From Lecture 3

The allele sharing kinship estimators and their expected values are

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

The standard kinship estimators and their expected values are

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

POPULATION STRUCTURE

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."

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

Patterson et al., 2006

The allele dosage, for the reference or the variant allele, at SNP l and individual j is X_{jl}. The mean and variance, over individuals, of the dosages for that allele are $2p_l$ and $2p_l(1-p_l)(1+f)$ from above.

The allele frequency $\tilde{p}_l$ for a sample of n individuals is $\tilde{p}_l = \sum_{j=1}^{n} X_{jl}/2$. Patterson et al. "normalize" the 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 $\mathbf{K} = \mathbf{XX}'/\mathbf{L}$ has diagonal elements

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

and off-diagonal elements

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

These elements are measures of genetic similarity (averaged over SNPs) of pairs of individuals (including self-similarity).

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.

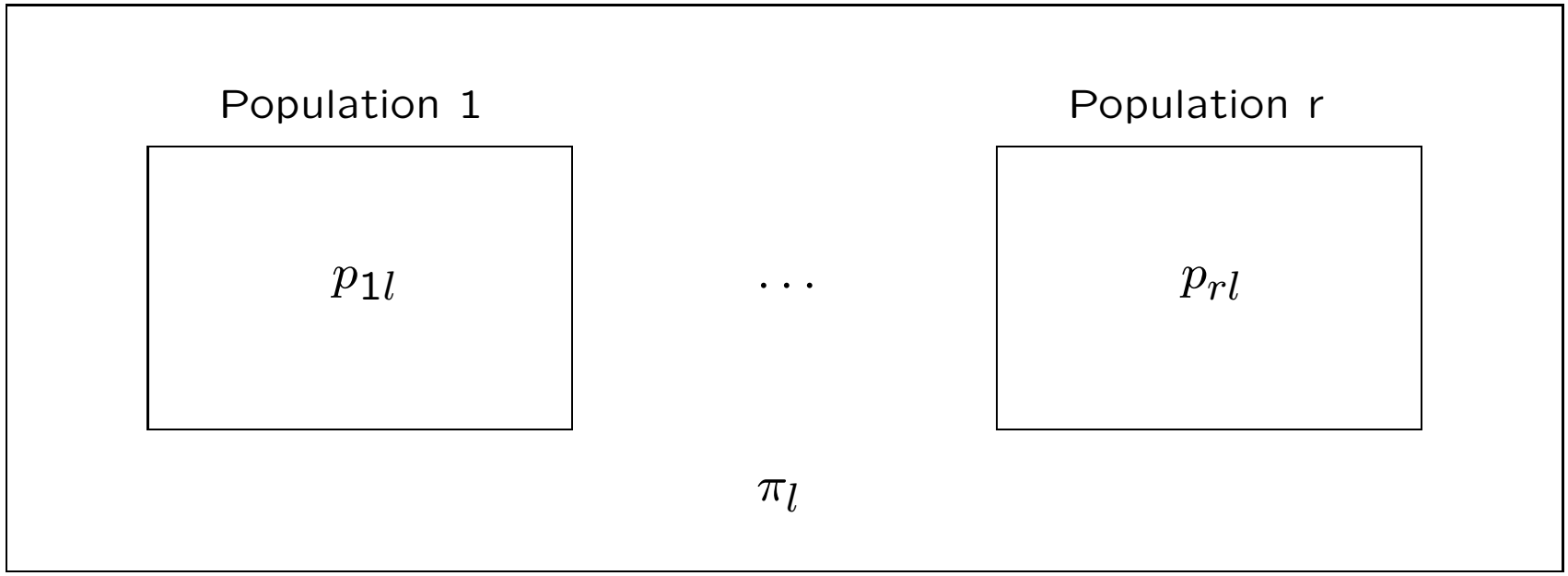
The eigenvalue decomposition of $\mathbf{K} = \mathbf{XX}'$ is

