = \frac{F_I - \theta_B}{1 - \theta_B}$$

Allelic Matching Proportions Within Populations

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

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

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

$$\begin{aligned}\tilde{A}^i_{Wl} &= \tilde{p}^2_{il} + (1-\tilde{p}_{il})^2 = 1 - 2\tilde{p}_{il}(1-\tilde{p}_{il}) \\ \tilde{A}^i_{Dl} &= 1 - \frac{2n_i}{2n_i - 1} 2\tilde{p}_{il} 1 - \tilde{p}_{il})\end{aligned}$$

These have expected values that follow the pattern:

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

Aside: Exact Allelic Sharing Proportions

If the sample has $2n_{il}$ alleles at SNP l, and if r_{il} of these are the designated type, the observed sharing proportion of distinct allele pairs within this sample, is

$$\begin{aligned}\tilde{A}^i_{Dl} &= \frac{1}{2n_{il}(2n_{il}-1)}[r_{il}(r_{il}-1)+(2n_{il}-r_{il})(2n_{il}-r_{il}-1)] \\ &= \frac{2n_{il}}{2n_{il}-1}\left[\frac{r_{il}}{2n_{il}}\left(\frac{r_{il}}{2n_{il}}-\frac{1}{2n_{il}}\right)+(1-\frac{r_{il}}{2n_{il}})\left(1-\frac{r_{il}}{2n_{il}}-\frac{1}{2n_{il}}\right)\right] \\ &= 1-\frac{2n_{il}}{2n_{il}-1}2\tilde{p}_{il}(1-\tilde{p}_{il})\end{aligned}$$

where $\tilde{p}_{il}$ is the sample frequency for the designated allele for this population.

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

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

Allele-based Estimate of F_{ST}

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

$$\widehat{\psi}^i_{DT} = \widehat{F}^i_{DT} \;\; = \;\; \frac{\sum_l(\tilde{A}^i_{Dl} - \tilde{A}_{Bl})}{\sum_l(1 - \tilde{A}_{Bl})}$$

and this has expected value $F^i_{DT} = (\theta^i_D - \theta_B)/(1 - \theta_B)$ which is the population-specific value for allelic data.

Average over populations:

$$\widehat{F}_{DT} = \widehat{\psi}_{DT} \;\; = \;\; \frac{\sum_l(\tilde{A}_{D_l} - \tilde{A}_{B_l})}{\sum_l(1 - \tilde{A}_{B_l})}$$

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

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

$$\widehat{\psi}^i_{DT} = \widehat{F}^i_{DT} \;=\; \frac{\sum_l(\tilde{A}^i_{Dl} - \tilde{A}_{Bl})}{\sum_l(1 - \tilde{A}_{Bl})}$$

and this has expected value $F^i_{DT} = (\theta^i_D - \theta_B)/(1 - \theta_B)$ 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 almost surely to the parametric value F^i_{DT}.

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

Alternative Computing Equations for F_{WT}

Now consider random pairs of alleles with $\tilde{A}_W$'s, rather random pairs of distinct alleles with $\tilde{A}_D$'s, For all sample sizes and r populations:

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

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 sample variance of sample allele frequencies over populations.

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

Alternative Estimates for F_{WT}

The population-specific estimates are

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

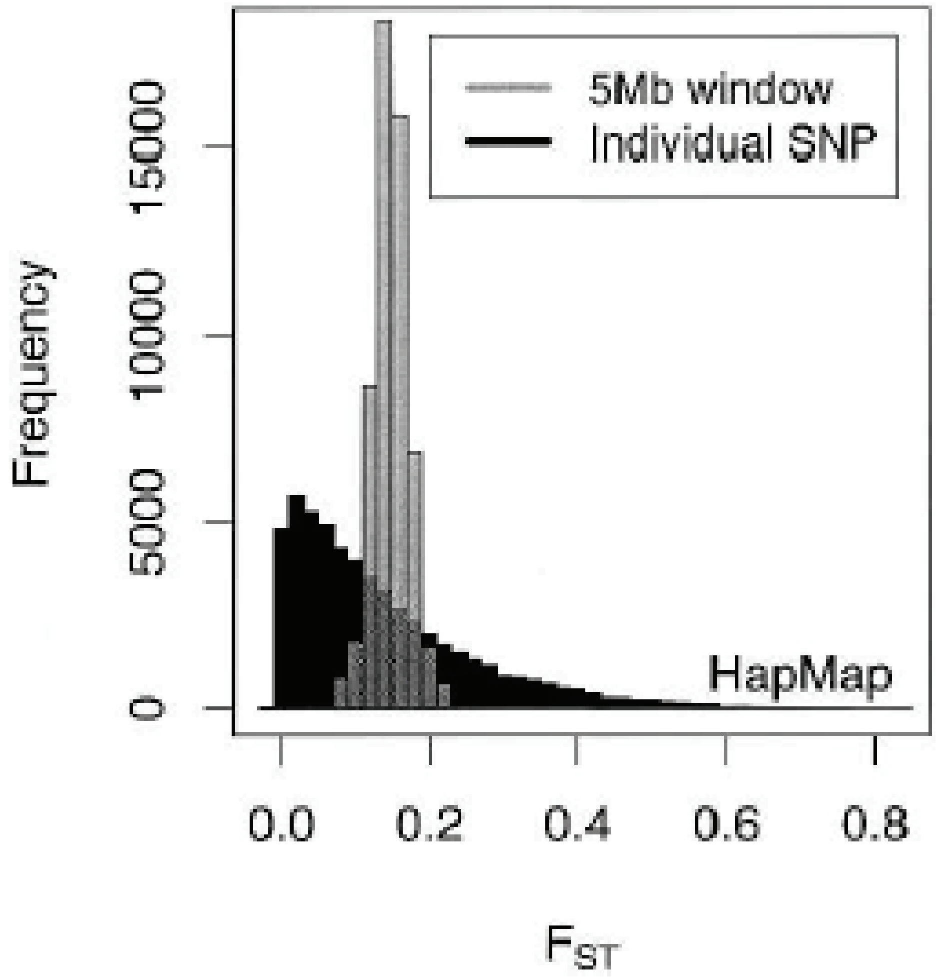
The global estimates are

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

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

But we would prefer to use genotypic data to avoid confounding with inbreeding and the effects of the number of populations and sample sizes.

