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 $\tilde{M}^i_{Wl}$ 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

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

Aside: Exact Allelic Matching Proportions

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

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

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

$$\begin{aligned}\tilde{M}^{ii'}_{Bl} &= \frac{1}{4n_i n_{i'}}\sum_{\substack{j=1\\ j\neq j'}}^{2n_i}\sum_{j'=1}^{2n_{i'}} x_{ju}x_{j'u} \\ &= \tilde{p}_{il}\tilde{p}_{i'l}+(1-\tilde{p}_{il})(1-\tilde{p}_{i'l})\end{aligned}$$

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

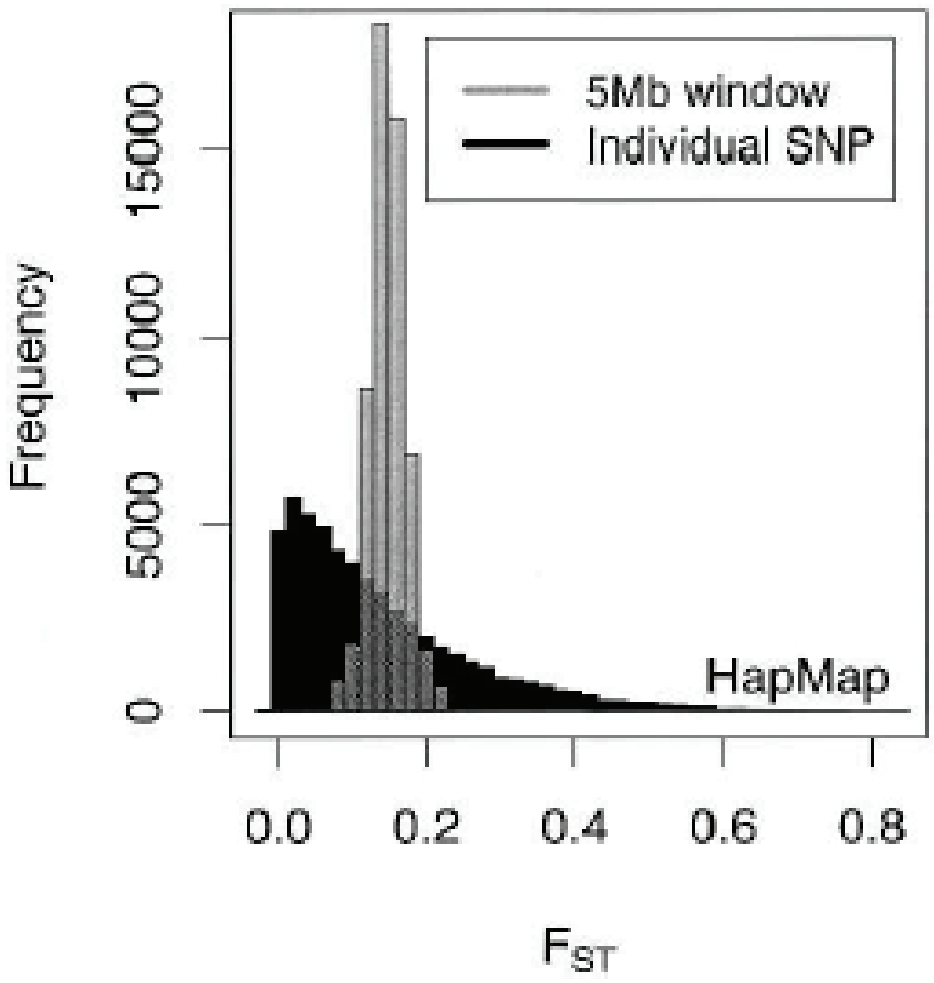
$$\hat{F}^i_{WT} = 1 - \frac{\sum_l \tilde{p}_{il}(1-\tilde{p}_{il})}{\sum_l [\bar{p}_l(1-\bar{p}_l) + \frac{1}{r}s_l^2]}$$

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 $r = 2$.