$$\mathbf{XX}' = \mathbf{UDU}'$$

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

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

Genetic Analysis: SNP l Allele Frequencies

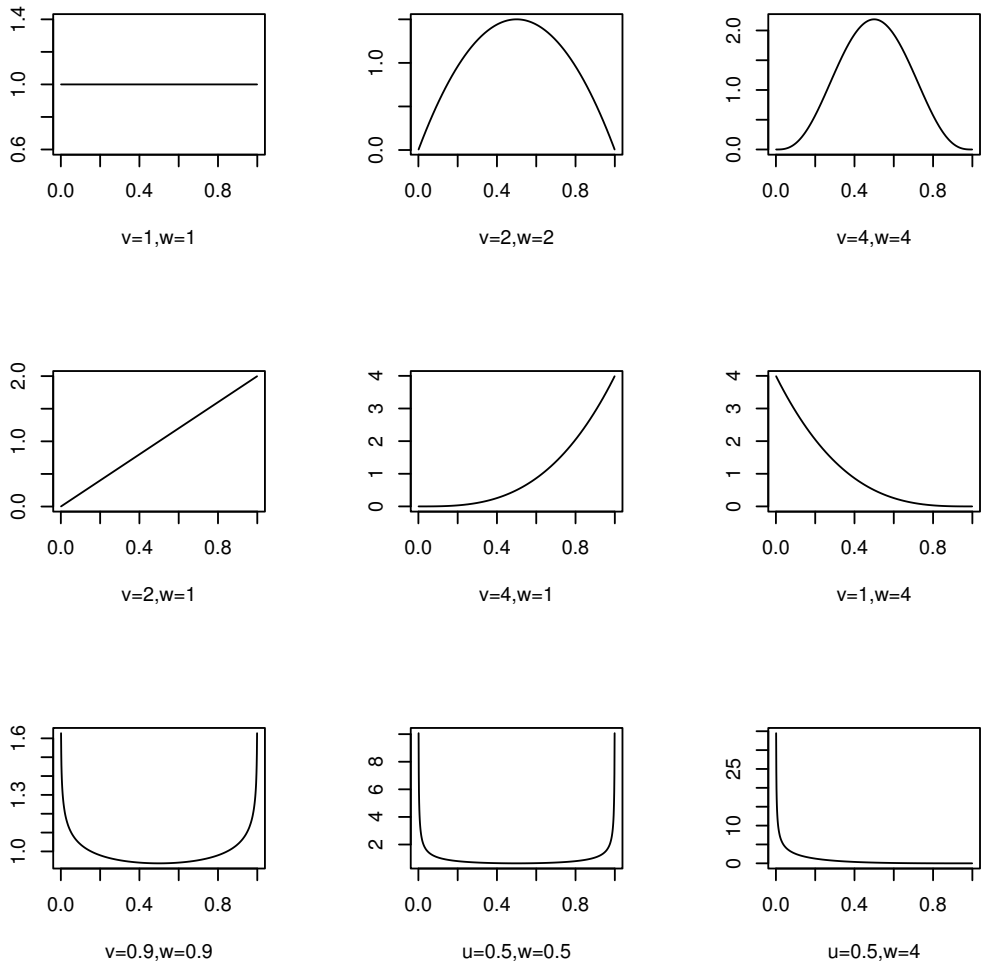


Among samples of n_i alleles from population i: counts for the SNP l reference allele follow a binomial distribution with mean p_{il} and variance $n_i p_{il}(1-p_{il})$. Sample allele frequencies $\tilde{p}_{il}$ have expected values p_{il} and (under HWE) variances $p_{il}(1-p_{il})/n_i$.

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.

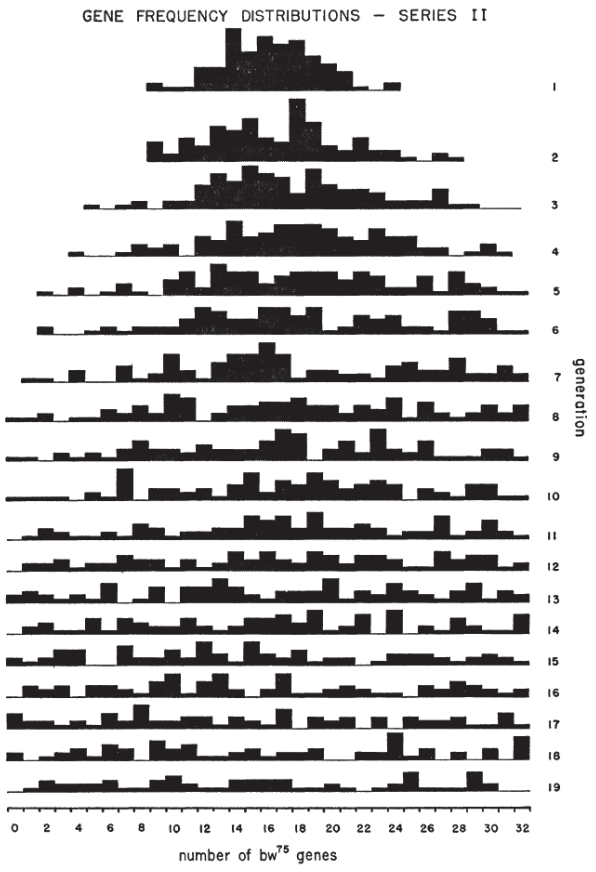
Beta distribution: Theoretical

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.

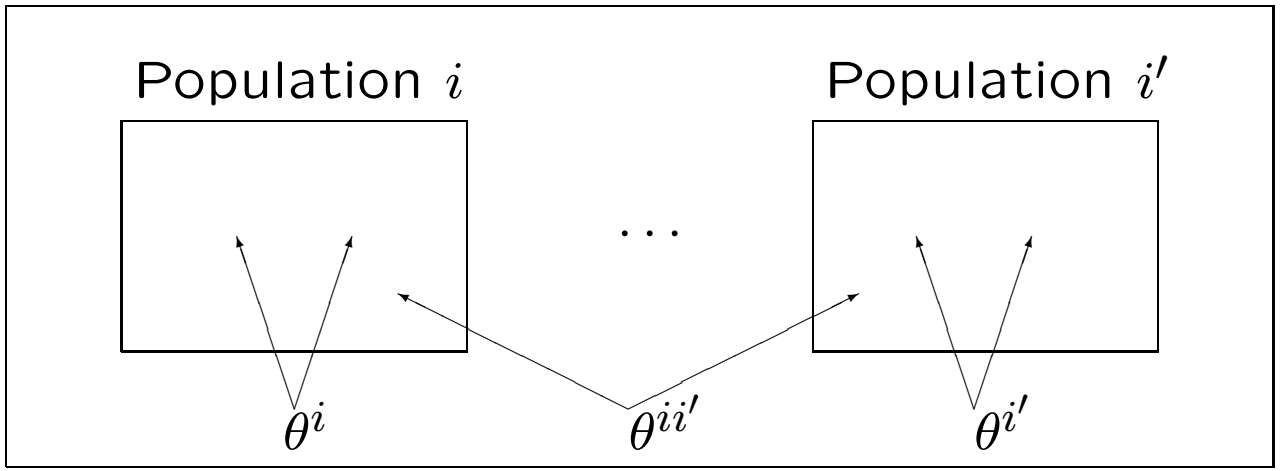
The target alleles may be in specified populations, and this leads to characterization of population structure, of they may be in specified individuals and this leads to characterization of inbreeding and relatedness.