Effect of Number of Loci



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

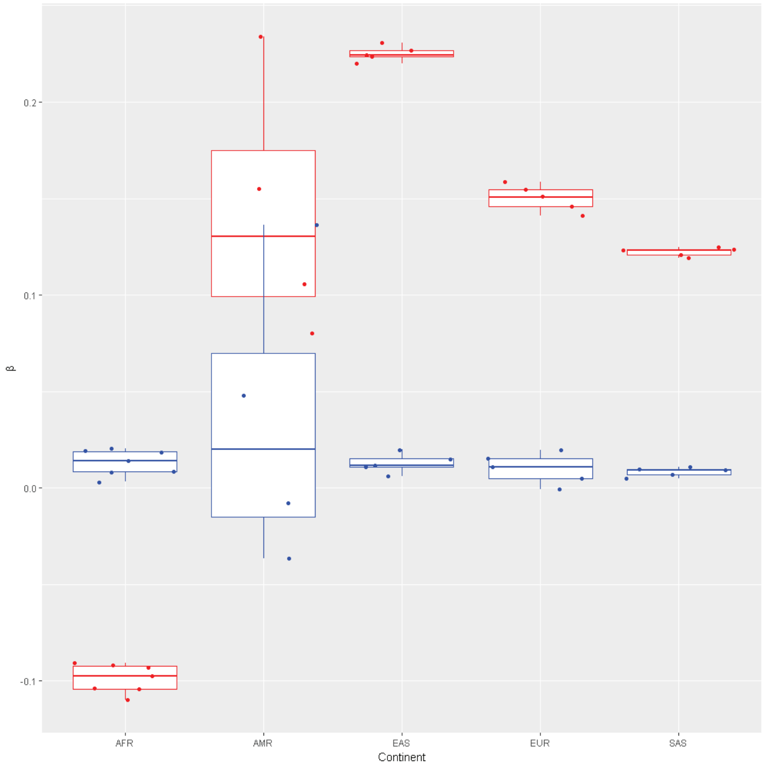
F_{ST} is relative, not absolute

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_{ST}, $\widehat{\psi}_{ST} = 0.013$, with Africa as a reference set than there is, $\widehat{\psi}_{ST} = -0.099$, 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_{ST}, $\widehat{\psi}_{ST} = 0.013$ with East Asia as a reference set than there is, $\widehat{\psi}_{ST} = 0.225$ 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



AFR AMR EAS EUR SAS

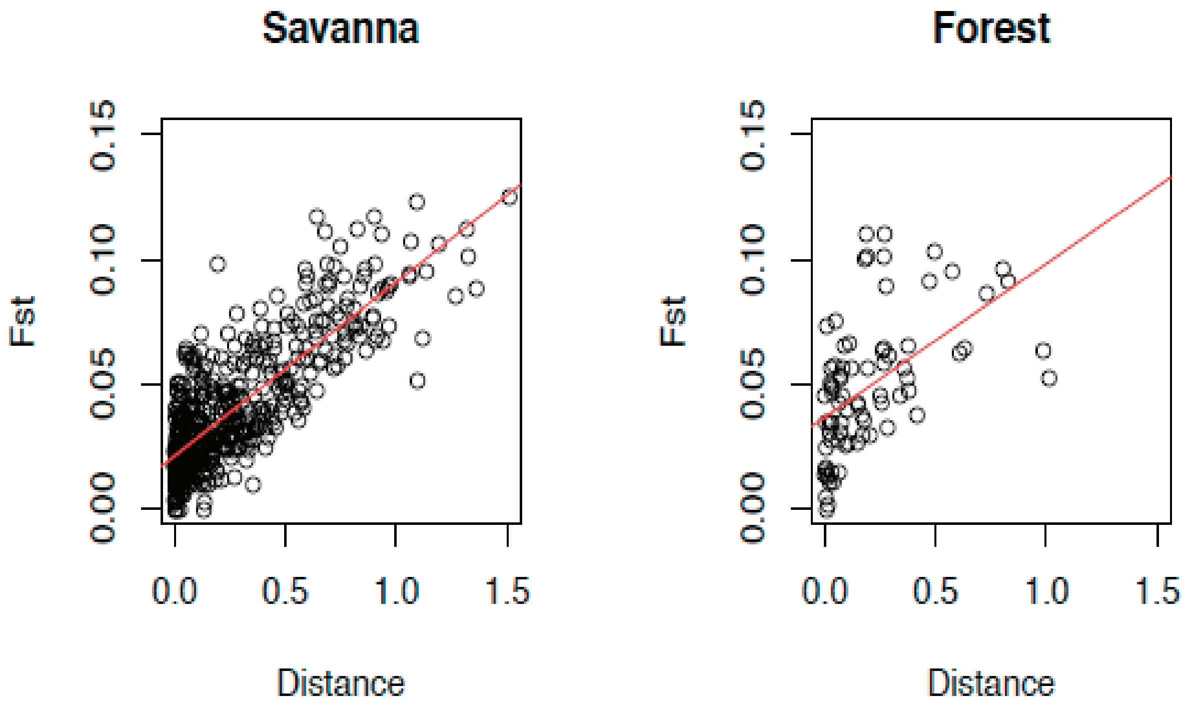
Blue box: Population relative to pairs of populations in same continent.