Effect of Number of Loci



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

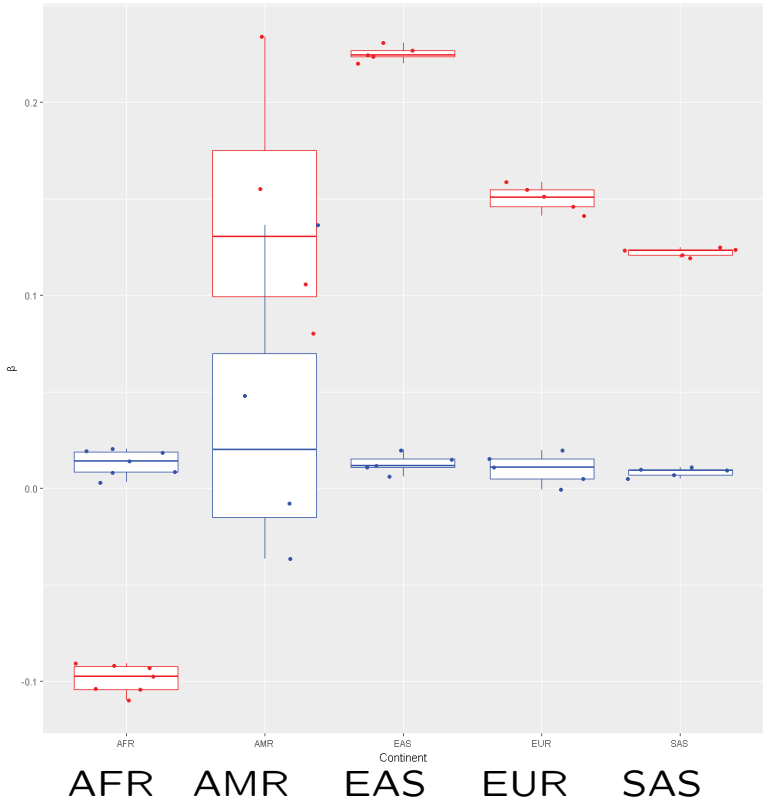
F_{WT} is relative, not absolute

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



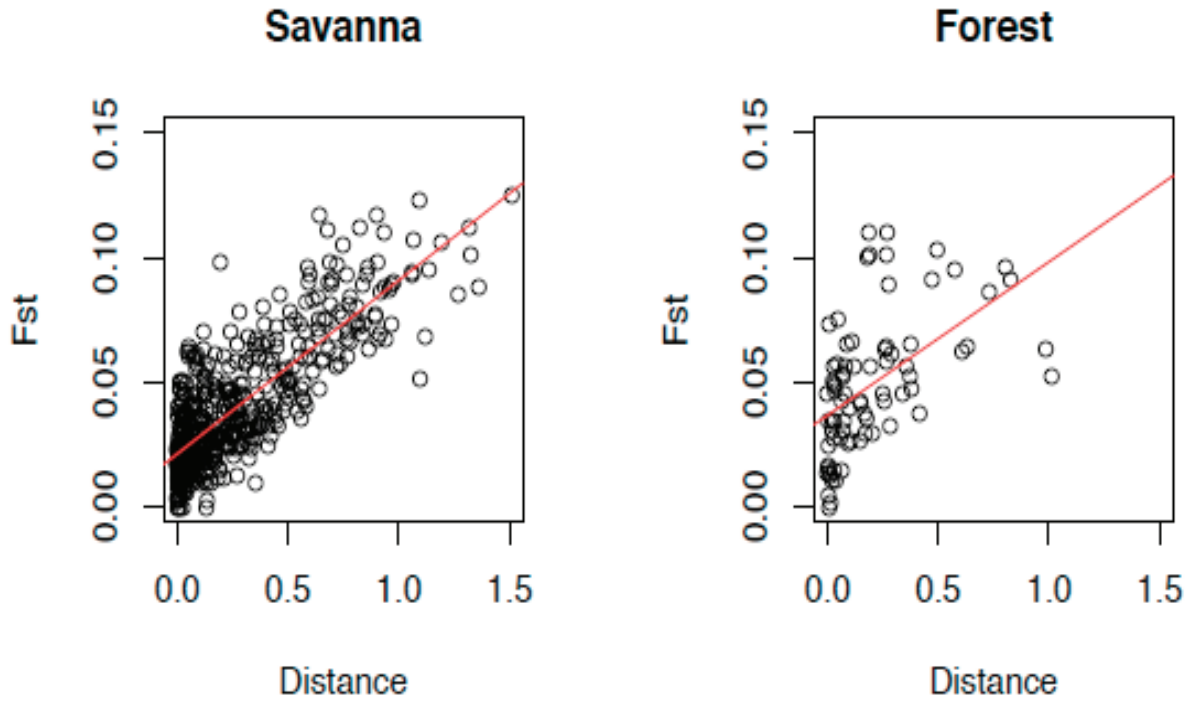
Blue box: Population relative to pairs of populations in same continent.

