Allelic matching within and between populations

Jérôme Goudet and Bruce Weir

Matching Proportions

If a sample of n_i alleles from population i have n_{iu} copies of allele u, the within-population sample matching proportion is

$$\tilde{M}_{Wi} = \frac{1}{n_i(n_i - 1)} \sum_u n_{iu}(n_{iu} - 1)$$

The sample matching proportion between populations i and i' is

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

For a set of r populations

$$\tilde{M}_W = \frac{1}{r} \sum_i \tilde{M}_{Wi} \quad , \quad \tilde{M}_B = \frac{1}{r(r-1)} \sum_{i \neq i'} \tilde{M}_{Bii'}$$

Allelic Indicators

A useful approach is to attach indicator variables to each allele. For the jth allele sampled from the ith population:

$$x_{ij} = \begin{cases} 1 & \text{allele of type } u \\ 0 & \text{otherwise} \end{cases}$$

Weir and Hill (2002) described a model that specifies expectations for these indicators:

$$\begin{aligned} \mathcal{E}(x_{ij}) &= p_u \\ \mathcal{E}(x_{ij}x_{ij'}) &= \theta_i p_u + (1-\theta_i)p_u^2, \;\; j \neq j' \\ \mathcal{E}(x_{ij}x_{i'j'}) &= \theta_{ii'} p_u + (1-\theta_{ii'})p_u^2, \;\; i \neq i' \end{aligned}$$

Expectation is over samples from each population and over replicates of each population.

Sample Allele Frequencies

The sample proportions of each profile type can be expressed as averages of indicator variables

$$\tilde{p}_{iu} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$$

and this leads to variances and covariances:

$$\begin{aligned}
\text{Var}(\tilde{p}_{iu}) &= p_u(1-p_u)\left(\theta_i + \frac{1-\theta_i}{n_i}\right) \\
\text{Cov}(\tilde{p}_{iu}, \tilde{p}_{iu'}) &= -p_u p_{u'}\left(\theta_i + \frac{1-\theta_i}{n_i}\right) \\
\text{Cov}(\tilde{p}_{iu}, \tilde{p}_{i'u}) &= p_u(1-p_u)\theta_{ii'} \\
\text{Cov}(\tilde{p}_{iu}, \tilde{p}_{i'u'}) &= -p_u p_{u'}\theta_{ii'}
\end{aligned}$$

Coancestries

Corresponding to ibs measures M there are ibd probabilities θ.

Taking expectations over samples from populations and over evolutionary replicates of populations:

$$\mathcal{E}(\tilde{M}_{Wi}) = 1 - H(1-\theta_i) \quad , \quad \mathcal{E}(\tilde{M}_W) = 1 - H(1-\theta_W)$$
$$\mathcal{E}(\tilde{M}_{Bii'}) = 1 - H(1-\theta_{ii'}) \quad , \quad \mathcal{E}(\tilde{M}_B) = 1 - H(1-\theta_B)$$

where $H = 1 - \sum_u p_u^2$. These suggest moment estimates

$$\hat{\beta}_{Wi} = \frac{\tilde{M}_i - \tilde{M}_B}{1 - \tilde{M}_B} \quad , \quad \mathcal{E}(\hat{\beta}_{Wi}) = \frac{\theta_i - \theta_B}{1 - \theta_B}$$
$$\hat{\beta}_W = \frac{\tilde{M}_W - \tilde{M}_B}{1 - \tilde{M}_B} \quad , \quad \mathcal{E}(\hat{\beta}_W) = \frac{\theta_W - \theta_B}{1 - \theta_B}$$

The usual F_{ST} is the same as $\beta_W = (\theta_W - \theta_B)/(1 - \theta_B)$.

WC84 Estimator

Under this (Weir and Hill) model, the WC84 estimator has expectation

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

where

$$\theta_W^c = \frac{\sum_i n_i^c \theta_i}{\sum_i n_i^c} \quad , \quad \theta_B^c = \frac{\sum_{i \neq i'} n_i n_{i'} \theta_{ii'}}{\sum_{i \neq i'} n_i n_{i'}}$$

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

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

If the WC84 model holds ($\theta_i = \theta$), or if $n_i = n$, or if n_c is large, then $Q = 0$ and $\mathcal{E}(\hat{\theta}_{WC}) = (\theta_W - \theta_B)/(1 - \theta_B)$.

Continent-Island Model

A continent of infinite population size sends migrant alleles to islands with finite population sizes N_i at rate m. Alleles mutate to a new state at rate μ.

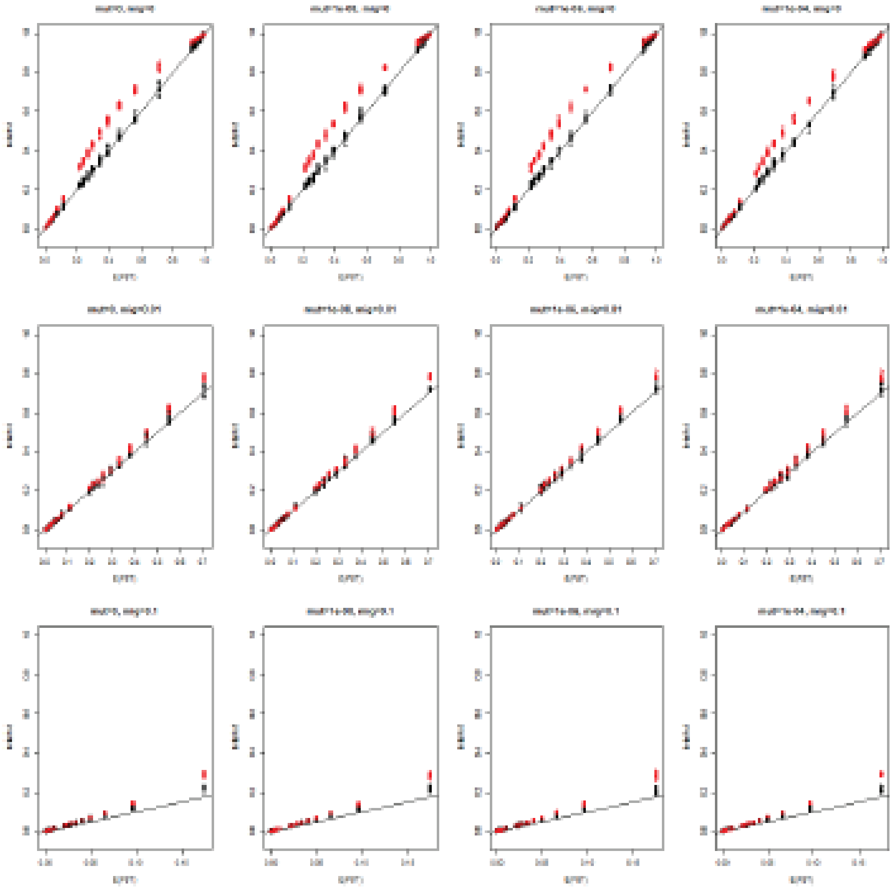
Within island i, at equilibrium,

$$\theta_{Wi} = \frac{(1-m)^2(1-\mu)^2}{2N_i - (2N_i - 1)(1-m)^2(1-\mu)^2}$$

and the values of $\theta_{Bii'}$ are all zero.

Simulations show that $\widehat{\beta}_{Wi}$ values are close to θ_{Wi} values but BayeScan estimates are too high.