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

Evolutionary Inferences

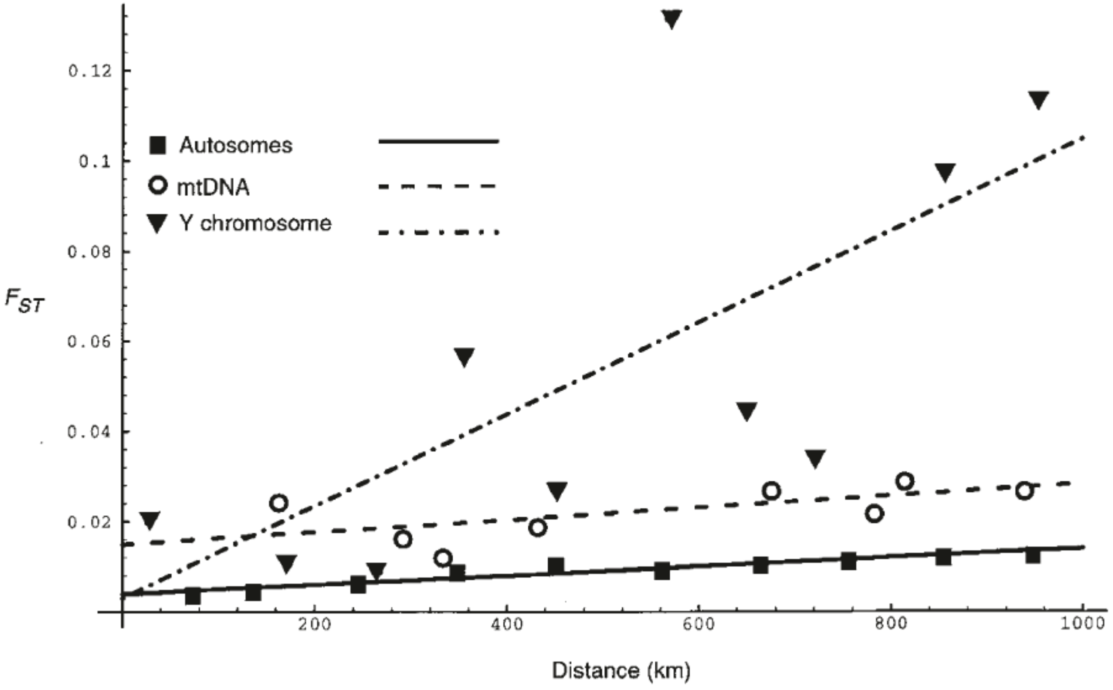
Aside: Geographic and Genetic Distances

From earlier slides, equilibrium values of F_{ST} for pairs of populations serve as measures of genetic distance between populations, and so may reflect geographic distances also.



Wasser S, et al. 2015. Science 349:84-87.

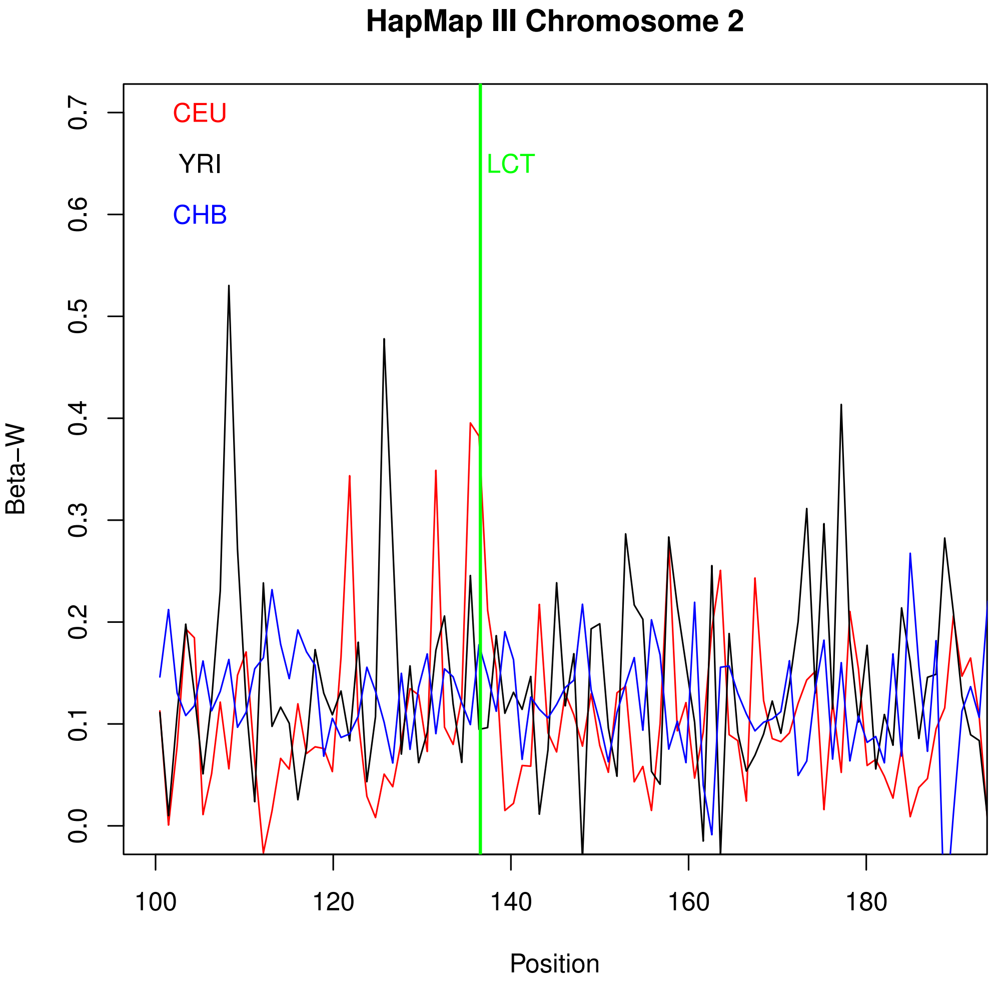
Aside: Human Migration Rates



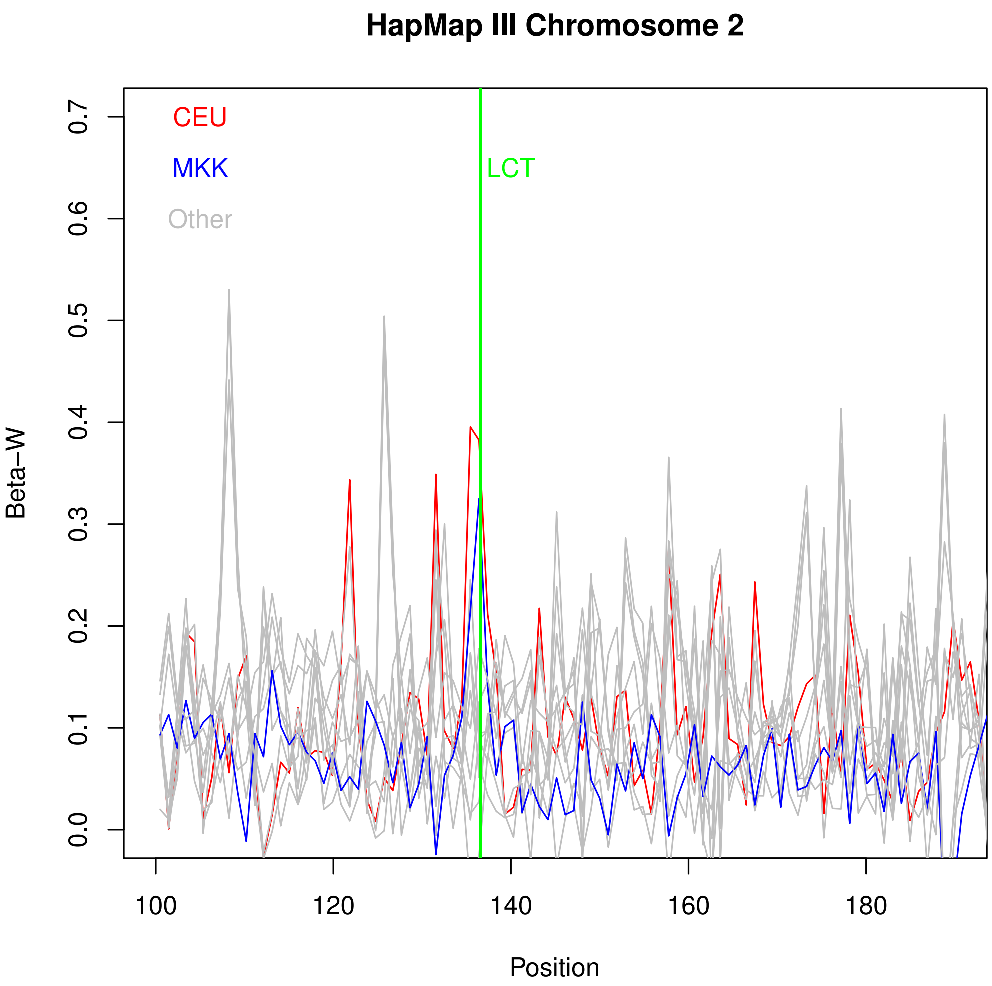
Suggests higher migration rate for human females among 14 African populations.

