Research Discussion Group: Tell us about yourself!

Name: Alejandro Ochoa

Summer Lab: John Storey's (genomics and statistics; I'm a PostDoc)

Undergrad institution: Massachusetts Institute of Technology
(PhD from Princeton University)

Home town: El Paso, TX / Ciudad Juárez, México.

Fun facts: Play guitar, used to be in a grad student rock band.
Love languages: English, Spanish, French; Latin, Medieval Spanish, Nahuatl; try to learn a little of many other languages.

F_{ST} generalized for arbitrary population structures

Summer Undergraduate Research Program

Alejandro Ochoa and John D. Storey

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

2016-06-09

F_{ST} and "island" models

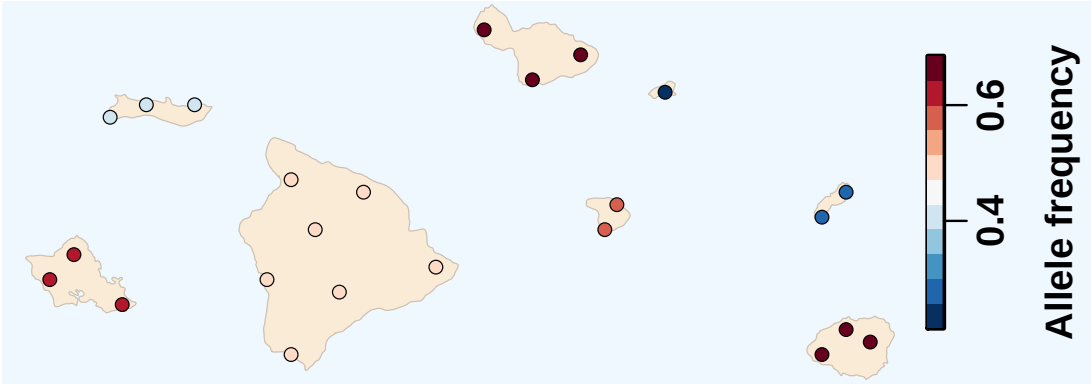
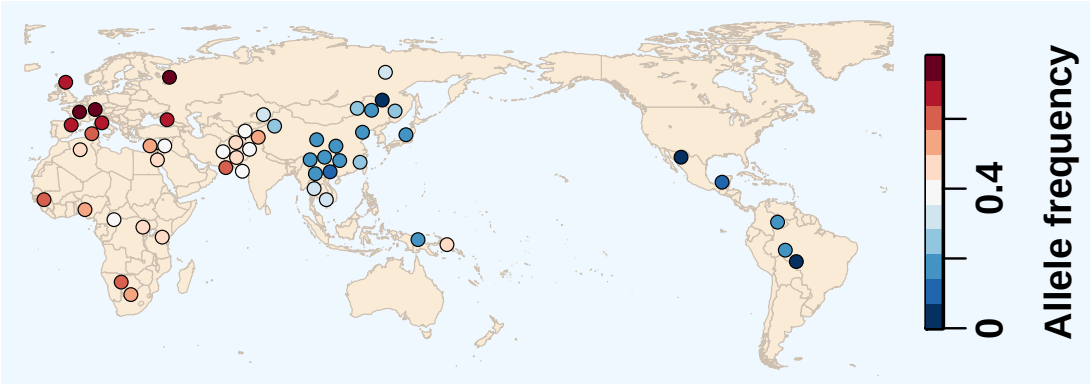


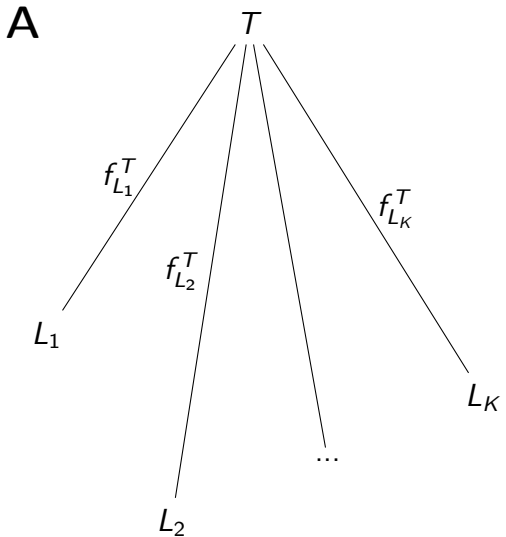
Illustration (not real data)