Island Model Estimates



HGDP Microsatellite Data

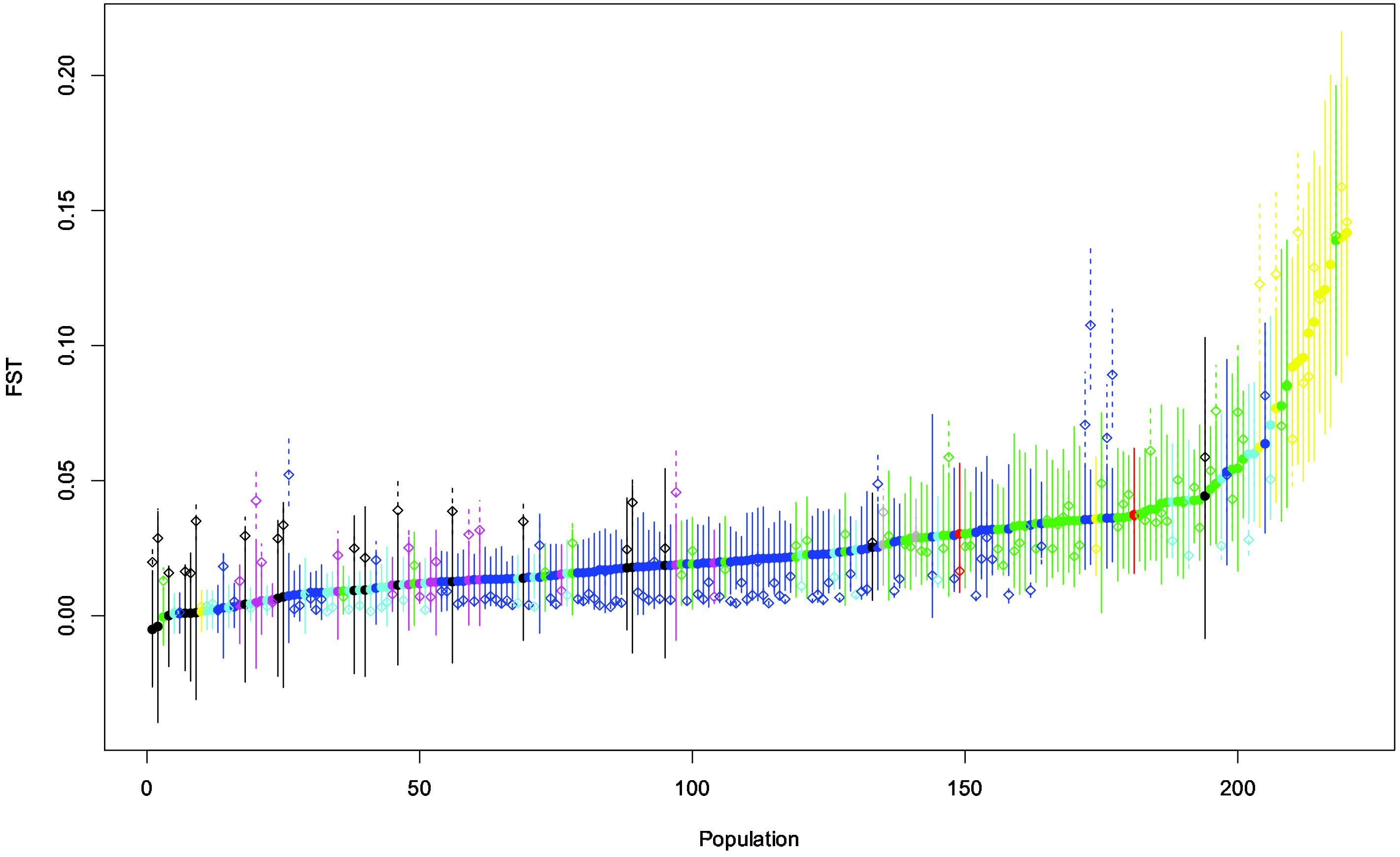
Pemberton, DeGiorgi and Rosenberg, 2013, G3 3:891-907.

Estimated F_{ST} as moment estimate β_{Wi}, with bootstrap over loci confidence intervals. (Solid dots).

Also by BayeScan with credible intervals. (empty diamonds)

Estimates colored by continent.

Human Genome Diversity STR Panel



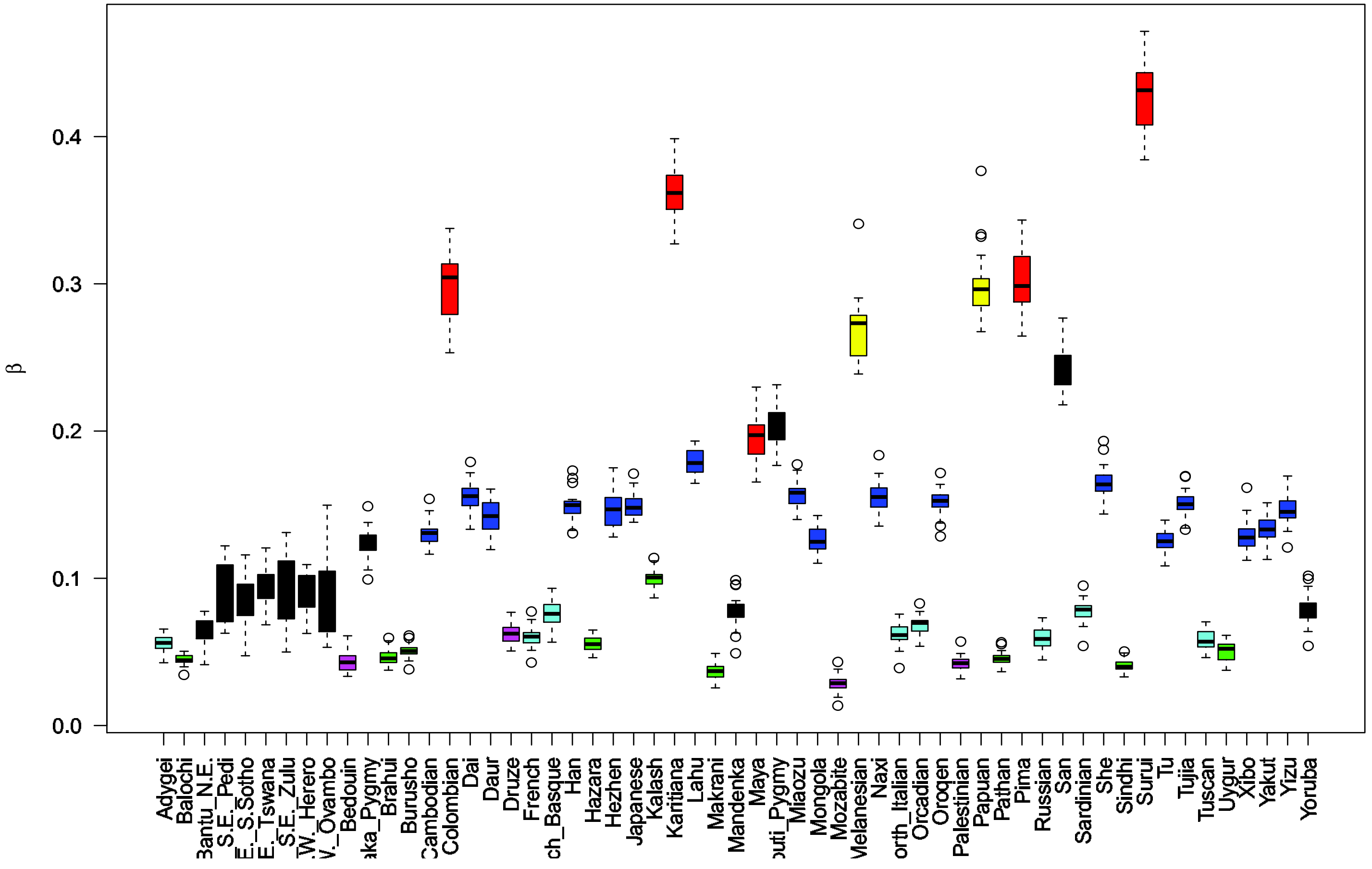
HGDP SNP Data

Li et al., 2008, Science 319:1100-1104.

