F_{ST} generalized for arbitrary population structures

New York Area Population Genomics Workshop 2016

Alejandro Ochoa and John D. Storey

Center for Statistics and Machine Learning, and
Lewis-Sigler Institute for Integrative Genomics,
Princeton University

2016-01-21

F_{ST} and "island" models

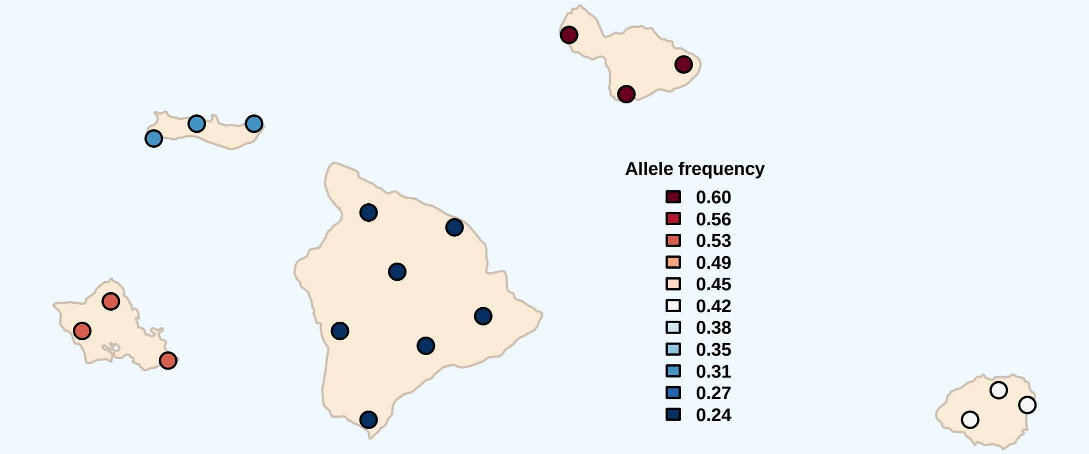
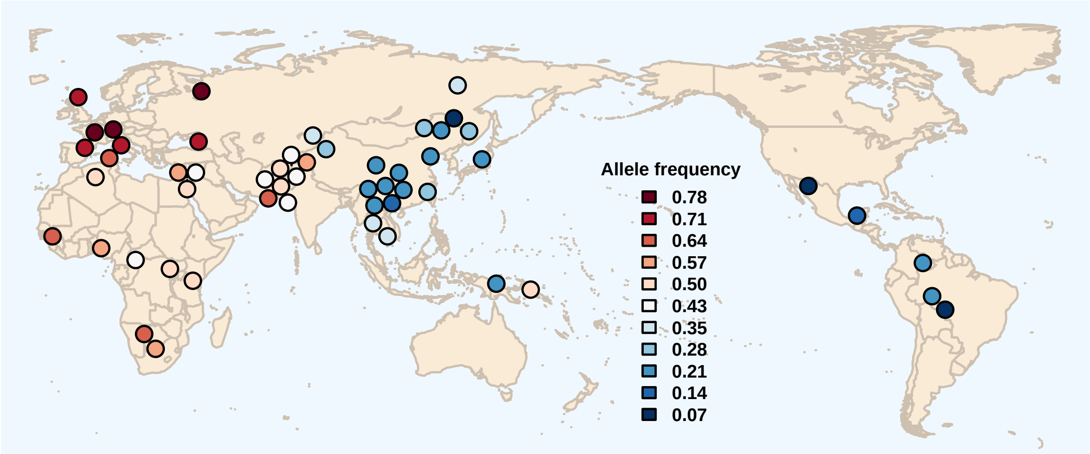
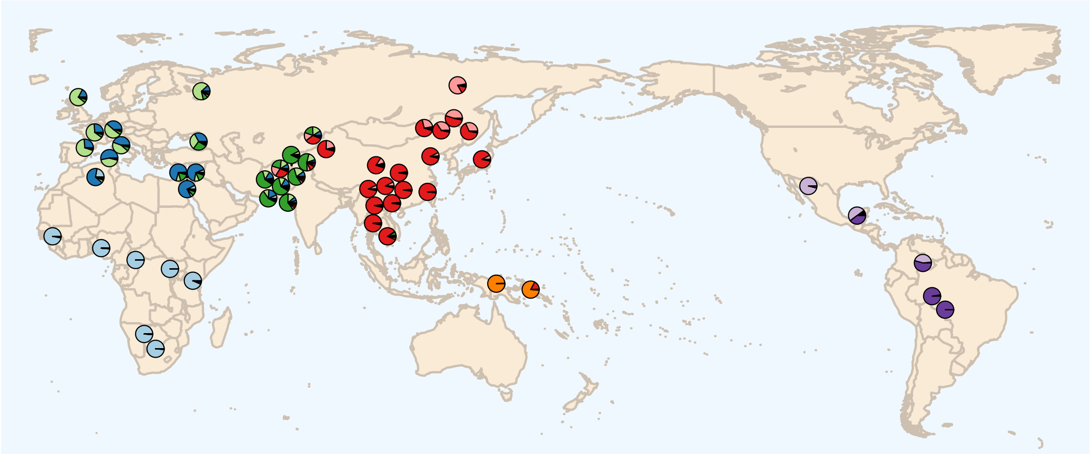