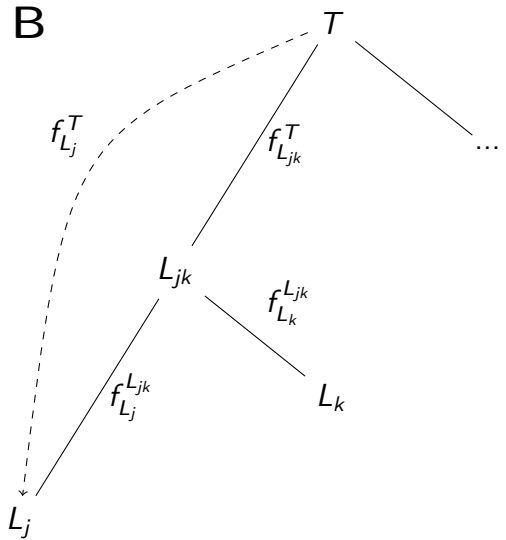
Allele frequencies in human populations



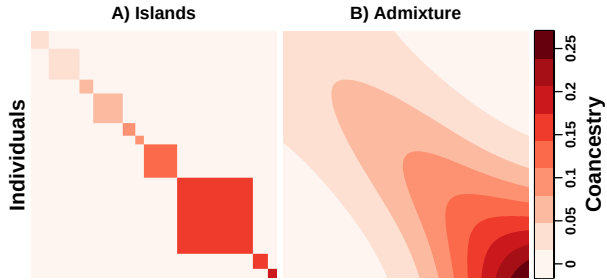
Median differentiation SNP (rs1528341)

Island model (A) versus arbitrary structure (B)





Our contribution



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

We generalize F_{ST} for **arbitrary populations**, in terms of **individuals**.

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

F_{ST} for arbitrary population structures

We propose

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

where

- $f_{L_j}^{T}$ = inbreeding coefficient of L_j relative to T
- $w_j \geq 0, \sum_{j=1}^{n} w_j = 1$ are weights

Backward compatible with island models.
Coherent with Wright's 1951 definition.

F_{ST} estimation under the island model

Weir-Cockerham and Hudson F_{ST} estimators using perfect AFs (π_{ij}) reduce to

$$\hat{p}_i = \frac{1}{n}\sum_{j=1}^{n}\pi_{ij},$$

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\pi_{ij} - \hat{p}_i)^2,$$

$$\hat{F}_{ST}^{\text{island}} = \frac{\sum_{i=1}^{m} s_i^2}{\sum_{i=1}^{m}\hat{p}_i(1-\hat{p}_i) + \frac{1}{n}s_i^2} \xrightarrow[m\to\infty]{\text{a.s.}} F_{ST}.$$

Under the island model, F_{ST} can be solved for:

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m} s_i^2\right] = \overline{p(1-p)}F_{ST},$$

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{p}_i(1-\hat{p}_i)\right] = \overline{p(1-p)}\left(1 - \frac{F_{ST}}{n}\right)$$

F_{ST} estimation under arbitrary coancestry

Weir-Cockerham and Hudson F_{ST} estimators using perfect AFs (π_{ij}) reduce to

$$\hat{p}_i = \frac{1}{n}\sum_{j=1}^{n}\pi_{ij},$$

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\pi_{ij}-\hat{p}_i)^2,$$

$$\hat{F}_{ST}^{\text{island}} = \frac{\sum_{i=1}^{m} s_i^2}{\sum_{i=1}^{m}\hat{p}_i(1-\hat{p}_i)+\frac{1}{n}s_i^2} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{n\left(F_{ST}-\bar{\theta}\right)}{n-1+F_{ST}-n\bar{\theta}}$$

Under the general coancestry model, system is underdetermined:

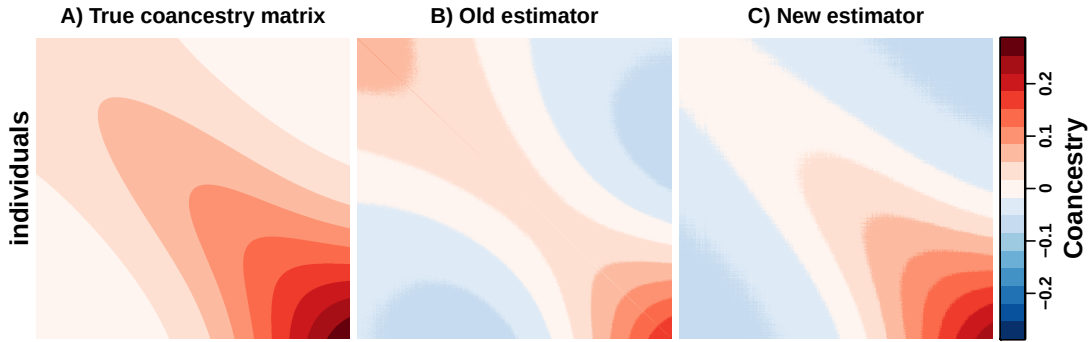
$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}s_i^2\right] = \overline{p(1-p)}\frac{n(F_{ST}-\bar{\theta})}{n-1},$$

$$\mathrm{E}\left[\frac{1}{m}\sum_{i=1}^{m}\hat{p}_i(1-\hat{p}_i)\right] = \overline{p(1-p)}(1-\bar{\theta}).$$

$\bar{\theta}$ = mean coancestry.
In islands, $\bar{\theta} = \frac{1}{n}F_{ST}$.

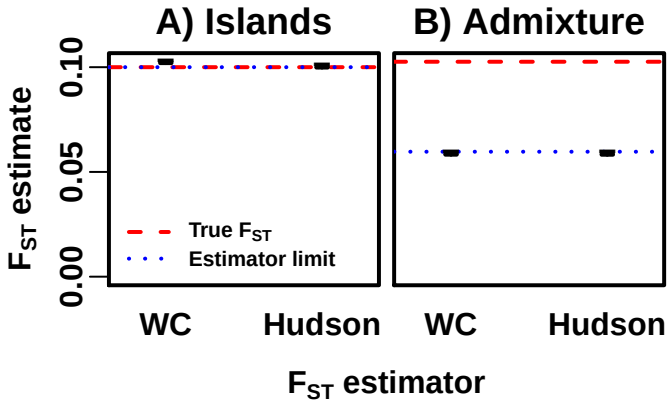
Bias estimating coancestry coefficients

A popular coancestry estimator ("old") is very distorted. We have a "new" estimator that is biased but not distorted. Illustrated in our admixture simulation:

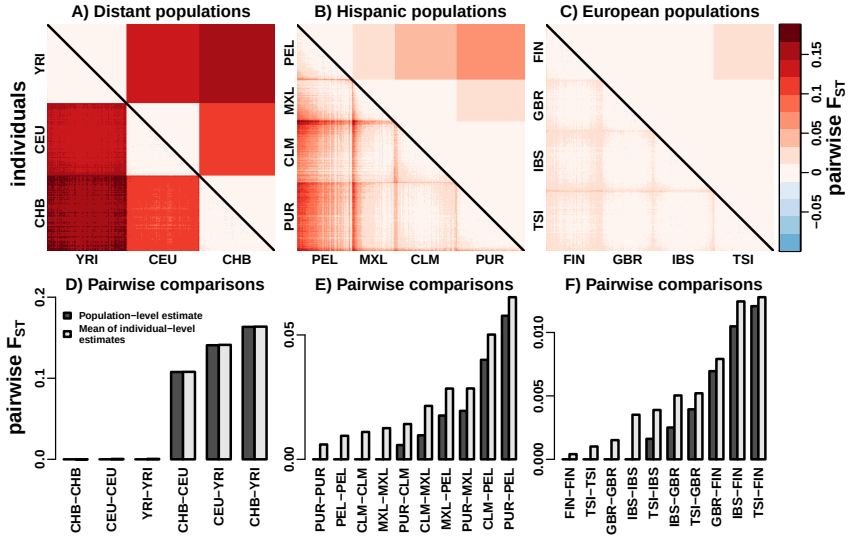


Bias estimating the generalized F_{ST}

The popular Weir-Cockerham (WC) and Hudson F_{ST} estimators, formulated for islands, are biased in our admixture simulation:



Population-level and Individual-level distances in 1000 Genomes



We have...

...generalized F_{ST} using parameters for individuals.

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

...found that common estimators are biased.

...used an admixture simulation and human data 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)