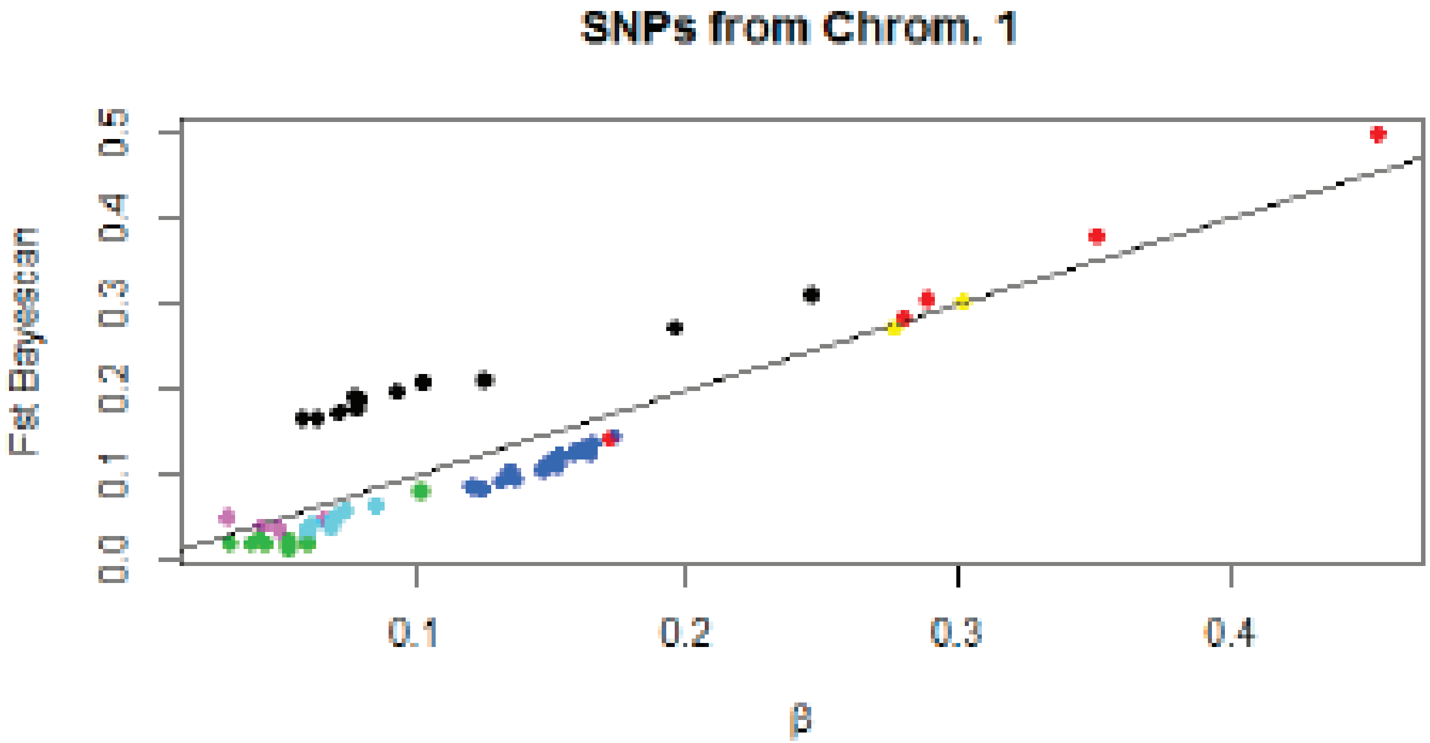
Moment estimates β_{Wi} for each population and for each chromosome. Box plots show variation among chromosomes.

African estimates larger than those for European. Suggests that SNPs were chosen to favor high MAF in Arica.

HGDP SNP Panel



Moment vs Bayes: HGDP Chr 1 SNPs

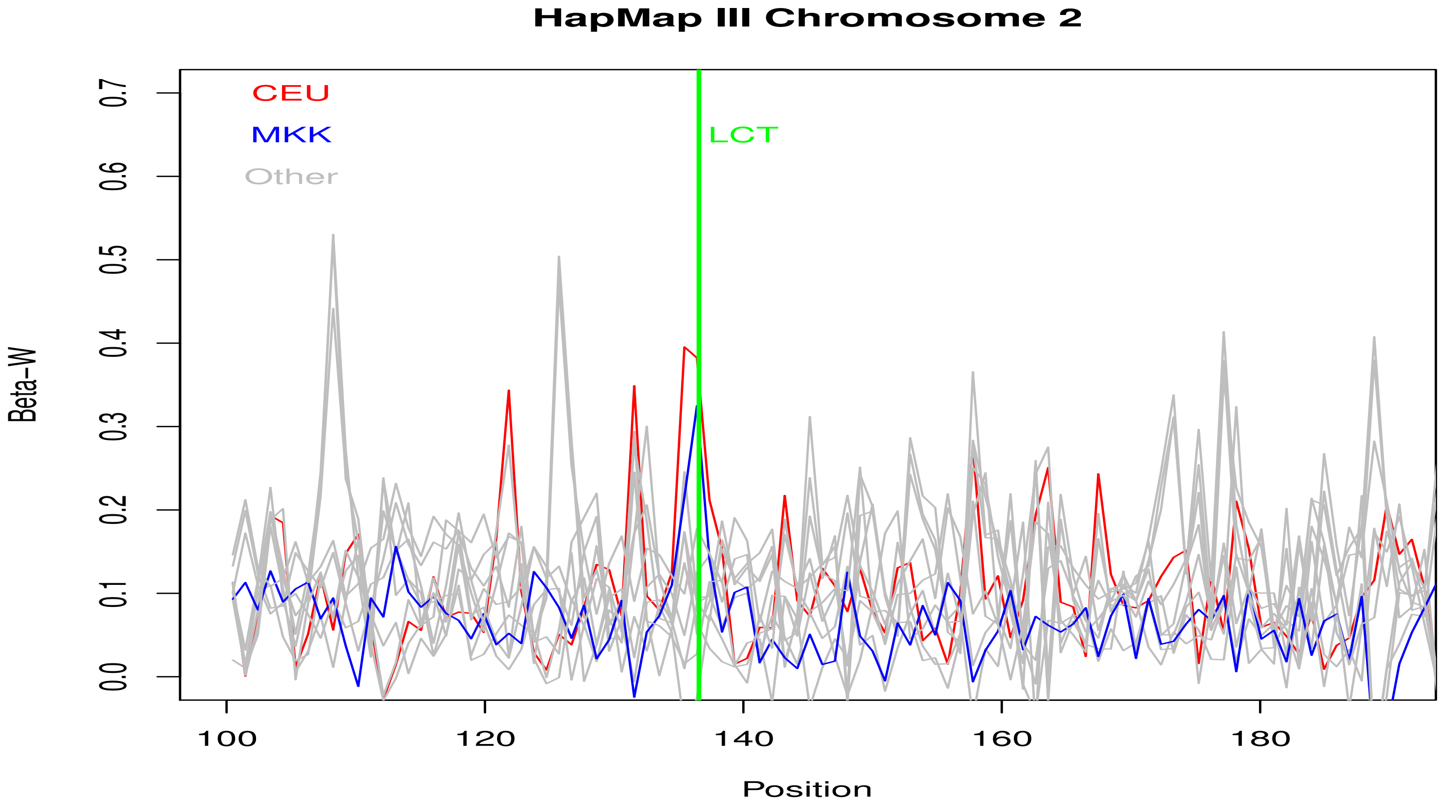


HapMap3 SNP Data

Altshuler et al. 2010.