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.

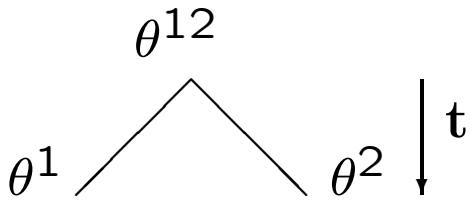
In the case of pure drift, where population i has constant size N_i and there is random mating, t generations after the population began drifting from an ancestral population in which $\theta^i = 0$

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

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):

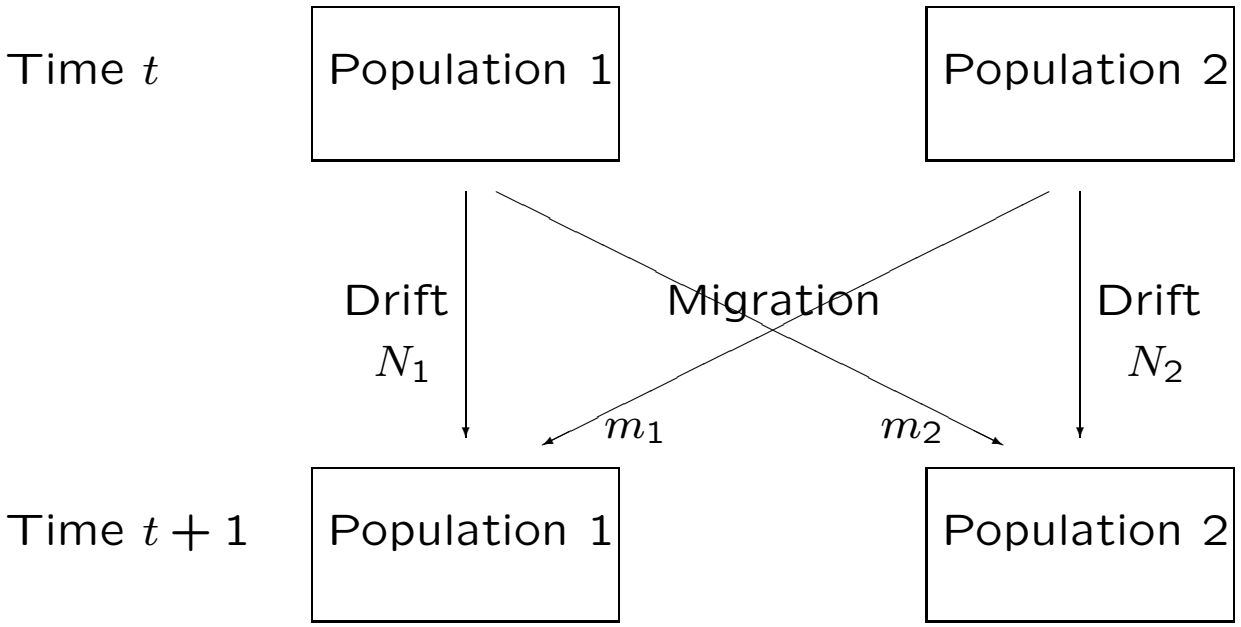


$$\theta^i \;=\; 1-(1-\theta^{12})\left(\frac{2N_i-1}{2N_i}\right)^t,\;\; i=1,2$$

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

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

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, 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_1 or m_2 are the proportions of alleles from population 2 or 1, 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

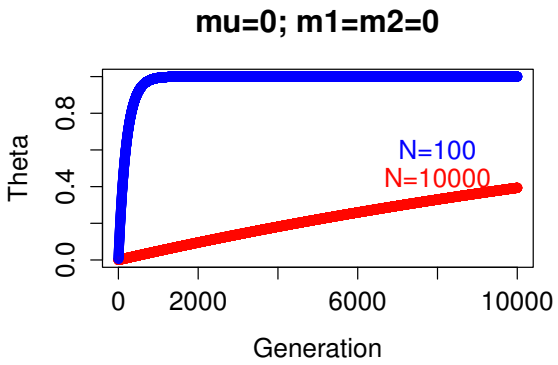
If there is no migration, the θ's tend to equilibrium values of