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

Evolutionary Inferences

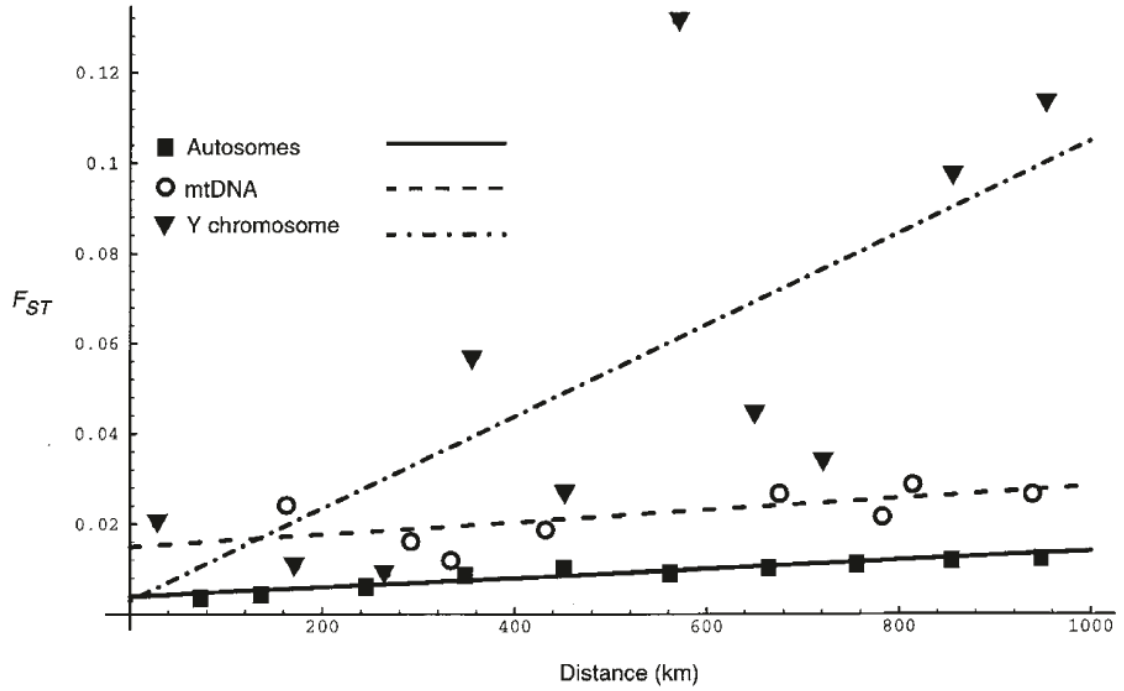
Aside: Geographic and Genetic Distances

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.