Moment estimates of β_{Wi} for SNP windows on chromosome 2. LCT gene location noted by dotted line, showing peaks for CEU and MKK populations.

HapMap3 Chr 2 SNPs



1000 Genomes SNP Data

Altshuler et al. 2012.

Native American data not used (small sample size).

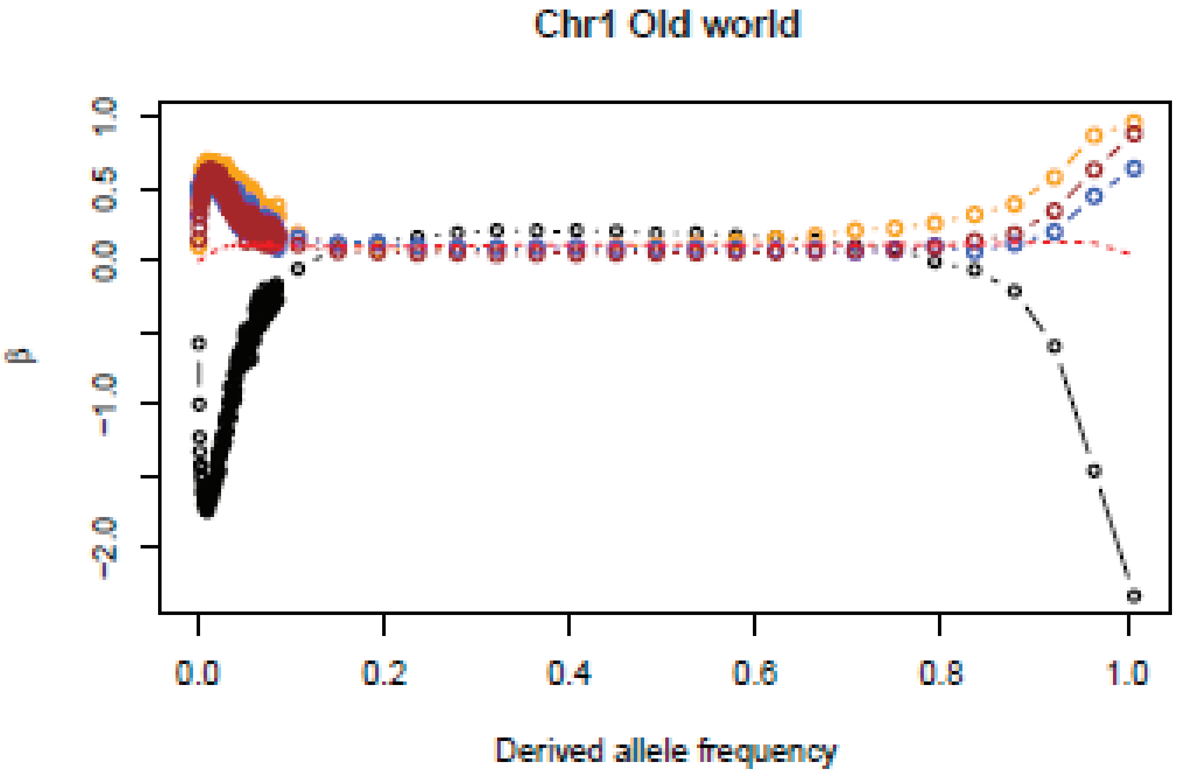
Estimates of β_{Wi} for each continent:

	Africa	EastAsia	Europe	SouthAsia	Overall
PlotColor	Black	Orange	Blue	Brown	Red
SampleSize	410	64	694	592	
Chr 1	-0.117	0.215	0.171	0.140	0.102
Chr 2	-0.124	0.223	0.165	0.146	0.102
Chr 6	-0.090	0.193	0.141	0.125	0.092
Chr 13	-0.113	0.210	0.153	0.134	0.096

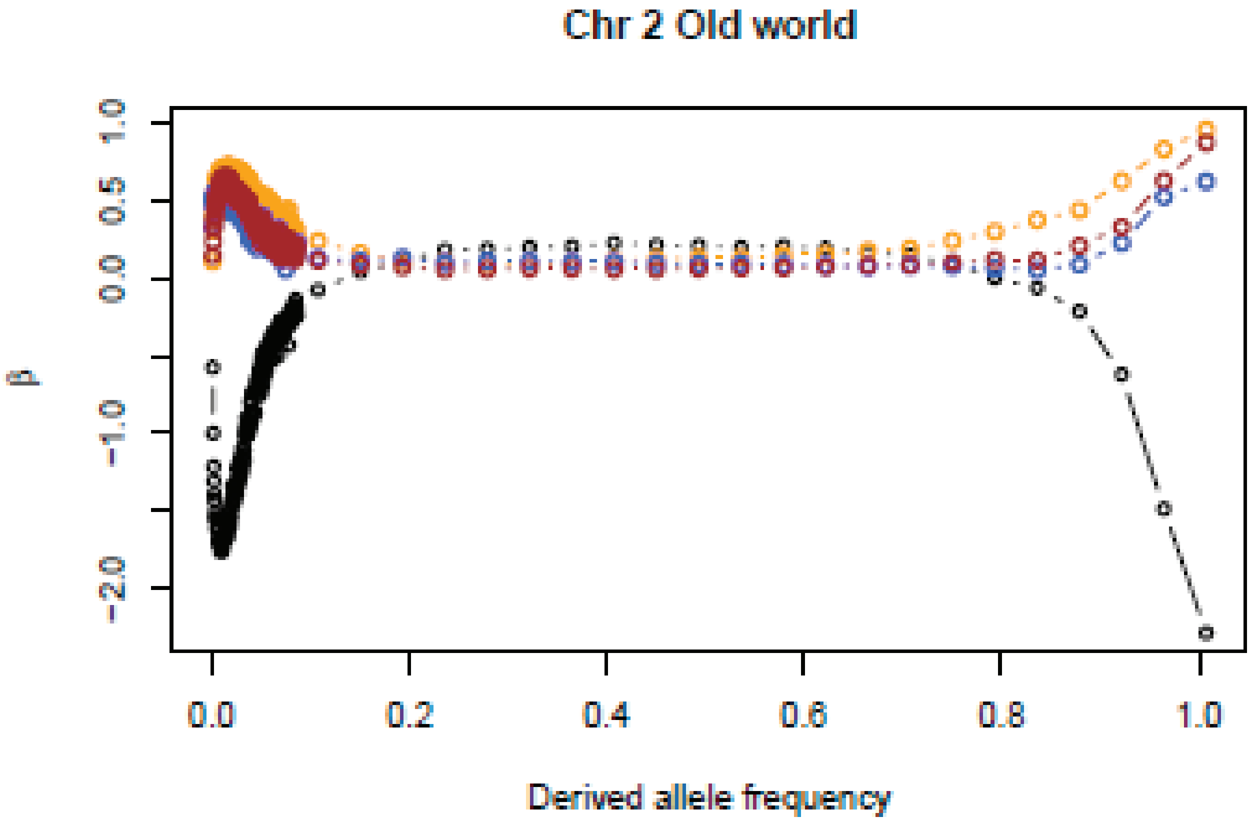
Estimate Details

Next slides plot estimates of β_{Wi} against derived frequency for each allele.