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

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

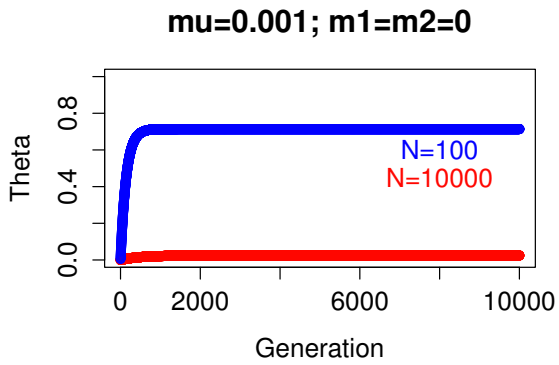
Drift, Mutation and Migration

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



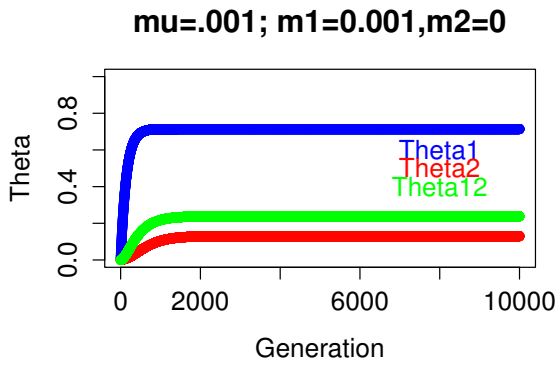
Drift Only

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



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, the ibd probability θ_W^i for any distinct pair of alleles within population i *relative to* the ibd probability between populations is

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

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

Averaging over populations:

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

and this is the global F_{WT} for alleles. This is the quantity often referred to as "F_{ST}".

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.

ψ^i is the probability two alleles drawn randomly from population i are ibd, and $\psi^{ii'}$ 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'.

Estimators for Populations

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 $\tilde{A}^i_{Wl}$ 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:

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

The expected value of this over replicates of the population is

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

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

Matching Proportions between Populations

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

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

The expected value of this over replicates of the population is

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

and, averaging over all pairs of populations

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

Aside: Exact Allelic Sharing Proportions

If the sample has $2n_{il}$ alleles at SNP l, and if r_{il} of these are the reference type, the observed sharing proportion of allele pairs (reference or non-reference) within this sample, is

$$\begin{aligned}\tilde{A}^{i}_{Wl} &= \frac{1}{2n_{il}(2n_{il}-1)}[r_{il}(r_{il}-1)+(2n_{il}-r_{il})(2n_{il}-r_{il}-1)] \\ &\approx \tilde{p}^2_{il}+(1-\tilde{p}_{il})^2\end{aligned}$$

where $\tilde{p}_{il}$ is the sample frequency for the reference 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:

$$\hat{\psi}^i_{WT} = \hat{F}^i_{WT} \;=\; \frac{(\tilde{A}^i_{Wl} - \tilde{A}_{Bl})}{(1 - \tilde{A}_{Bl})}$$

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:

$$\hat{F}_{WT} = \hat{\psi}_{WT} \;=\; \frac{\tilde{A}_{W_l} - \tilde{A}_{B_l}}{1 - \tilde{A}_{B_l}}$$

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

$$\hat{\psi}^i_{WT} = \hat{F}^i_{WT} = \frac{\sum_l (\tilde{A}^i_{Wl} - \tilde{A}_{Bl})}{\sum_l (1 - \tilde{A}_{Bl})}$$

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 almost surely to the parametric value F^i_{ST}.

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

Alternative Computing Equations for F_{WT}

For large sample sizes and r populations:

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

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:

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

Alternative Estimates for F_{WT}

The population-specific estimates are

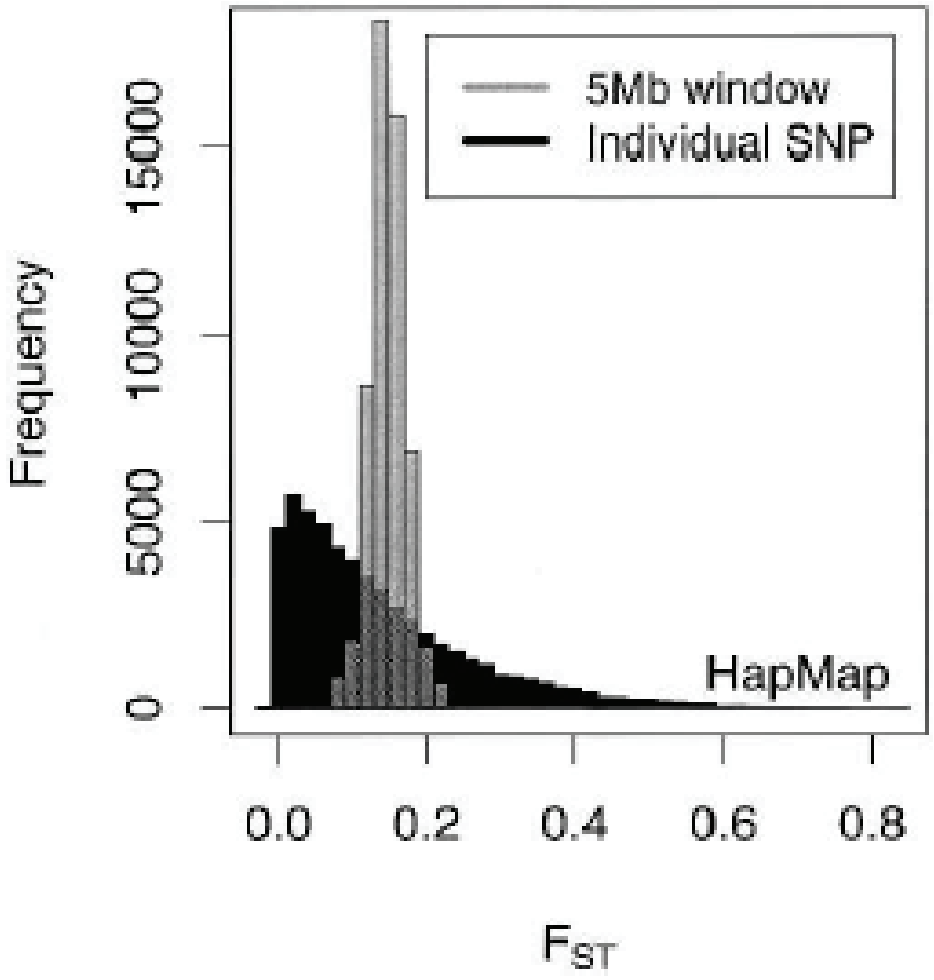
$$\widehat{F}^{i}_{WT} \;=\; 1 - \frac{\sum_l \tilde{p}_{il}(1-\tilde{p}_{il})}{\sum_l [\bar{p}_l(1-\bar{p}_l) + \frac{1}{r}s_l^2]}$$