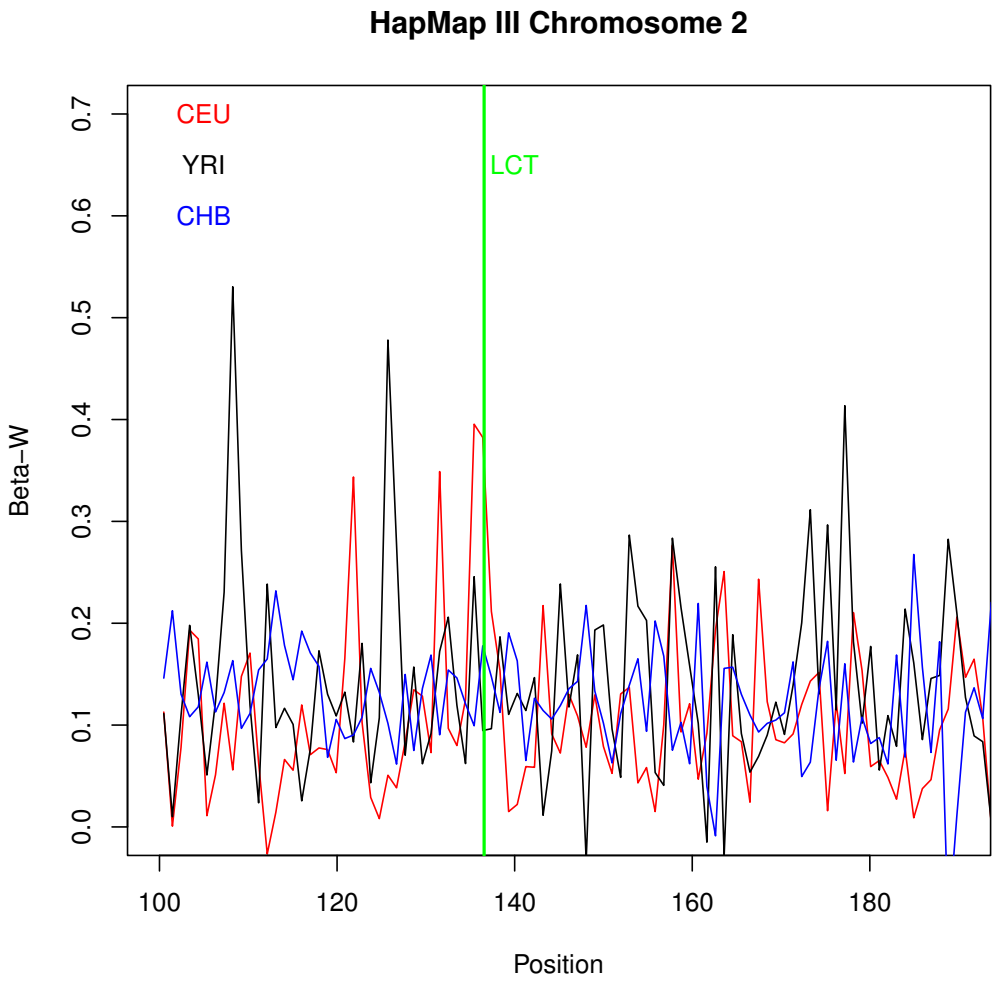
Aside: Human Migration Rates



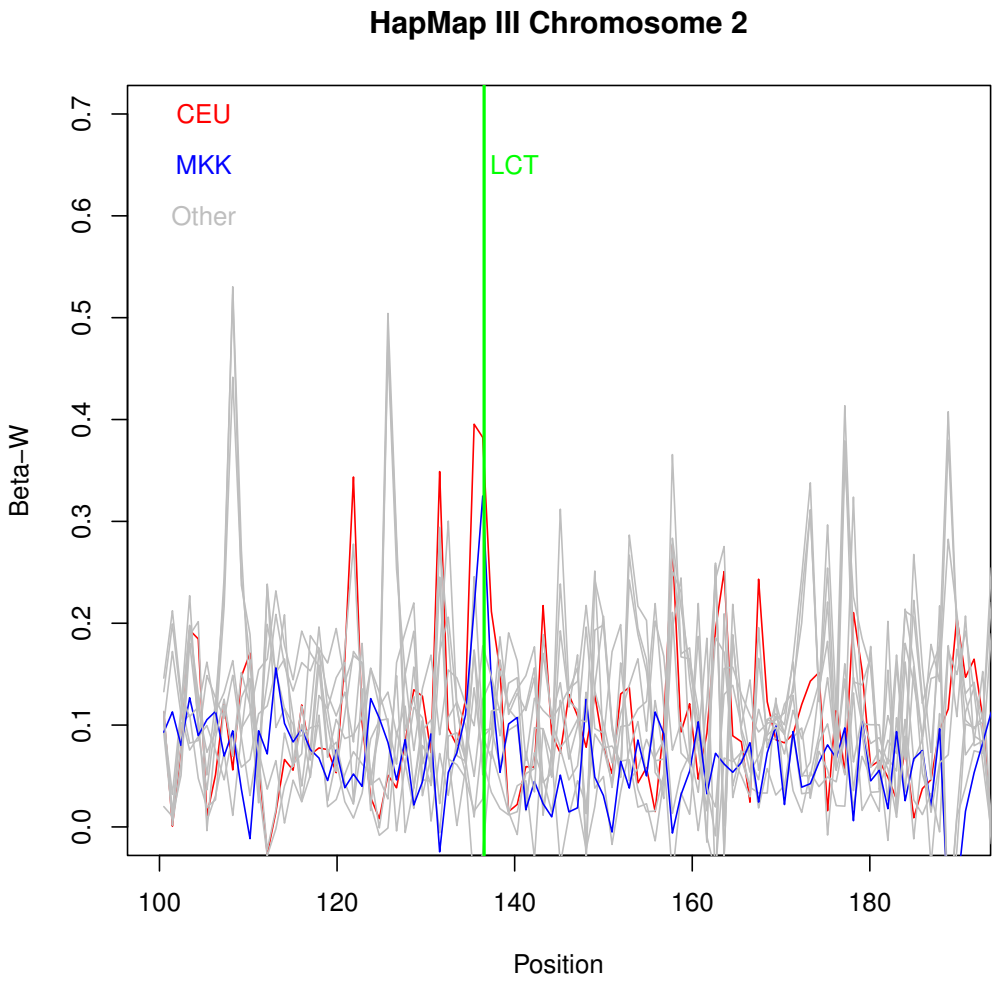
Suggests higher migration rate for human females among 14 African populations.