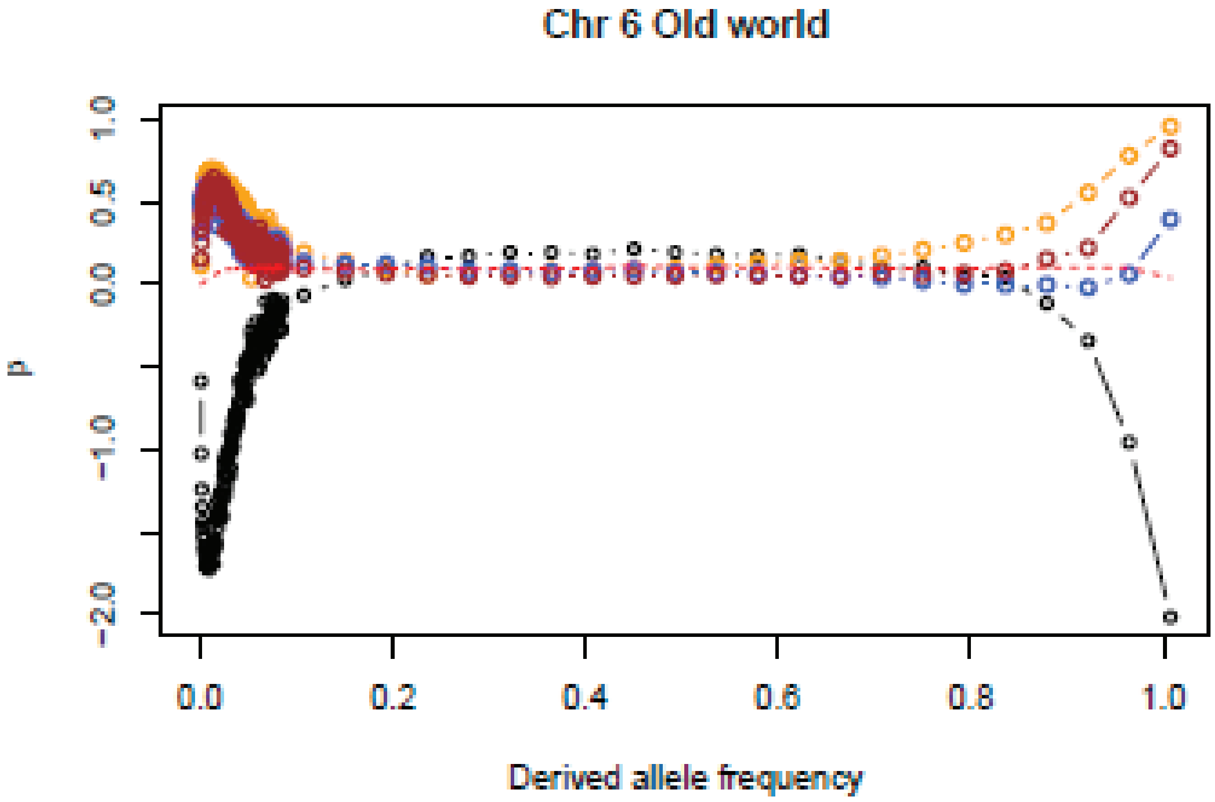
Old World 1000Genomes Chr 1



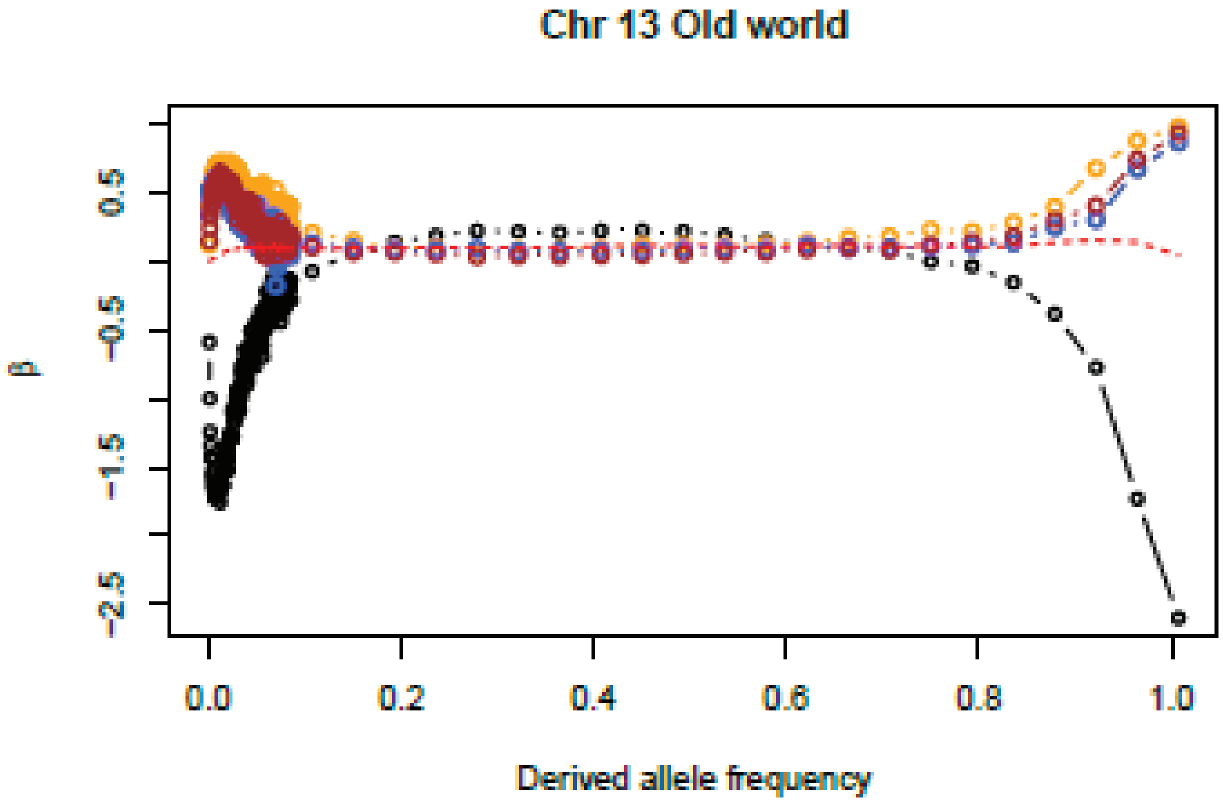
Old World 1000Genomes Chr 2



Old World 1000Genomes Chr 6



Old World 1000Genomes Chr 13



Private Alleles

Suppose population i has MAF x at some SNP, but no other population is polymorphic at that position. The population matching proportions for this SNP are

$$\begin{aligned}
M_{Wi} &= x^2 + (1-x)^2 \\
M_{Wi'} &= 1, i' \neq i \\
M_{Bii'} &= (1-x), i \neq 1 \\
M_{Bi'i''} &= 1, i'' \neq i' \neq i \\
M_W &= 1 - \frac{2x(1-x)}{r} \\
M_B &= 1 - \frac{2x}{r}
\end{aligned}$$

and the F_{ST} values are

$$\begin{aligned}
\beta_{Wi} &= 1 - r(1-x) \\
\beta_{Wi'} &= 1, i' \neq i \\
\beta_W &= x
\end{aligned}$$

Private Alleles

For a population with a private SNP allele, at that locus $\beta_i < 0$ for $x < (r-1)/r$.