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.

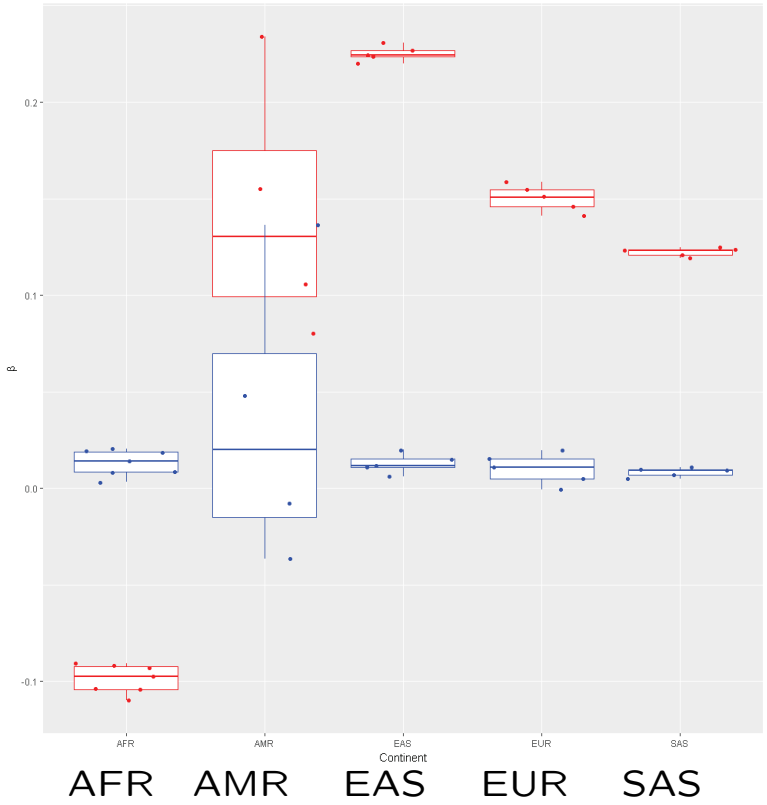
F_{WT} is relative, not absolute

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



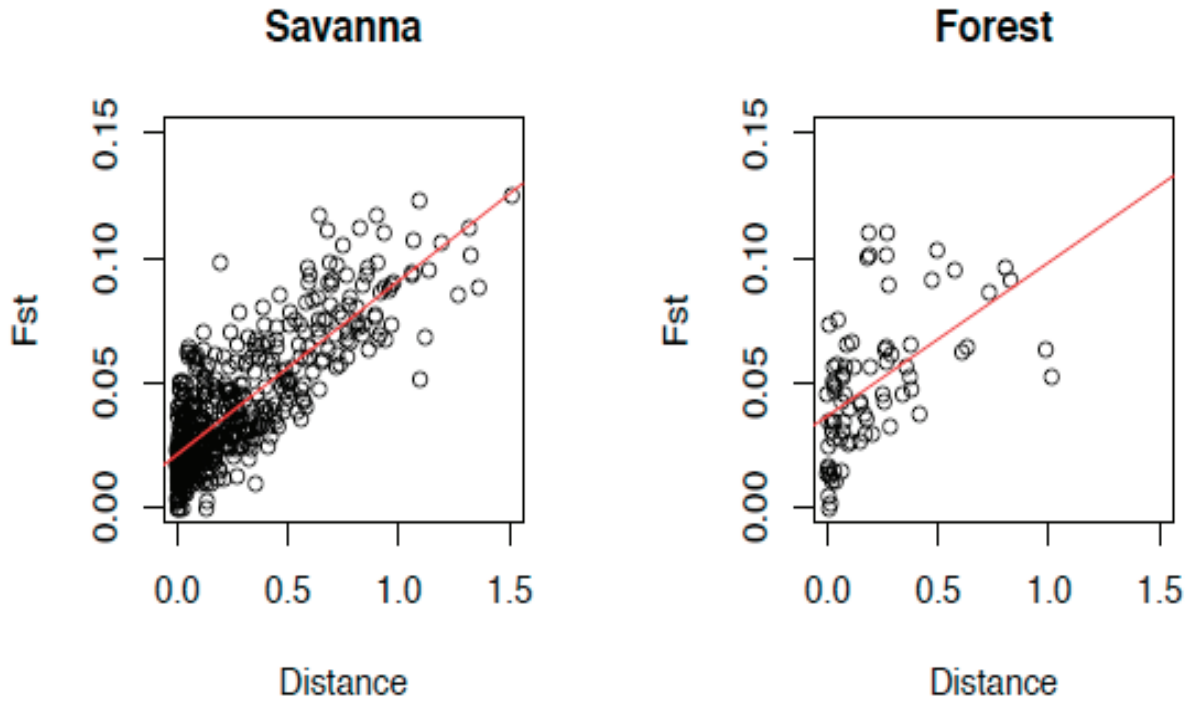
Blue box: Population relative to pairs of populations in same continent.