Seielstad MT, et al. 1998. Nature Genetics 20:278-280.

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



$\hat{\beta}_{WT}$ in LCT Region: 11 Populations



MKK Population

"The Maasai are a pastoral people in Kenya and Tanzania, whose traditional diet of milk, blood and meat is rich in lactose, fat and cholesterol. In spite of this, they have low levels of blood cholesterol, and seldom suffer from gallstones or cardiac diseases.

Analysis of HapMap 3 data using Fixation Index (Fst) identified genomic regions and single nucleotide polymorphisms (SNPs) as strong candidates for recent selection for lactase persistence and cholesterol regulation in 143156 founder individuals from the Maasai population in Kinyawa, Kenya (MKK). The strongest signal identified by all three metrics was a 1.7 Mb region on Chr2q21. This region contains the gene LCT (Lactase) involved in lactase persistence."

Wagh et al. 2012. PLoS One 7: e44751

Aside: F-statistics from Heterozygosities

Nei defined three “heterozygosities” H_0, H_S, H_T in terms of population allele frequencies p_i and they estimated Wright’s F-statistics with ratios of estimators for these.

$$
\begin{aligned}
\tilde{H}_0 &= 1 - \tilde{A}_I = 1 - \tilde{H}_I \\
\tilde{H}_S &= 1 - \tilde{A}_W = \frac{1}{r}\sum_{i=1}^{r} 2\tilde{p}_i(1-\tilde{p}_i) \\
\tilde{H}_T &= 1 - \tilde{A}_T = 2\bar{p}(1-\bar{p})
\end{aligned}
$$

Aside: Nei and Chesser

To remove sample-size effects, Nei and Chesser modified these to

$$\begin{aligned}
\widehat{H}_0 &= 1 - \tilde{A}_I \\
\widehat{H}_S &= 1 - \tilde{A}_S \\
\widehat{H}_T &= 1 - \tilde{A}_T + \frac{1}{2r^2}\sum_{i=1}^{r}\frac{1}{n_i}(1 + \tilde{A}_I - 2\tilde{A}_S)
\end{aligned}$$

Their F-statistics estimators are

$$\widehat{F}_{IS} = 1 - \frac{\widehat{H}_0}{\widehat{H}_S}; \widehat{F}_{IT} = 1 - \frac{\widehat{H}_0}{\widehat{H}_T}; \widehat{F}_{ST} = 1 - \frac{\widehat{H}_S}{\widehat{H}_T}$$

Aside: F-statistics from Heterozygosities

Although these estimators look simple and intuitive, they have expectations that depend on the number of populations:

$$\mathcal{E}(\hat{F}_{IS}) = \frac{F_I - \theta_S}{1 - \theta_S}$$

$$\mathcal{E}(\hat{F}_{IT}) = \frac{(F_I - \theta_B) - \frac{1}{r}(\theta_S - \theta_B)}{(1 - \theta_B) - \frac{1}{r}(\theta_S - \theta_B)}$$

$$\mathcal{E}(\hat{F}_{ST}) = \frac{(\theta_S - \theta_B) - \frac{1}{r}(\theta_S - \theta_B)}{(1 - \theta_B) - \frac{1}{r}(\theta_S - \theta_B)}$$

Aside: Weir & Cockerham 1984 Model

W&C assumed all populations have equal evolutionary histories ($\theta_S^i = \theta$, all i) and are independent ($\theta_B^{ii'} = 0$, all $i' \neq i$), and they worked with overall allele frequencies that were weighted by sample sizes

$$\bar{p}_l = \frac{1}{\sum_i n_i} \sum_i n_i \tilde{p}_{il}$$

Aside: Weir & Cockerham 1984 Model: Genotypic Data

WC84 defined three mean squares:

$$\begin{aligned}\text{MSP} &= \frac{n_T}{r-1}(\tilde{A}^*_W - \tilde{A}^*_T)\\ \text{MSI} &= \frac{n_T}{n_T - r}[\frac{1}{2}(1+\tilde{A}^*_I) - \tilde{A}^*_W]\\ \text{MSG} &= \frac{1}{2}(1-\tilde{A}^*_I)\end{aligned}$$

where $n_T = \sum_{i=1}^n n_i$, $\tilde{A}^*_I = \sum_{i=1}^r n_i \tilde{A}^i_I / n_T$, $\tilde{A}^*_W = \sum_{i=1}^r n_i \tilde{A}^i_W / n_T$, $\tilde{A}^*_T = \sum_{i=1}^r n_i \tilde{A}^i_T / n_T$.