Illustration (not real data)

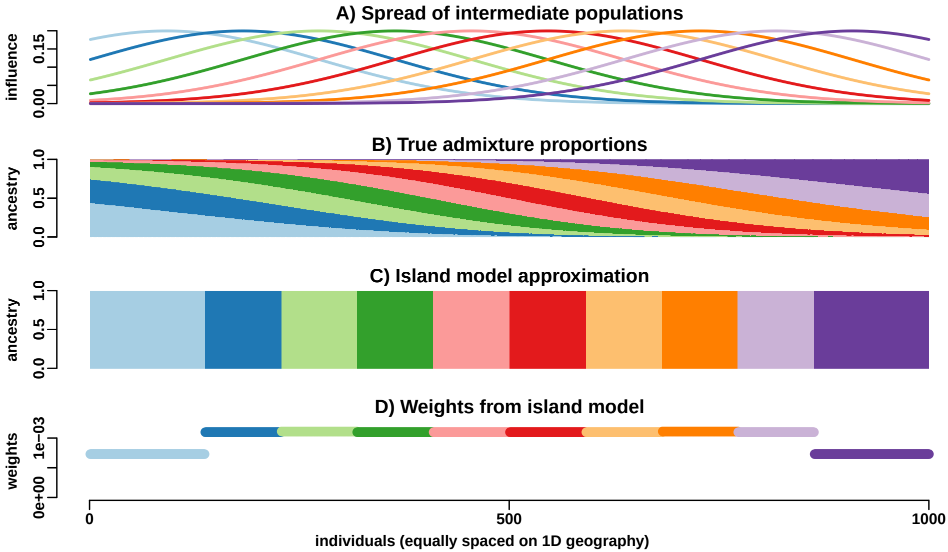
Allele frequencies in human populations



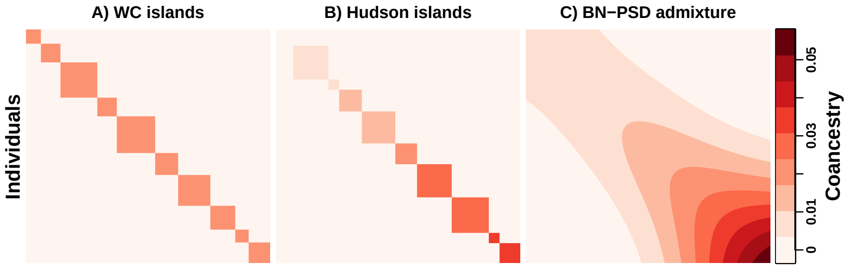
Admixture in human populations



Our admixture simulation



Our contribution



Previous F_{ST} definitions/estimators assume subdivided, independent populations.

We generalize F_{ST} for **arbitrary populations**, in terms of **individuals**, using **inbreeding** and **kinship** coefficients.

We characterize the **bias** of popular **estimators**, through theory and simulations.