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

Evolutionary Inferences

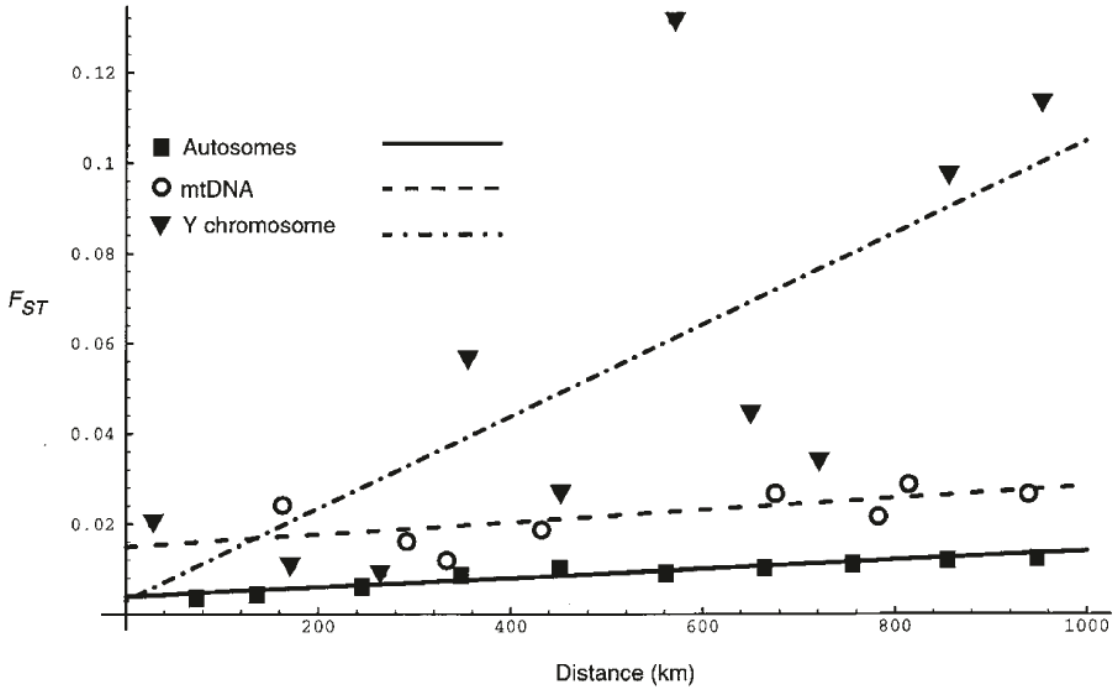
Aside: Geographic and Genetic Distances

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.205. Science 349:84-87.