The WC84 F-statistics estimators are

$$\begin{aligned}\hat{F}_{IS} &= \frac{\text{MSI} - \text{MSG}}{\text{MSI} + \text{MSG}}\\ \hat{F}_{IT} &= \frac{(\text{MSP} - \text{MSI}) + n_c(\text{MSI} - \text{MSG})}{(\text{MSP} - \text{MSI}) + n_c(\text{MSI} + \text{MSG})}\\ \hat{F}_{ST} &= \frac{(\text{MSP} - \text{MSI})}{(\text{MSP} - \text{MSI}) + n_c(\text{MSI} + \text{MSG})}\end{aligned}$$

Aside: Weir & Cockerham 1984 Model: Genotypic Data

In general, the WC84 estimators of F_{IS}, F_{IT}, F_{ST} have complicated expected values depending on the dimensions of the data. In the special case of the WC84 model of $\theta_S^i = \theta_S, F_I^i = F_I$, or if the sample sizes are the same, the expected values are the same as for allele-sharing:

$$\begin{aligned}
\mathcal{E}(\hat{F}_{IS}) &= \frac{F_I - \theta_S}{1 - \theta_S} \\
\mathcal{E}(\hat{F}_{IT}) &= \frac{F_I - \theta_B}{1 - \theta_B} \\
\mathcal{E}(\hat{F}_{ST}) &= \frac{\theta_S - \theta_B}{1 - \theta_B}
\end{aligned}$$

With the further assumption of independent populations, $\theta_B = 0$:

$$\begin{aligned}
\mathcal{E}(\hat{F}_{IS}) &= \frac{F_I - \theta_S}{1 - \theta_S} \\
\mathcal{E}(\hat{F}_{IT}) &= F_I \\
\mathcal{E}(\hat{F}_{ST}) &= \theta_S
\end{aligned}$$

Aside: Weir & Cockerham 1984 Model: Allelic Data

For allelic data, the mean square between populations is the same, MSP, as for genotypic data. The mean square within populations, MSW, is a combination of MSI and MSG for genotypic data:

$$\text{MSW} = \frac{n_T}{2n_T - r}(1 - \tilde{A}^*_W)$$

In the special case of Hardy-Weinberg equilibrium (HWE) $F_I = \theta_S$ and

$$\begin{aligned}\mathcal{E}(\text{MSB}) &= p_l(1-p_l)[(1-\theta_S) + n_c\theta_S] \\ \mathcal{E}(\text{MSW}) &= p_l(1-p_l)(1-\theta_S)\end{aligned}$$

where $n_c = (\sum_i n_i - \sum_i n_i^2 / \sum_i n_i)/(r-1)$.

Aside: Weir & Cockerham 1984 Model: Allelic Data

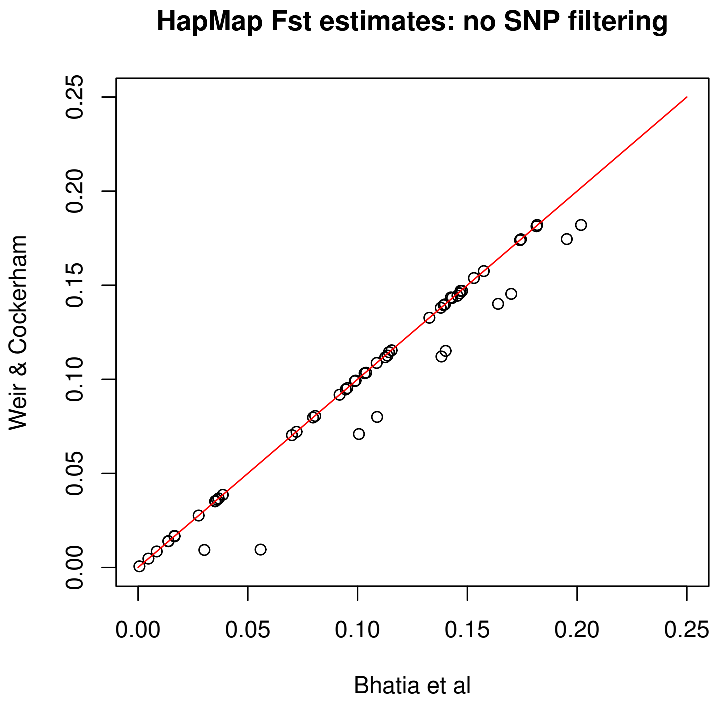
The WC84 allele-based estimator of θ_S (or F_{ST}) is

$$\hat{\theta}_S = \frac{\sum_l (\text{MSB} - \text{MSW})}{\sum_l (\text{MSB} + (n_c - 1)\text{MSW})}$$

and, when there is HWE, it has an expected value of $(\theta_S - \theta_B)/(1 - \theta_B)$ as for the allele-sharing estimator.

If there is not HWE, this allele-based estimator is of a mixture of inbreeding and coancestry values.

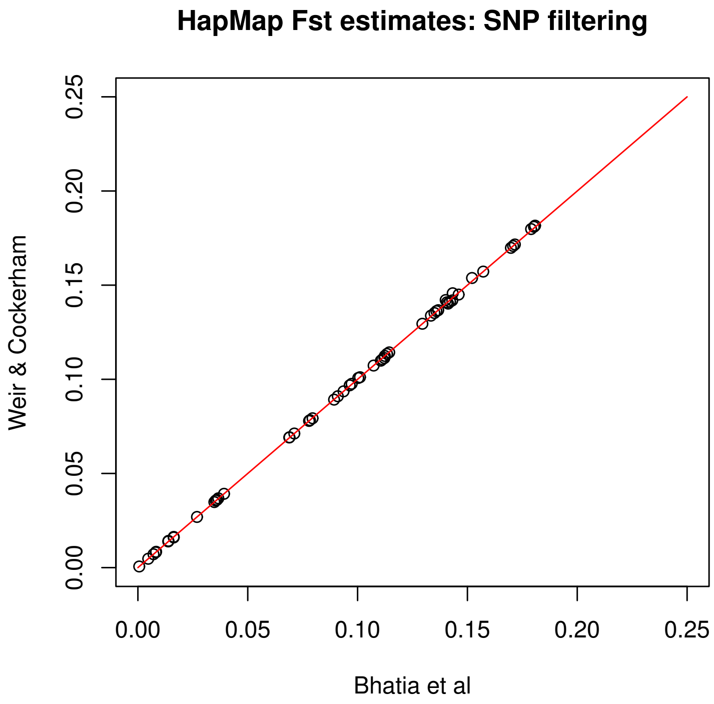
Aside: WC84 vs Allele-sharing Estimators (allelic data)



F_{WT} estimates for HapMap III, using all 87,592 SNPs on chromosome 1.

