F_{ST} and kinship for arbitrary population structures

Alejandro Ochoa

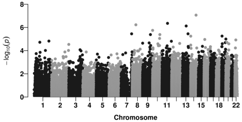
John D. Storey Lab

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

2017-02-13

Why study F_{ST} and kinship?

Human genetics is fascinating!

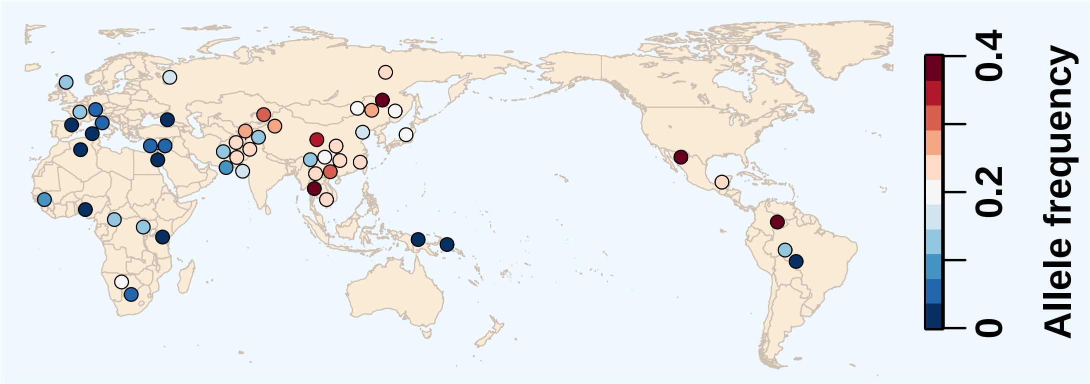


Pop. structure confounds association studies (GWAS)

Heritability of complex traits

Animal and plant breeding

F_{ST} measures population structure and differentiation



Median differentiation SNP (rs11692531)
$\hat{F}_{ST} \approx 0.081$ using Weir-Cockerham estimator
Human Genome Diversity Project (HGDP)

F_{ST} in the independent subpopulation model

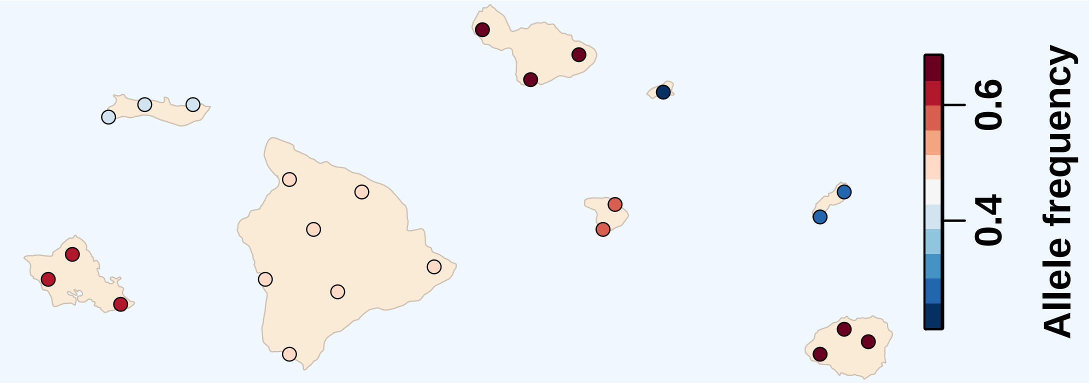
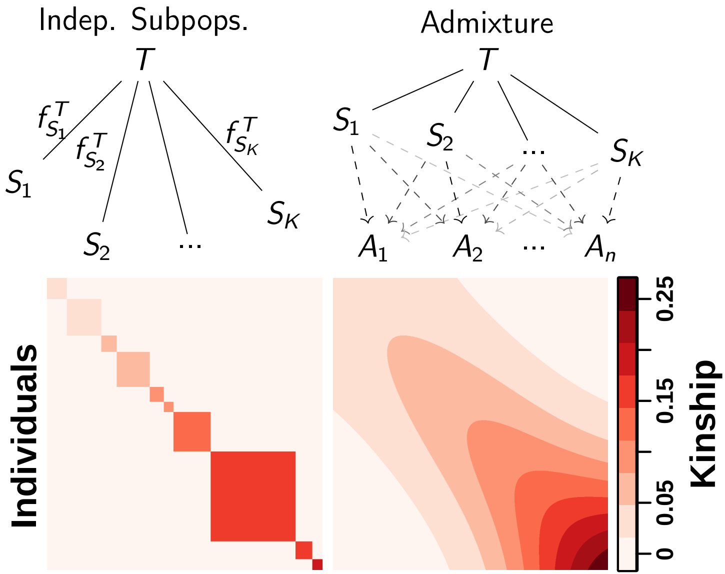


Illustration.

$$F_{\mathrm{ST}} = \frac{\mathrm{Var}\left(p_i^S \middle| T\right)}{p_i^T\left(1 - p_i^T\right)}.$$

F_{ST} estimation is constrained to independent subpopulations



Our contribution

Previous F_{ST} definitions/estimators assume independent subpopulations.