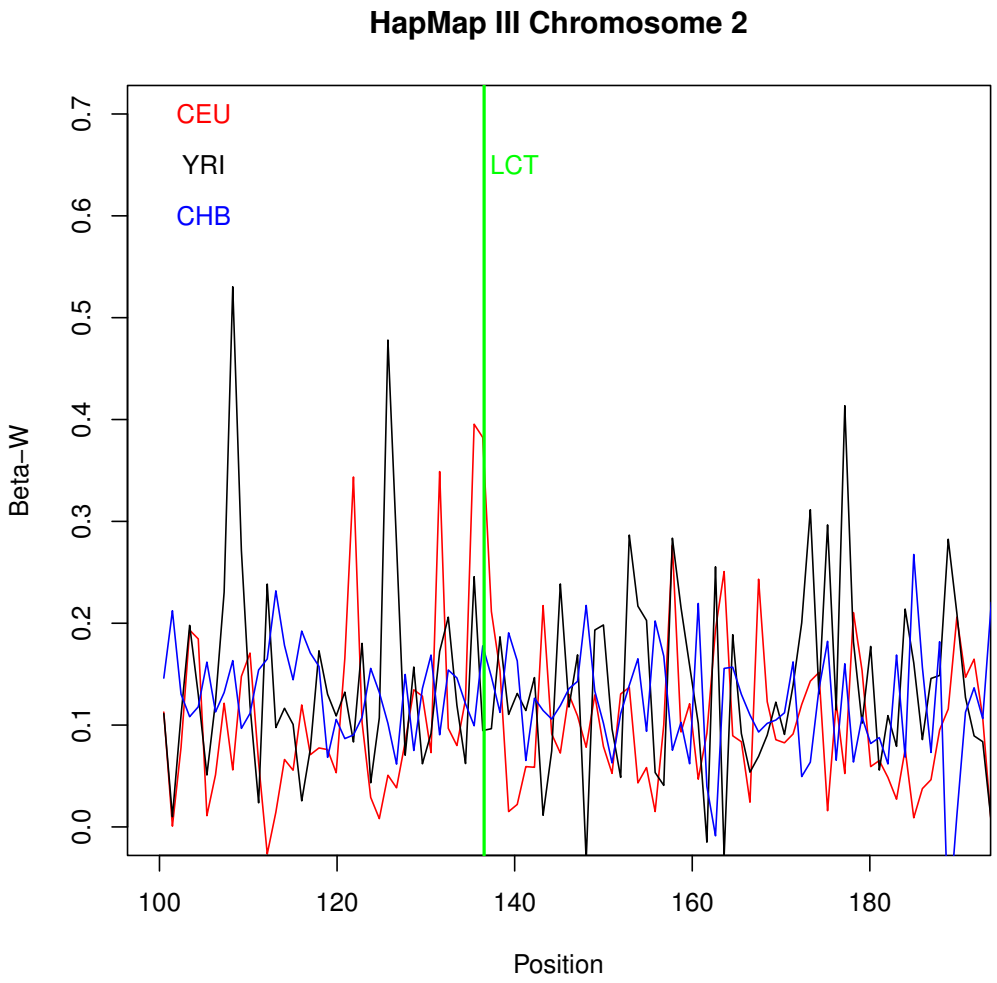
Aside: Human Migration Rates



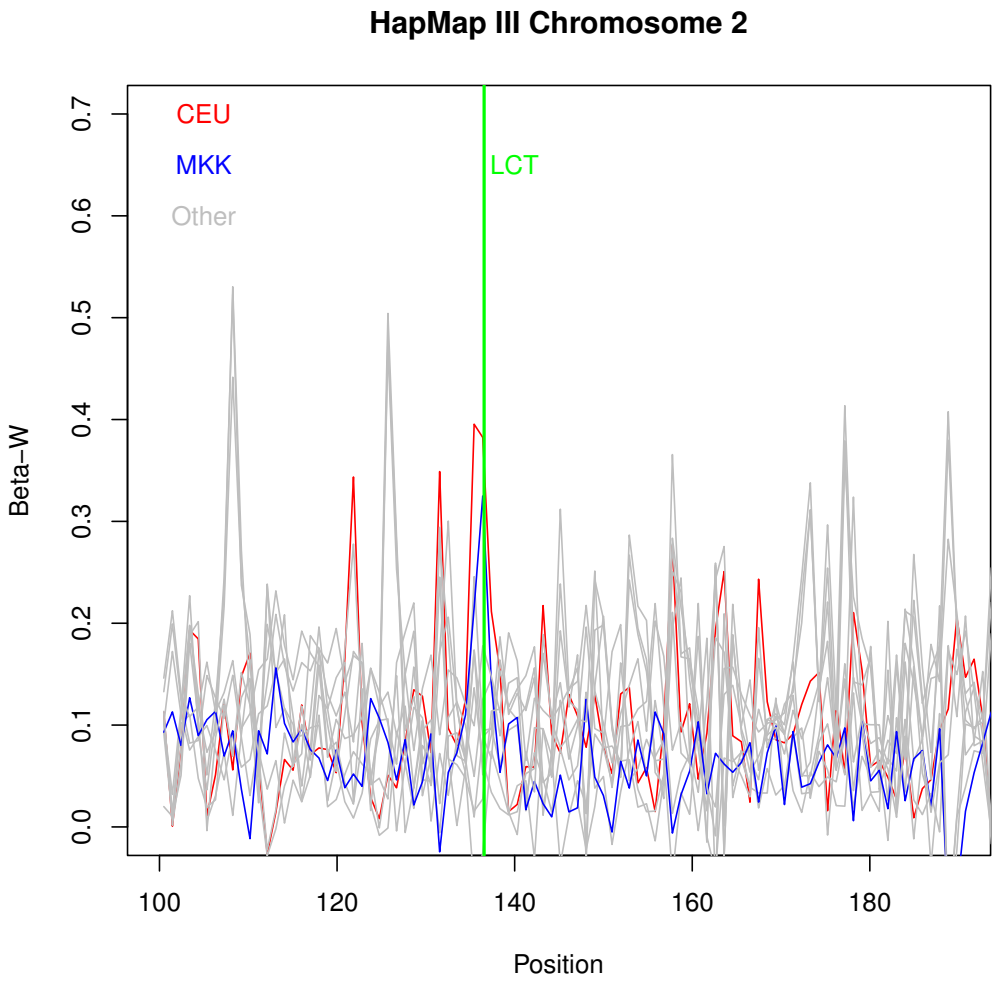
Suggests higher migration rate for human females among 14 African populations.

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

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



$\widehat{\beta}_{WT}$ 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: Weir & Cockerham 1984 Model

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

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

If $\theta = 0$, these weighted means have minimum variance.