A population with many SNPs with private alleles at low to intermediate frequencies will likely have a negative estimate of β_{Wi}. How negative will depend on how many populations are sampled.

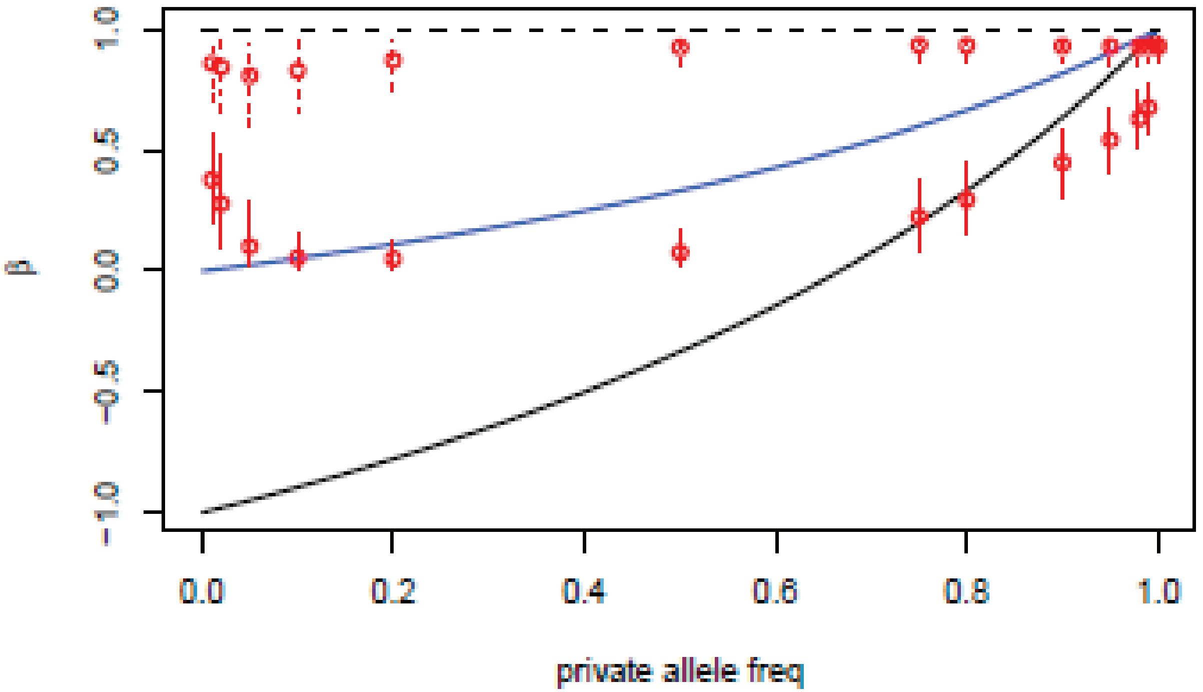
Estimation procedures must allow population-specific β_{Wi} values to be negative - not a feature of current Bayesian methods that, in essence, regard populations as independent $\beta_B = 0$.

Two Populations

The next slide shows estimates from two populations, one of which has a private allele.

The BayeScan means and 95% credible intervals are in red.

Private Alleles: Two Populations



Human Evolution

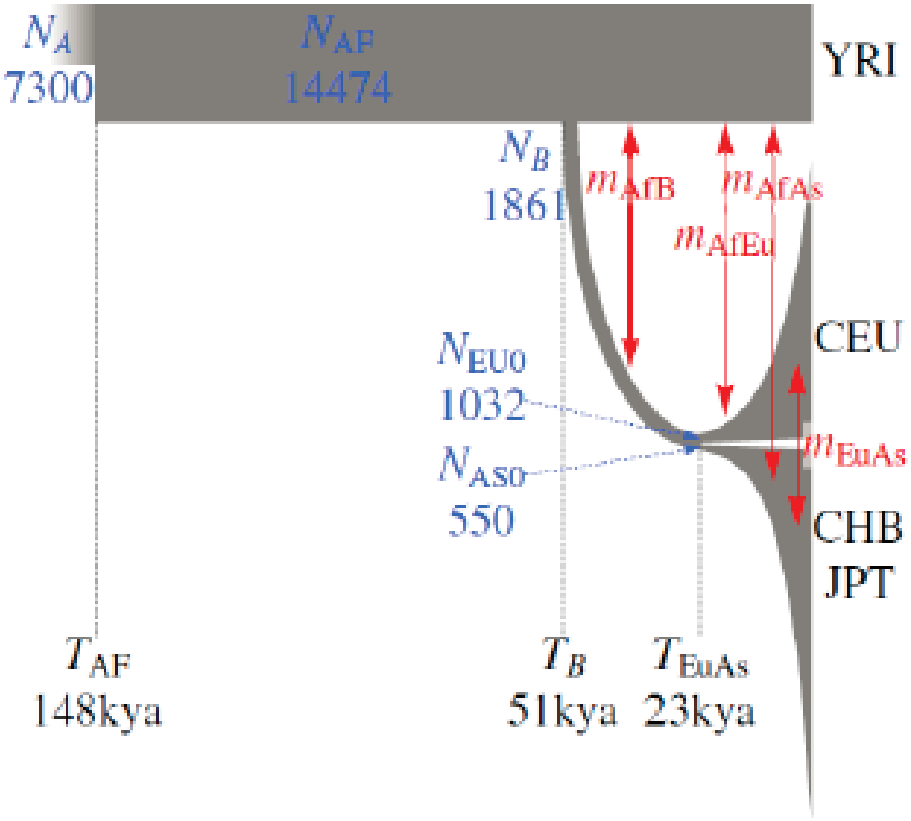
Several published scenarios for world colonization by *Homo sapiens*:

Gutenkunst et al. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLoS Genetics 5: e1000695.

Gravel et al. 2011. Demographic history and rare allele sharing among human populations. PNAS 108:1193-11988.

Gravel et al. gave the scenario on next slide.

Gravel et al.



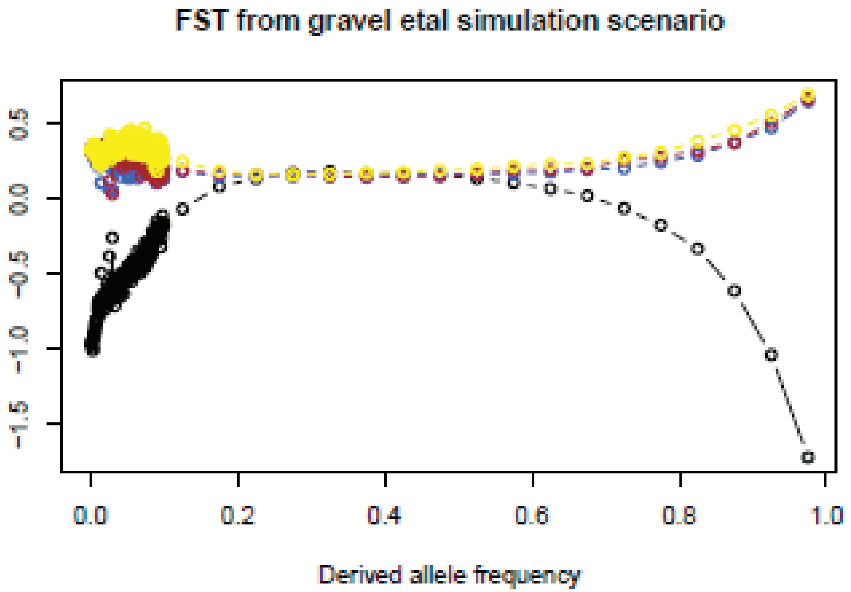
N_A
7300
N_{AF}
14474
YRI
N_B
1861
m_{AfB}
m_{AfAs}
m_{AfEu}
CEU
N_{EU0}
1032
N_{AS0}
550
m_{EuAs}
CHB
JPT
T_{AF}
148kya
T_B
51kya
T_{EuAs}
23kya

Gravel et al. Parameters

Simulated data with software of Liang, Zollner and Abecasis, 2007, Bioinformatics 23:1565-1567.