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 β approach described here, 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

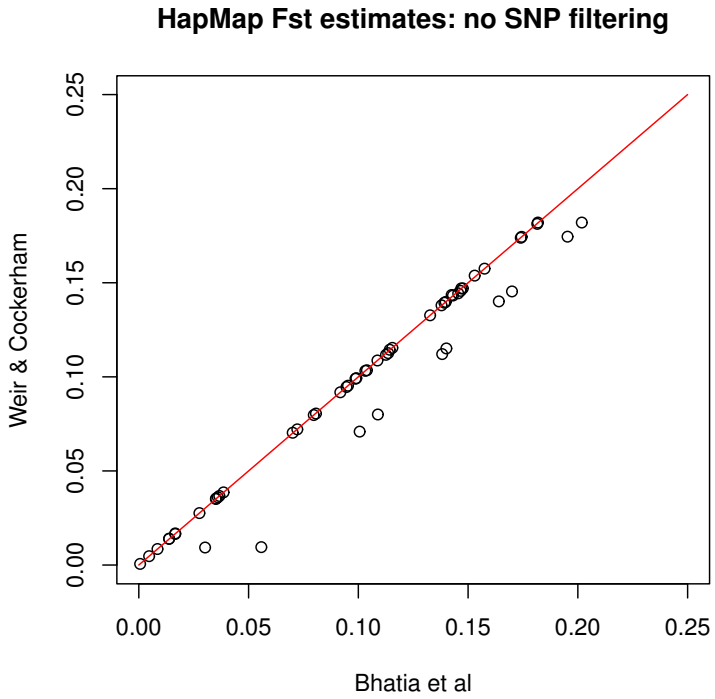
$$\theta_{Wc} = \frac{\sum_i n_i^c \theta^i}{\sum_i n_i^c} \quad , \quad \theta_{Bc} = \frac{\sum_{i \neq i'} n_i n_{i'} \theta^{ii'}}{\sum_{i \neq i'} n_i n_{i'}}$$

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

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

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

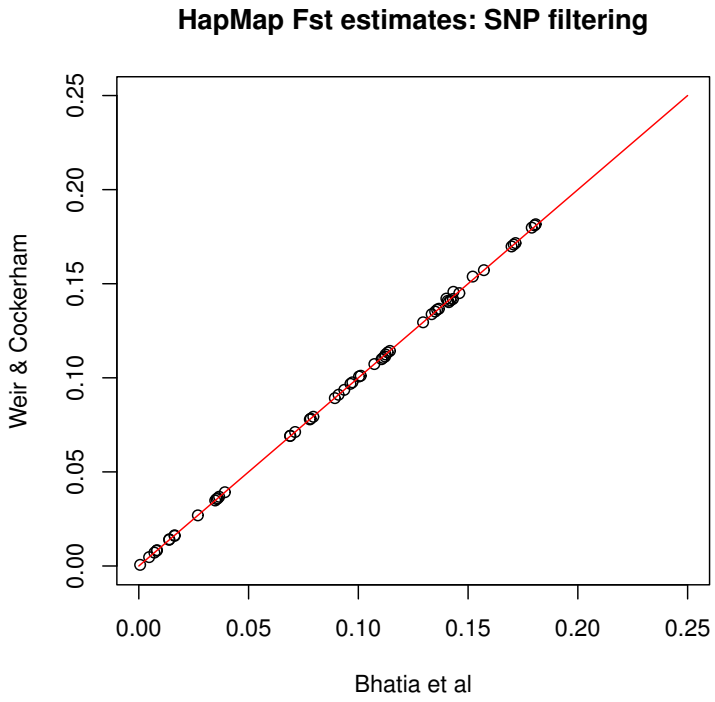
Aside: WC84 vs Beta Allele-based Estimators



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.