Aside: Weir & Cockerham 1984 Model

Two mean squares were constructed for each allele:

$$\begin{aligned}
\mathrm{MSB}_l &= \frac{1}{r-1}\sum_{i=1}^{r} n_i(\tilde{p}_{il}-\bar{p}_l)^2 \\
\mathrm{MSW}_l &= \frac{1}{\sum_i (n_i-1)}\sum_i n_i\tilde{p}_{il}(1-\tilde{p}_{il})
\end{aligned}$$

These have expected values

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

where $n_c = (\sum_i n_i - \sum_i n_i^2/\sum_i n_i)/(r-1)$. The Weir & Cockerham *weighted* allele-based estimator of θ (or F_{WT}) is

$$\hat{\theta}_{WC} = \frac{\sum_l(\mathrm{MSB}_l - \mathrm{MSW}_l)}{\mathrm{MSB}_l + (n_c-1)\mathrm{MSW}_l}$$

Aside: Weir & Cockerham 1984 Estimator

Under the allele sharing model, the Weir and Cockerham estimator has expectation

$$\mathcal{E}(\hat{\theta}_{\mathrm{WC}}) = \frac{\theta_{Wc} - \theta_{Bc} + Q}{1 - \theta_{Bc} + Q} \quad \text{instead of} \quad \frac{\theta_W - \theta_B}{1 - \theta_B}$$

where

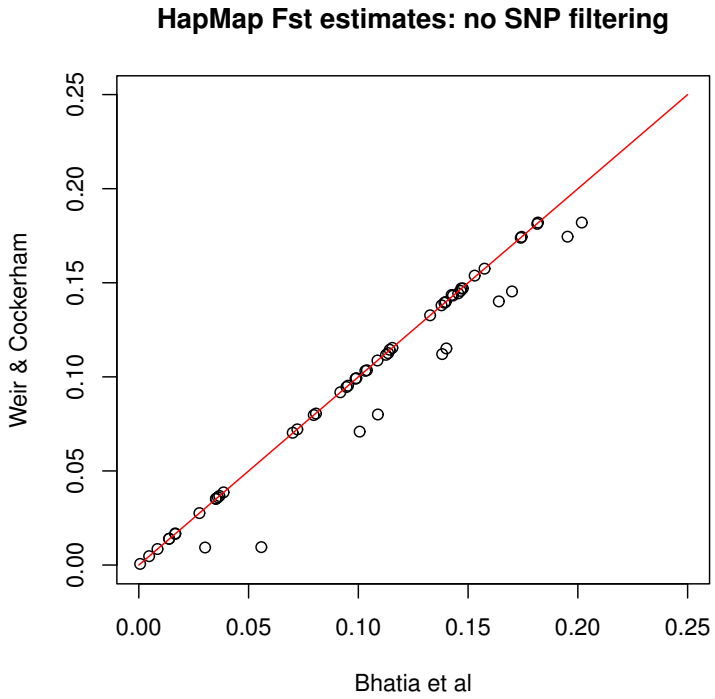
$$\theta_{Wc} = \frac{\sum_i n_i^c \theta^i}{\sum_i n_i^c} \quad , \quad \theta_{Bc} = \frac{\sum_{i \neq i'} n_i n_{i'} \theta^{ii'}}{\sum_{i \neq i'} n_i n_{i'}}$$

$$n_i^c = n_i - \frac{n_i^2}{\sum_i n_i} \quad , \quad n_c = \frac{1}{r-1} \sum_i n_i^c$$

$$Q = \frac{1}{(r-1)n_c} \sum_i \left(\frac{n_i}{\bar{n}} - 1 \right) \theta^i$$

If the Weir and Cockerham model holds ($\theta^i = \theta$), or if $n_i = n$, or if n_c is large, then $Q = 0$.

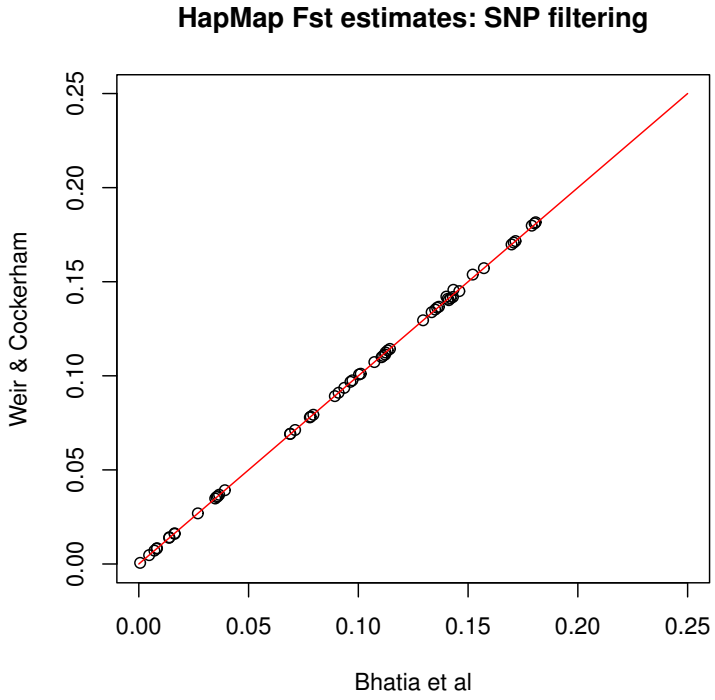
Aside: WC84 vs Beta Allele-based Estimators



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

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

Aside: WC vs Unweighted Estimator



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.