An unstructured population

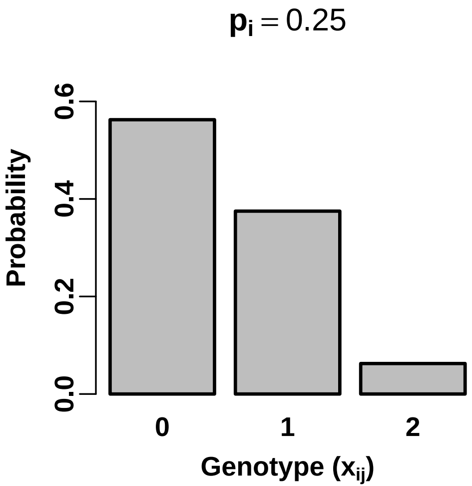
A population is "unstructured" if its individuals mate randomly.

In a large population, genotypes

$$x_{ij} \sim \text{Binomial}(2, p_i),$$

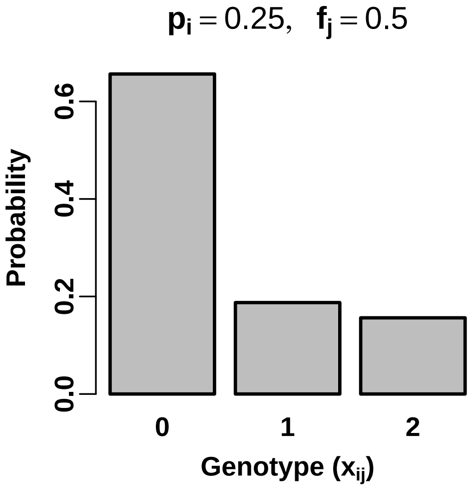
at SNP i with reference allele frequency p_i, for any individual j.

This is "Hardy-Weinberg Equilibrium".



Inbreeding rises in structured and small populations

"Inbreeding coefficient" f_j: probability that the two alleles of individual j at a random SNP are "identical by descent" (IBD) **given** an ancestral population.



Kinship coefficients quantify relatedness

"Kinship coefficient" φ_{jk}: probability that one allele of individual j and one of individual k, at a random SNP, are IBD, **given** an ancestral population.

Kinship given **unrelated founders**

j, k relation	φ_{jk}
self	$1/2$
child	$1/4$
sibling	$1/4$
half sibling	$1/8$
uncle or nephew	$1/8$
first cousins	$1/16$
second cousins	$1/64$
unrelated	0

What is F_{ST}? Wright (1951)

Given a "subdivided" population...

T: total population
S: a subpopulation of T
I: an individual in S

We define these coefficients:

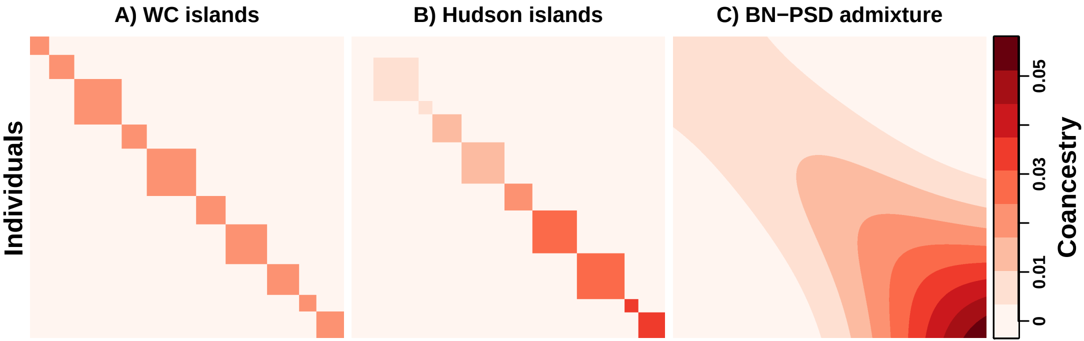
F_{IT}: total inbreeding (of I relative to T)
F_{IS}: local inbreeding (of I relative to S)
F_{ST}: inbreeding due to the population structure (of S relative to T).

These coefficients are related by:

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}).$$

F_{ST} is the inbreeding coefficient that individuals in S would have, relative to T, if they mated randomly.