1. We generalize F_{ST} for **arbitrary populations**, in terms of **individuals**.
2. We characterize the **bias** of popular **estimators** under arbitrary population structure, through theory and simulations.
3. We develop a **new estimator** of kinship and F_{ST} for arbitrary population structures.

Confusion: three versions of F_{ST}

Definition 1: F_{ST} as a measure of **relatedness** in a population

$$F_{\text{ST}} = \bar{f}_S^T = \theta^T \quad \text{or} \quad \bar{\theta}^T.$$

Initially estimated from pedigrees.

Definition 2: F_{ST} as a **parameter** controlling allelic variance

$$F_{\text{ST}} = \frac{\text{Var}\left(p_i^S \middle| T\right)}{p_i^T\left(1 - p_i^T\right)}.$$

Def. 1 $\Rightarrow$ Def. 2 with F_{ST}

- Shared across loci i.
- No μ or selection.

Definition 3: F_{ST} as a **statistic** of locus-specific variance

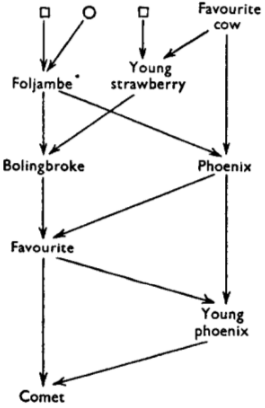
$$F_{\text{ST},i} = \frac{\hat{\sigma}_i^2}{\bar{p}_i(1 - \bar{p}_i)}.$$

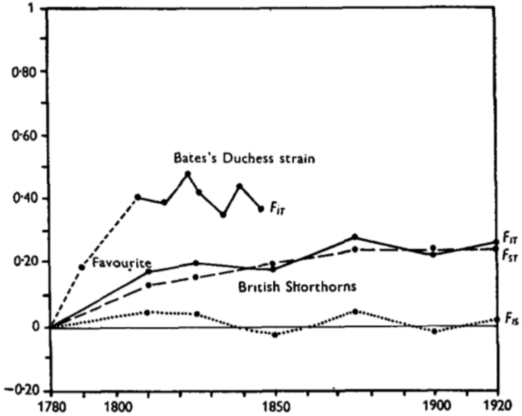
Goals:

- Varies per locus i.
- Measures μ and selection.

Our generalized definition corresponds most closely to **Definition 1**.

Wright's F_{ST} in cattle

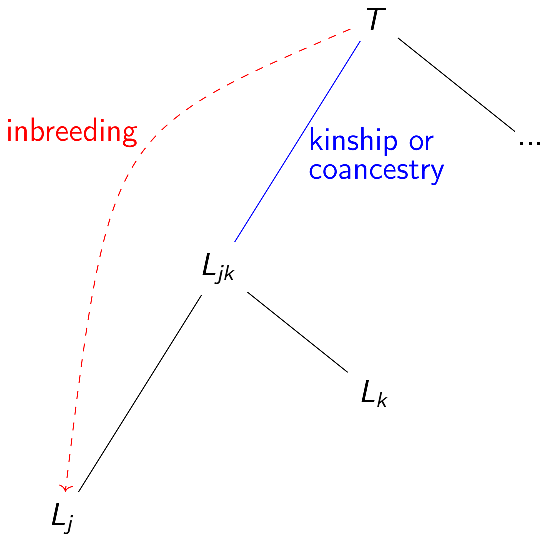




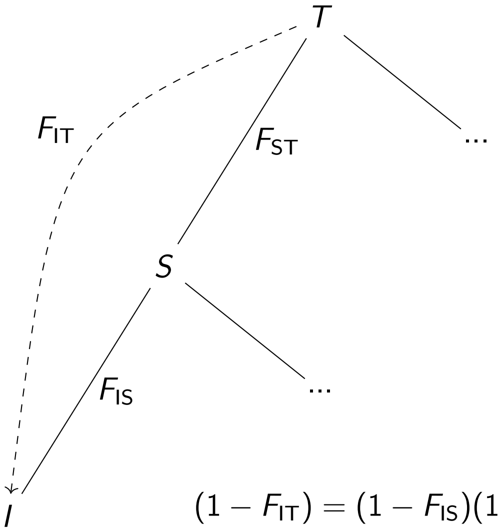
Populations:
T: Shorthorn
S: Dutchess strain

Wright (1951)

Populations related by a tree



F_{ST} in a subdivided population: Wright (1951)



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

Admixed populations have complex structures

US individuals are often admixed from populations across the world.