Bhatia et al, 2013, Genome Research 23:1514-1521.

Aside: WC84 vs Allele-sharing Estimators (allelic data)



F_{WT} estimates for HapMap III, using the 42,463 SNPs on chromosome 1 that have at least five copies of the minor allele in samples from all 11 populations.

WC vs WG

Between Populations

Anova: $\text{SSP} = 2\sum_{i=1}^{r} n_i(\tilde{p}_i - \bar{p}^*)^2 = 2n_T\bar{p}^*(1-\bar{p}^*) - \sum_{i=1}^{r} 2n_i\tilde{p}_i(1-\tilde{p}_i)$

Het: $\text{SSP} = n_T(\tilde{H}_T^* - \tilde{H}_S^*)$

AS: $\text{SSP} = n_T(\tilde{A}_W^* - \tilde{A}_T^*)$

Equal: $\text{SSP} = rn(\tilde{A}_W - \tilde{A}_T)$

$\text{MSP} = \frac{1}{r-1}(\text{SSP})$

$\mathcal{E}(\text{MSP}) = \frac{\pi(1-\pi)}{r-1}\{\sum_{i=1}^{r}(1-\frac{n_i}{n_T})[(1-F_I^i)+2(F_I^i-2\theta_S^i)]+(r-1)n_c(\theta_S^{**}-\theta_B^*)\}$

Equal: $\mathcal{E}(\text{MSP}) = \pi(1-\pi)[(1-F_I)+2(F_I-\theta_S)+2n(\theta_S-\theta_B)]$

WC84: $\mathcal{E}(\text{MSP}) = \pi(1-\pi)[(1-F_I)+2(F_I-\theta_S)+2n_c(\theta_S-\theta_B)]$

WC vs WG

Between Individuals within Populations

Anova: SSI $= \sum_{i=1}^{r} [2n_i\tilde{p}_i(1-\tilde{p}_i) - \frac{1}{2}n_i\tilde{H}_I^i]$

Het: SSI $= n_T(\tilde{H}_S^* - \frac{1}{2}\tilde{H}_I^*)$

AS: SSI $= n_T[\frac{1}{2}(1+\tilde{A}_I^*) - \tilde{A}_W^*] = 2\sum_{i=1}^{r}(n_i-1)[(1-\tilde{A}_I^i) + 2(\tilde{A}_I^i - \tilde{A}_S^i)]$

Equal: SSI $= rn[\frac{1}{2}(1+\tilde{A}_I) - \tilde{A}_W]$

MSI $= \frac{1}{n_T - r}$(SSI)

$\mathcal{E}$(MSI) $= \frac{\pi(1-\pi)}{n_T-r}\sum_{i=1}^{r}(n_i-1)[(1-F_I^i) + 2(F_I^i - \theta_S^i)]$

Equal or WC84: $\mathcal{E}$(MSI) $= \pi(1-\pi)[(1-F_I) + 2(F_I - \theta_S)]$

WC vs WG

Within Individuals

Anova: SSG $= \frac{1}{2}\sum_{i=1}^{r} n_i \tilde{H}_I^i$

Het: SSG $= \frac{1}{2} n_T \tilde{H}_I^*$

AS: SSG $= n_T \frac{1}{2}(1 - \tilde{A}_I^*) = \frac{1}{2}\sum_{i=1}^{r} n_i (1 - \tilde{A}_I^i)$

Equal: SSG $= nr\frac{1}{2}(1 - \tilde{A}_I)$

MSG $= \frac{1}{n_T}$(SSG)

$\mathcal{E}$(MSG) $= \frac{\pi(1-\pi)}{n_T}\sum_{i=1}^{r} n_i (1 - F_I^i)$

Equal or WC84: $\mathcal{E}$(MSG) $= \pi(1-\pi)(1 - F_I)$

Forensic Match Probabilities

For loci at equilibrium under drift and mutation, allele frequencies p_A in populations follow a Beta distribution with parameters $(1-\theta)\pi_A/\theta$ and $(1-\theta)(1-\pi_A)/\theta$ where θ is the probability two distinct alleles in the population are ibd. This has mean π_A. This also assumes HWE.

As a consequence, the probability an allele is of type A given that n previous alleles have been found to have n_A of type A is

$$\Pr(A|n_A, n) = \frac{n_A\theta + (1-\theta)\pi_A}{1 + (n-1)\pi_A}$$

This allows the construction of conditional genotype probabilities. For example

$$\Pr(AA|AA) = \frac{\Pr(AAAA)}{\Pr(AA)} = \frac{\Pr(A)\Pr(A|A)\Pr(A|AA)\Pr(A|AAA)}{\Pr(A)\Pr(A|A)}$$

Forensic Match Probabilities

The match probabilities are

$$\begin{aligned}
\Pr(AA|AA) &= \frac{(3\theta + (1-\theta)\pi_A)(2\theta + (1-\theta)\pi_A}{(1+\theta)(1+2\theta)} \\
\Pr(AB|AB) &= \frac{2(\theta + (1-\theta)\pi_A)(\theta + (1-\theta)\pi_B}{(1+\theta)(1+2\theta)}
\end{aligned}$$

For example, if $\pi_A = \pi_B = 0.1, \theta = 0.03$
then $\Pr(AA) = 0.01, \Pr(AB) = 0.02$
but $\Pr(AA|AA) = 0.027, \Pr(AB|A) = 0.03$.

These equations require π's and θ. If sample allele frequencies $\tilde{p}$'s are used then it is appropriate to use estimates of ψ.

Worldwide Autosomal-STR Survey

Buckleton et al, Forensic Sci Int, 2016 compiled a survey of 250 published papers showing allele frequencies at 24 forensic STR markers from 446 populations in 8 ancestral groups. Represents data from 494,473 individuals.

The ancestral groups were identified by a combination of clustering and geographic criteria.

Moment estimates were obtained for each locus l in each population i from

$$\widehat{\psi}^{i}_{WT_l} = \frac{\tilde{M}^{i}_{W_l} - \tilde{M}_{T_l}}{1 - \tilde{M}_{T_l}}$$

The "T" may refer to the group of populations with the same continental ancestry, or it may refer to the entire set of populations.