Set parameters to correspond to Gravel et al.

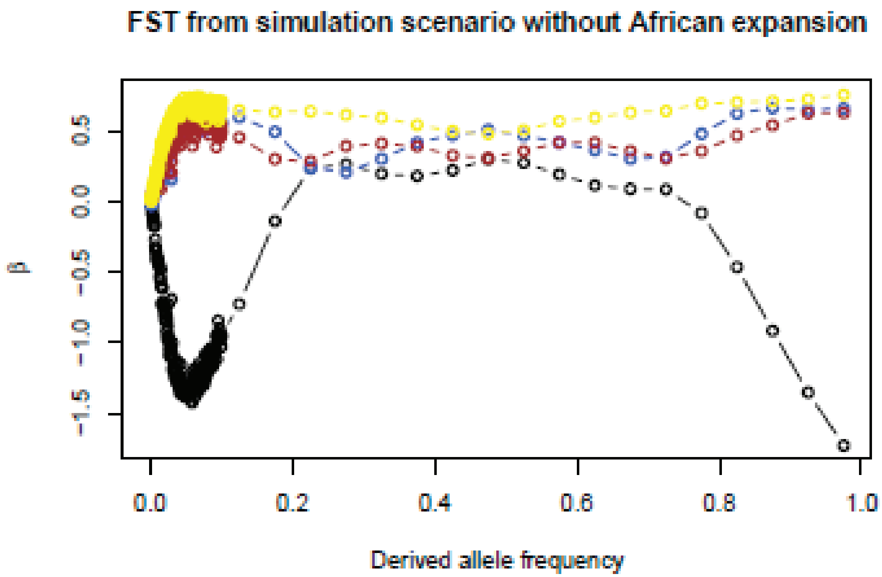
Gravel et al. Parameters



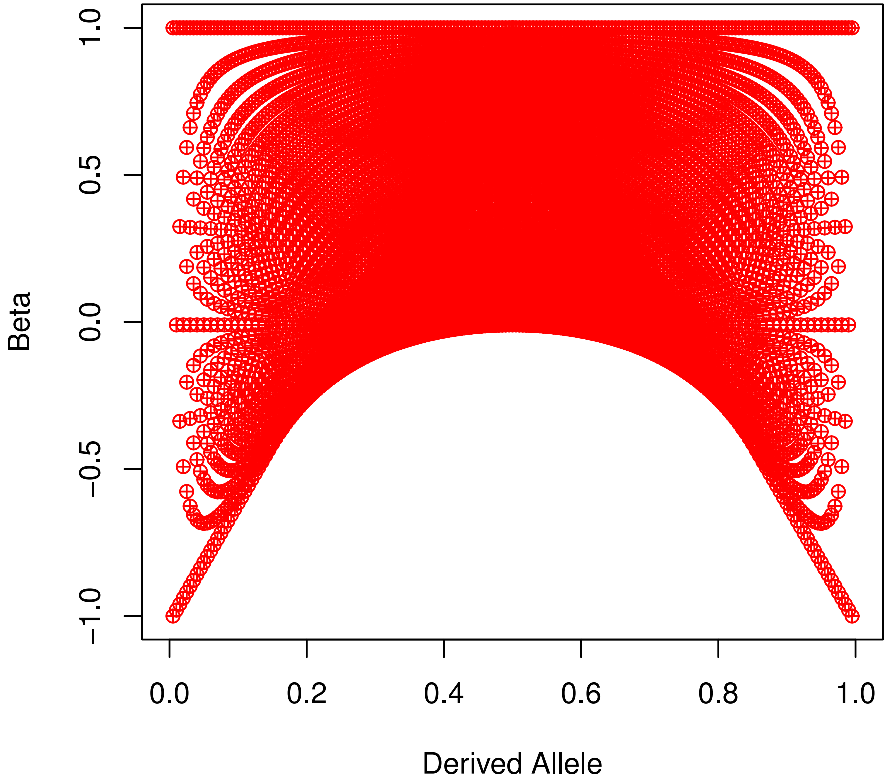
Modified Parameters

Then simulated data without the initial bottleneck followed by expansion in Africa.