- European: UK, Ireland, Germany, Italy
- African: West Africa
- Hispanic: Puerto Rico, Mexico
- Asian: China, India

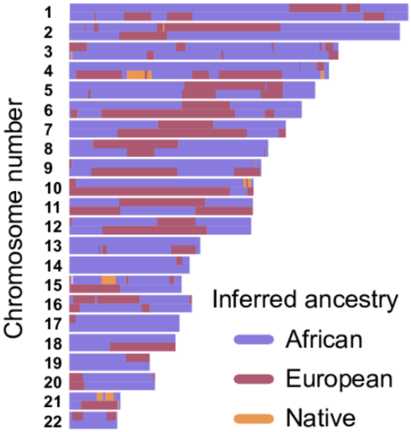
African-Americans and Hispanics are recently admixed (5-15 generations ago) from differentiated populations.

Admixture proportions vary (admix. LD) $\Rightarrow$ complex kinship.

GWAS and heritability estimation in multiethnic or admixed data?

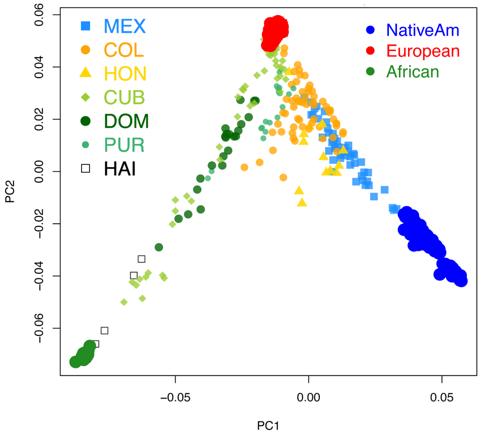
Recently admixed populations

African-Americans



Baharian, *et al.* (2016)

Hispanics



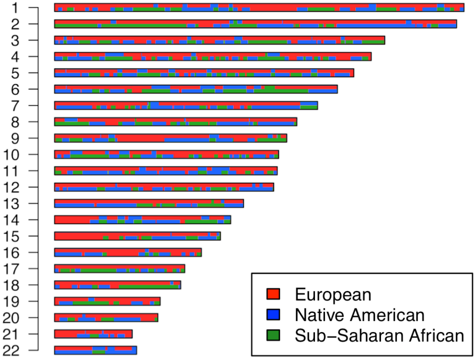
Moreno-Estrada, *et al.* (2013)

Admixed siblings from different populations?

Lucy and Maria, UK

Ochoa brothers, MX

High Admixture LD:

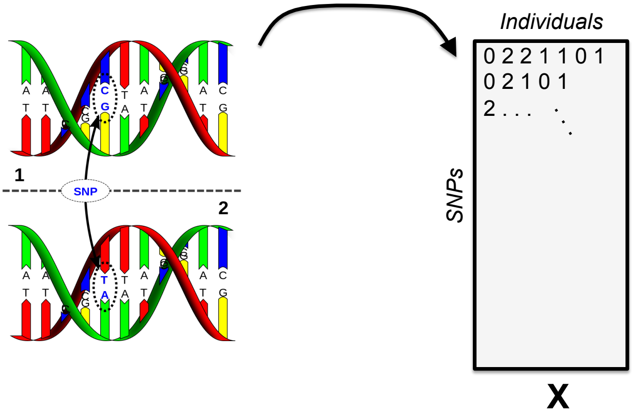


Moreno-Estrada, *et al.* (2013)

Solution: treat every individual as its own population!