STR Survey: $\hat{\psi}$ Values for Groups and Loci

	Geographic Region								
Locus	Africa	AusAb	Asian	Cauc	Hisp	IndPK	NatAm	Poly	Aver.
CSF1PO	0.003	0.002	0.008	0.008	0.002	0.007	0.055	0.026	0.011
D1S1656	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.000	0.011
D2S441	0.000	0.000	0.002	0.003	0.021	0.000	0.000	0.000	0.020
D2S1338	0.009	0.004	0.011	0.017	0.013	0.003	0.023	0.005	0.031
D3S1358	0.004	0.010	0.009	0.006	0.012	0.040	0.079	0.001	0.025
D5S818	0.002	0.013	0.009	0.008	0.014	0.018	0.044	0.007	0.029
D6S1043	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.016
D7S820	0.004	0.021	0.010	0.007	0.007	0.046	0.030	0.005	0.026
D8S1179	0.003	0.007	0.012	0.006	0.002	0.031	0.020	0.008	0.019
D10S1248	0.000	0.000	0.000	0.002	0.004	0.000	0.000	0.000	0.007
D12S391	0.000	0.000	0.000	0.003	0.020	0.000	0.000	0.000	0.010
D13S317	0.015	0.016	0.013	0.008	0.014	0.025	0.050	0.014	0.038
D16S539	0.007	0.002	0.015	0.006	0.009	0.005	0.048	0.004	0.021
D18S51	0.011	0.012	0.014	0.006	0.004	0.010	0.033	0.003	0.018
D19S433	0.009	0.001	0.009	0.010	0.014	0.000	0.022	0.014	0.023
D21S11	0.014	0.012	0.013	0.007	0.006	0.023	0.067	0.018	0.021
D22S1045	0.000	0.000	0.007	0.001	0.000	0.000	0.000	0.000	0.015
FGA	0.002	0.009	0.012	0.004	0.007	0.016	0.021	0.006	0.013
PENTAD	0.008	0.000	0.012	0.012	0.002	0.017	0.000	0.000	0.022
PENTAE	0.002	0.000	0.017	0.006	0.003	0.012	0.000	0.000	0.020
SE33	0.000	0.000	0.012	0.001	0.000	0.000	0.000	0.000	0.004
TH01	0.022	0.001	0.022	0.016	0.018	0.014	0.071	0.017	0.071
TPOX	0.019	0.087	0.016	0.011	0.007	0.018	0.064	0.031	0.035
VWA	0.009	0.007	0.017	0.007	0.012	0.022	0.028	0.005	0.023
All Loci	0.006	0.014	0.010	0.007	0.008	0.018	0.043	0.011	0.022

Asise: Multiple Alleles

Matching of two distinct alleles within individual j in popn i

$$\tilde{A}^i_j = \tfrac{1}{2}\sum_u X^i_{ju}(X^i_{ju}-1),\ \ \mathcal{E}(\tilde{A}^i_j) = A + (1-A)F^i_j$$

Average within-individual matching in popn i

$$\tilde{A}^i_I = \tfrac{1}{n_i}\sum_{j=1}^{n_i}\tilde{A}^i_j,\ \ \mathcal{E}(\tilde{A}^i_I) = A + (1-A)F^i_I$$

Average over popns of within-individual matching

$$\tilde{A}^I = \tfrac{1}{r}\sum_{i=1}^{r}\tilde{A}^i_I,\ \ \mathcal{E}(\tilde{A}^I) = A + (1-A)F^I$$

Matching of one allele from each of individuals j, j' in popn i

$$\tilde{A}^i_{jj'} = \tfrac{1}{4}\sum_u X^i_{ju}X^i_{j'u},\ \ \mathcal{E}(\tilde{A}^i_{jj'}) = A + (1-A)\theta^i_{jj'}$$

Average between-individual matching in popn i

$$\tilde{A}^i_S = \tfrac{1}{n_i(n_i-1)}\sum_{j=1}^{n_i}\sum_{j'=1,j'\neq j}^{n_i}\tilde{A}^i_{jj'},\ \ \mathcal{E}(\tilde{A}^i_S) = A + (1-A)\theta^i_S$$

Average over popns of between-individual within-popn matching

$$\tilde{A}^S = \tfrac{1}{r}\sum_{i=1}^{r}\tilde{A}^i_S,\ \ \mathcal{E}(\tilde{A}^S) = A + (1-A)\theta^S$$

Matching of two distinct alleles, ignoring genotypes, within popn i

$$\tilde{A}^i_W = \tfrac{2n_i}{2n_i-1}\sum_u \tilde{p}^2_{iu} - \tfrac{1}{2n_i-1},\ \ \mathcal{E}(\tilde{A}^i_W) = A + (1-A)\theta^i_W$$

Average over popns of within-popn allele matching, ignoring genotypes

$$\tilde{A}^W = \tfrac{1}{r}\sum_{i=1}^{1}\tilde{A}^i_W,\ \ \mathcal{E}(\tilde{A}^W) = A + (1-A)\theta^W$$

Matching of an allele from individual j in popn i with an allele from individual j' in popn i'.

$$\tilde{A}^{ii'}_{jj'} = \tfrac{1}{4}\sum_u X^i_{ju}X^{i'}_{j'u},\ \ \mathcal{E}(\tilde{A}^{ii'}_{jj'}) = A(1-A)\theta^{ii'}_{jj'}$$

Matching of one allele from each of popns i, i'