Modified Parameters

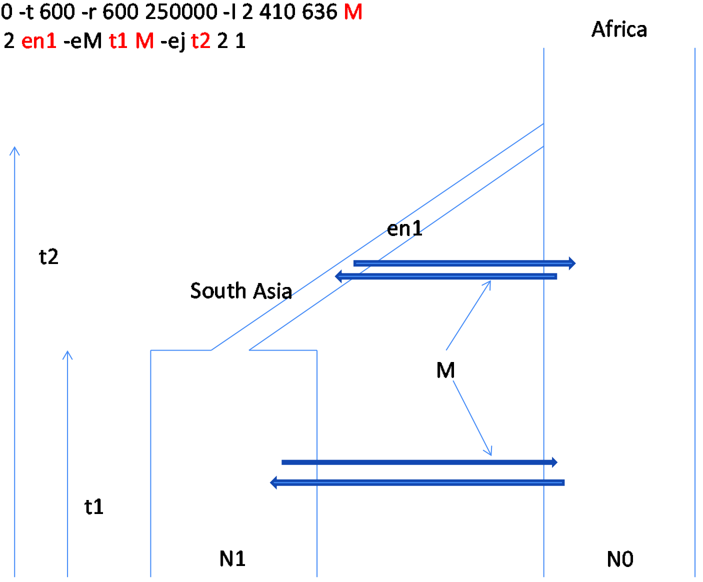


Data viewpoint

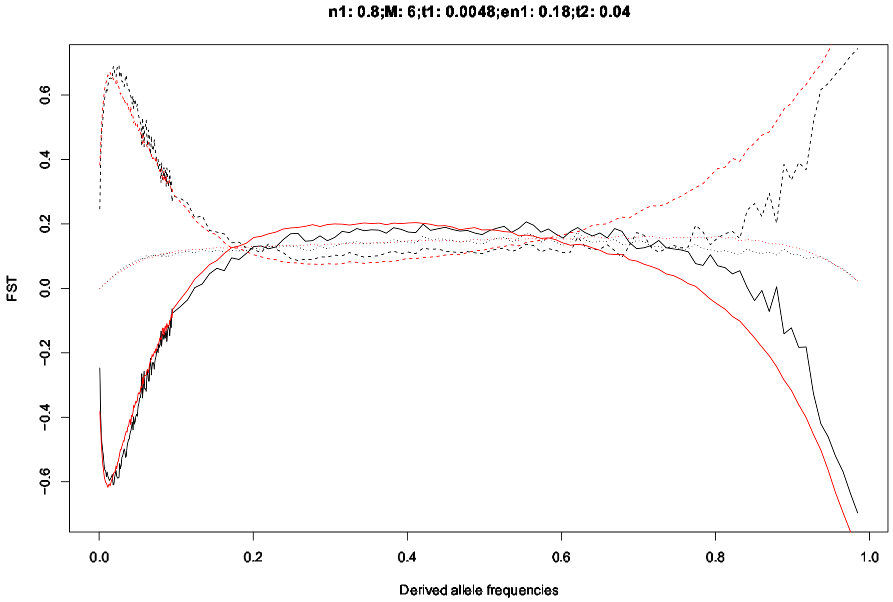


Model viewpoint

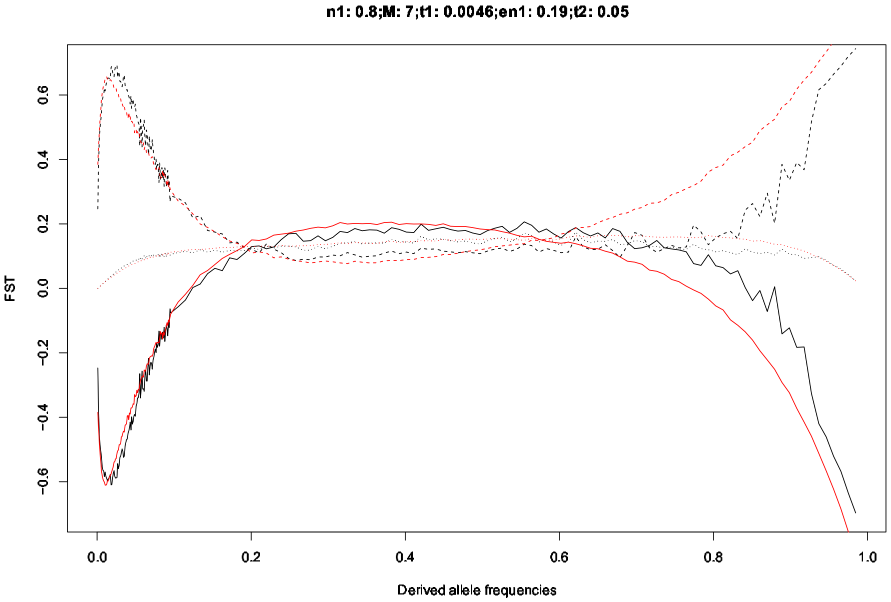
```
ms1 1046 1000 -t 600 -r 600 250000 -I 2 410 636 M
-n 2 N1 -en t1 2 en1 -eM t1 M -ej t2 2 1
```



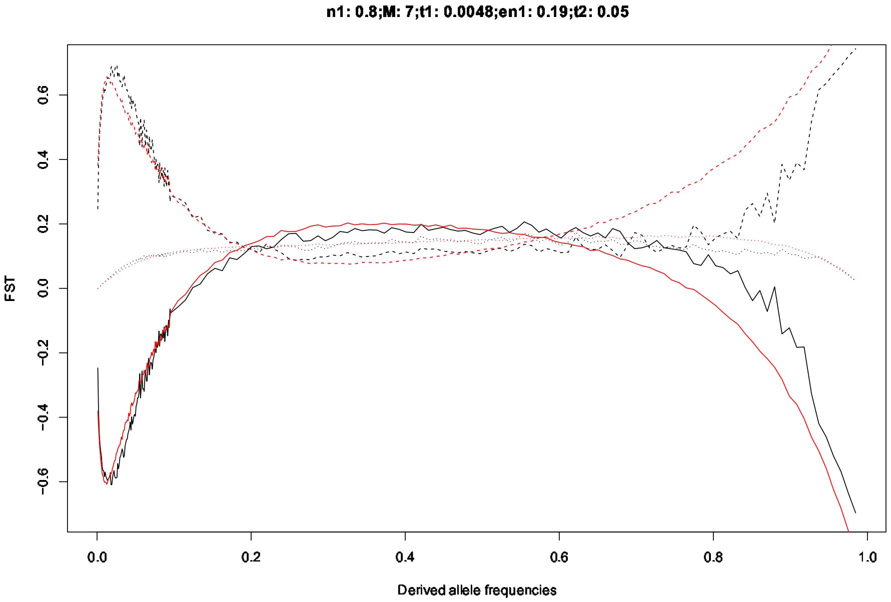
Two Populations



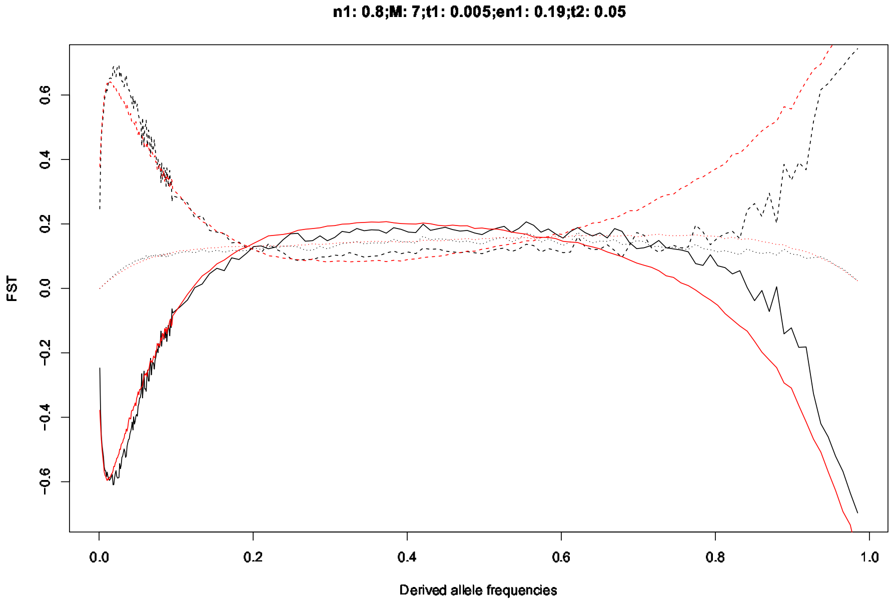
Two Populations



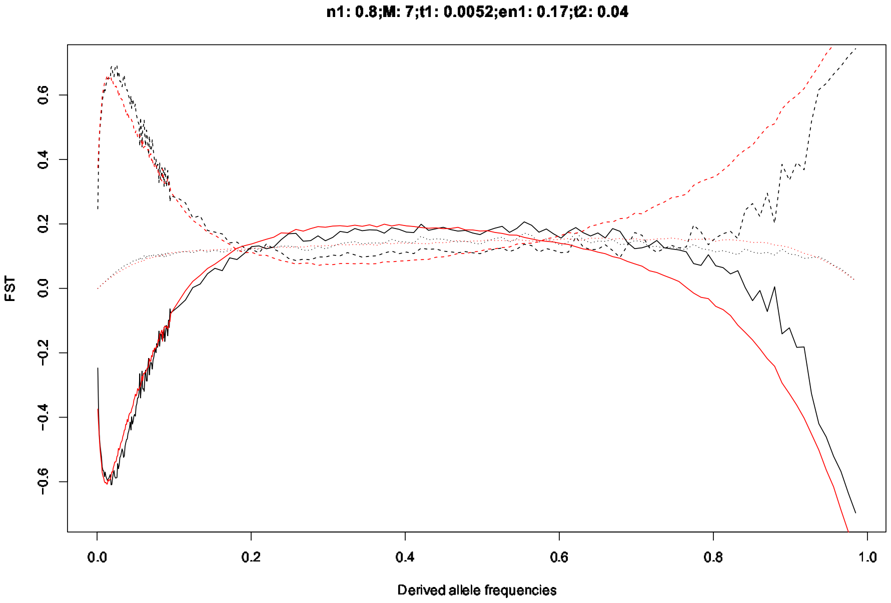
Two Populations



Two Populations



Two Populations


n1: 0.8;M: 7;t1: 0.0052;en1: 0.17;t2: 0.04
FST
Derived allele frequencies