Comparison of models assumed for F_{ST} estimation



Kinship model for genotypes

Let T be the ancestral population. In the absence of selective pressures, allele frequencies drift randomly from the ancestral frequency p_i^T, with covariances modulated by the kinship coefficients:

$$\begin{aligned} \mathrm{E}[x_{ij} | T] &= 2p_i^T, \\ \mathrm{Var}(x_{ij} | T) &= 2p_i^T(1 - p_i^T)(1 + f_j^T), \\ \mathrm{Cov}(x_{ij}, x_{ik} | T) &= 4p_i^T(1 - p_i^T)\varphi_{jk}^T. \end{aligned}$$

Note that $\varphi_{jj}^T = \frac{1}{2}(1 + f_j^T)$.

(Wright 1921, Malécot 1948, Wright 1951, Jacquard 1970).

Individual-level analogs of F_{IT}, F_{IS}, F_{ST}

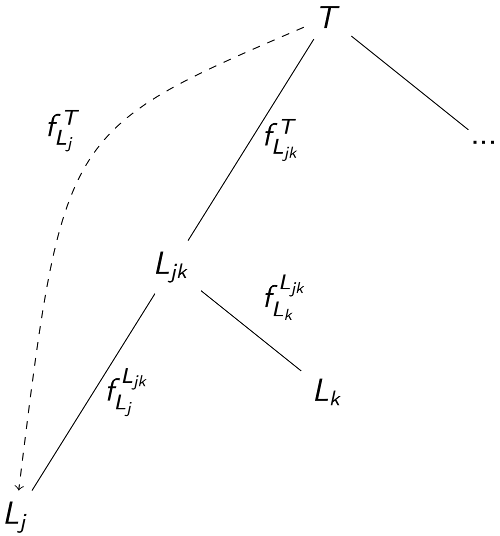
"Total" coef., analogous to F_{IT}:
f_j^T and φ_{jk}^T are relative to T.

"Local" coef., analogous to F_{IS}:
$f_j^{L_j}$ is relative to L_j,
$\varphi_{jk}^{L_{jk}}$ is relative to L_{jk}.

"Structural" coef., analogous to F_{ST}:

$$f_{L_j}^T = \frac{f_j^T - f_j^{L_j}}{1 - f_j^{L_j}},$$

$$f_{L_{jk}}^T = \frac{\varphi_{jk}^T - \varphi_{jk}^{L_{jk}}}{1 - \varphi_{jk}^{L_{jk}}}.$$



F_{ST} for arbitrary population structures

We propose

$$F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j f_{L_j}^T,$$

where $\sum_{j=1}^{n} w_j = 1$ are non-negative weights.

Backward compatible with island models (needs specific weights), and coherent with Wright's original definition.

Local inbreeding is removed on an **individual** basis!

"Coancestry" model and individual allele frequencies

This restricted model assumes the existence of "individual-specific allele frequencies" π_{ij}, modulated by "coancestry" coefficients θ^T_{jk}:

$$\begin{aligned} x_{ij} | \pi_{ij} &\sim \text{Binomial}(2, \pi_{ij}), \\ \mathrm{E}[\pi_{ij} | T] &= p_i^T, \\ \mathrm{Cov}(\pi_{ij}, \pi_{ik} | T) &= p_i^T (1 - p_i^T) \theta^T_{jk}. \end{aligned}$$

This model excludes local relationships. Given these assumptions, coancestry and kinship coefficients are the same:

$$\theta^T_{jk} = \begin{cases} \varphi^T_{jk} & \text{if} \quad j \neq k, \\ 2\varphi^T_{jj} - 1 = f^T_j & \text{if} \quad j = k. \end{cases}$$

Bias estimating marginal allele variance

The term $p_i(1-p_i)$ recurs in our models. The simplest estimator is biased:

$$\hat{p}_i = \sum_{j=1}^{n} w_j \pi_{ij} \quad \Rightarrow$$

$$\mathrm{E}[\hat{p}_i(1-\hat{p}_i)] = p_i(1-p_i)(1-\bar{\theta}),$$

where $\bar{\theta} = \sum_{j=1}^{n} \sum_{k=1}^{n} w_j w_k \theta_{jk}$ is the mean coancestry across individuals in our data. Since $0 \leq \bar{\theta} \leq 1$, the bias is always downward.