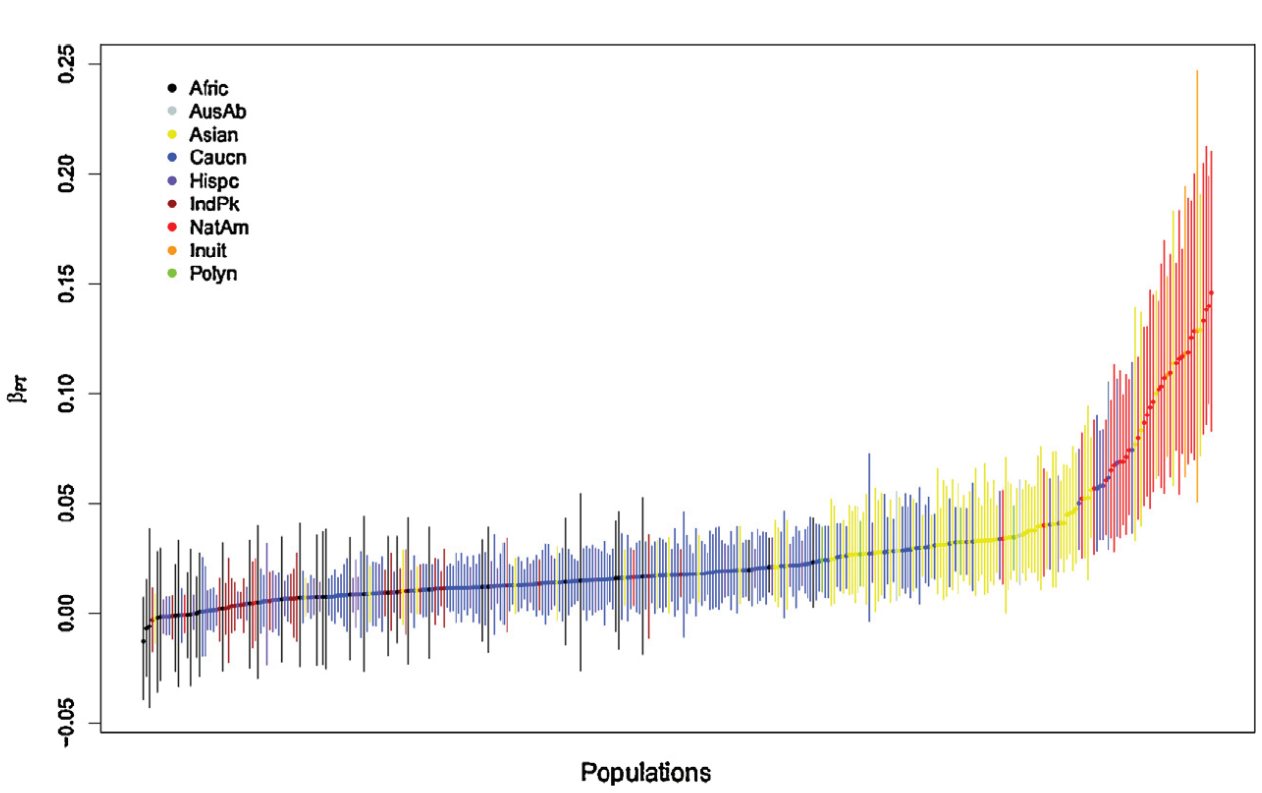
$$\tilde{A}^{ii'}_B = \tfrac{1}{n_n n_{i'}}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_{i'}}\tilde{A}^{ii'}_{jj'} = \sum_u \tilde{p}_{iu}\tilde{p}_{i'u},\ \ \mathcal{E}(\tilde{A}^{ii'}_B) = A + (1-A)\theta^{ii'}_B$$

Average over pairs of popns of between-popn-pair matching

$$\tilde{A}^B = \tfrac{1}{r(r-1)}\sum_{i=1}^{r}\sum_{i'=1,i'\neq i}^{r}\tilde{A}^{ii'}_B,\ \ \mathcal{E}(\tilde{A}^B) = A + (1-M)\theta^B$$

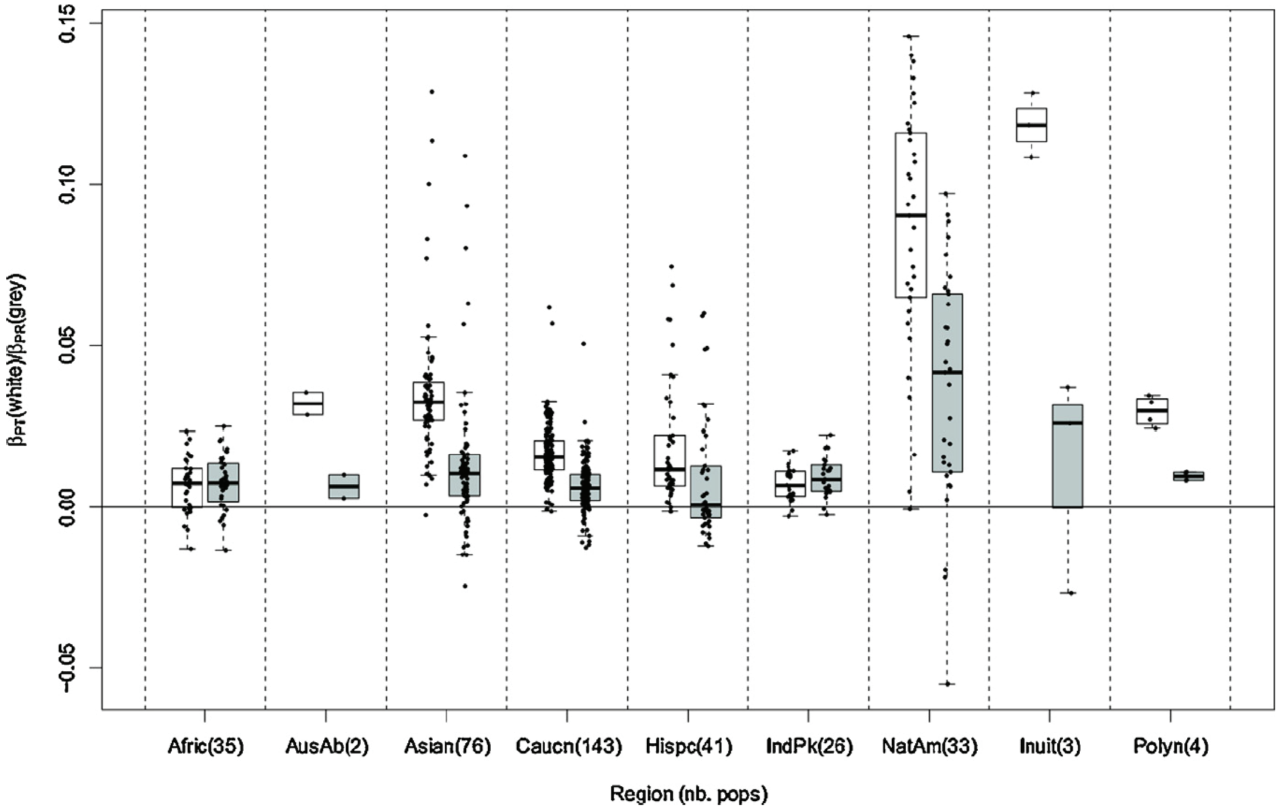
$X^ij u$ is dosage of allele u for individual j in population i.

Population-specific STR Estimates



Each vertical line represents one popn. The length of the line is the 95% confidence interval obtained by bootstrapping over loci.

F_{ST} is relative, not absolute



The white box plot is for within-popn matching compared to average matching among all pairs of popns. The grey box plot is for within-popn matching compared to average matching among all pairs of popns in the same region.