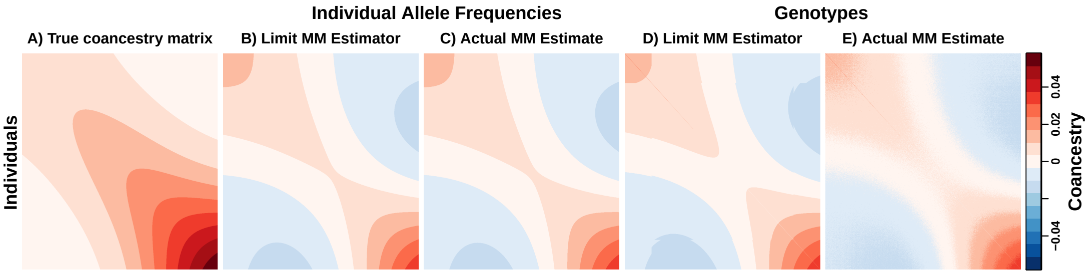
The same things happens if we use genotypes ($\bar{\theta}$ replaced by $\bar{\varphi}$).

Bias estimating kinship/coancestry coefficients

The popular kinship estimator from genotypes, and its limit as $m \to \infty$, are

$$\hat{\varphi}_{jk} = \frac{\sum_{i=1}^{m} (x_{ij} - 2\hat{p}_i)(x_{ik} - 2\hat{p}_i)}{4\sum_{i=1}^{m} \hat{p}_i(1 - \hat{p}_i)} \xrightarrow{\text{a.s.}} \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1 - \bar{\varphi}},$$

where $\bar{\varphi}_j = \sum_{k=1}^{n} w_k \varphi_{jk}$ and $\bar{\varphi} = \sum_{j=1}^{n} \sum_{k=1}^{n} w_j w_k \varphi_{jk}$. Bias in admixture sim.:

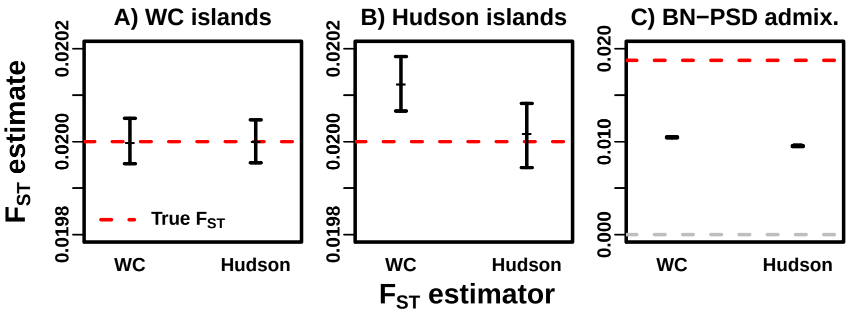


Bias estimating the generalized F_{ST}

A "simple" F_{ST} estimator, derived from $\hat{\theta}_{jj}$, is also biased as $m \to \infty$:

$$\hat{F}_{\text{ST}} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} w_j (\pi_{ij} - \hat{p}_i)^2}{\sum_{i=1}^{m} \hat{p}_i (1 - \hat{p}_i)} \xrightarrow{\text{a.s.}} \frac{F_{\text{ST}} - \bar{\theta}}{1 - \bar{\theta}}.$$

WC and Hudson F_{ST} estimators are similarly biased in our admixture simulation:



In this work, we...

...generalized F_{ST} using IBD probabilities for individuals.

...connected F_{ST}, kinship coefficients, and admixture models.

...proved almost sure convergence of simple estimators to biased quantities.

...used an admixture simulation to illustrate biases.

Our models could lead to more robust estimators.

Thanks!

John D. Storey
Andrew Bass
Irineo Cabreros
Chee Chen
Sean Hackett
Wei Hao
Emily Nelson

Neo Christopher Chung
(Wroclaw University of Life Sciences)

PRINCETON UNIVERSITY

CENTER FOR STATISTICS AND MACHINE LEARNING