SNP data

Example:	Genotype	CC	CT	TT
	x_{ij}	0	1	2



An unstructured population

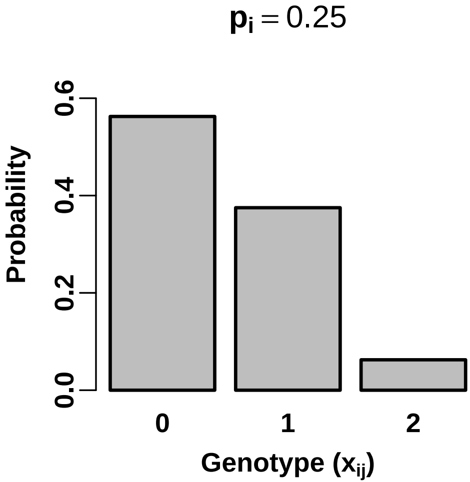
Individuals mate randomly.

In a large population T, genotypes

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

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

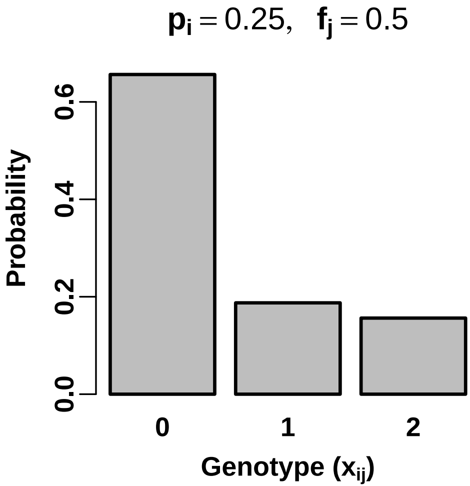
This is "Hardy-Weinberg Equilibrium".



Inbreeding coefficient f_j^T

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

A structured population has $f_j^T > 0$.



Kinship coefficients φ_{jk}^T

φ_{jk}^T: Probability that one allele of individual j and one of individual k, at a random SNP, are IBD, **given** an ancestral population T.

Local kinship,
given **unrelated founders**

j, k relation	φ_{jk}^T
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

Kinship model for genotypes

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

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

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

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

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

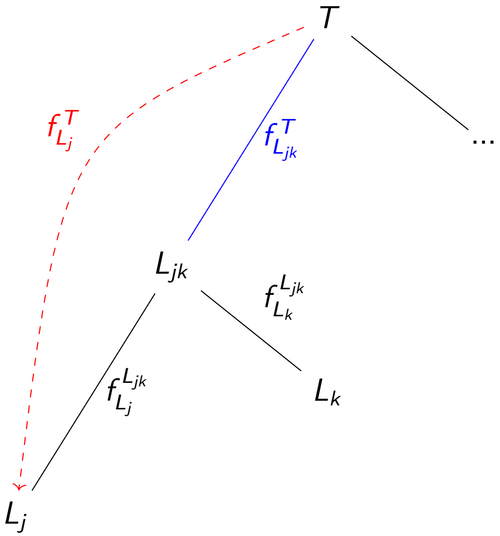
"Total" coef., analogous to F_{IT}:
f_j^T and φ_{jk}^T are relative to T.

"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

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

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 F_{ST} for subpopulations.
Coherent with Wright's 1951 definition.

Coancestry model and individual allele frequencies

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

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

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

$$\theta_{jk}^T = \begin{cases} \varphi_{jk}^T & \text{if} \quad j \neq k, \\ f_j^T = 2\varphi_{jj}^T - 1 & \text{if} \quad j = k. \end{cases} \qquad F_{\mathrm{ST}} = \sum_{j=1}^{n} w_j \theta_{jj}^T$$

F_{ST} estimation under independent subpopulations

Weir-Cockerham and Hudson F_{ST} estimators with π_{ij} simplify to

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

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

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

Under independent subpopulations, F_{ST} can be solved for:

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

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

F_{ST} estimation under arbitrary coancestry

Weir-Cockerham and Hudson F_{ST} estimators with π_{ij} simplify to

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

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

$$\hat{F}_{\mathrm{ST}}^{\mathrm{indep}} = \frac{\sum_{i=1}^{m} \hat{\sigma}_i^2}{\sum_{i=1}^{m} \hat{p}_i^T \left(1 - \hat{p}_i^T \right) + \frac{1}{n} \hat{\sigma}_i^2}$$

$$\xrightarrow[m \to \infty]{\text{a.s.}} \frac{n \left(F_{\mathrm{ST}} - \bar{\theta}^T \right)}{n - 1 + F_{\mathrm{ST}} - n \bar{\theta}^T}$$

Under the general coancestry model, system is underdetermined:

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

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

$\bar{\theta}^T$: mean coancestry.

In independent subpopulations $\bar{\theta}^T = \frac{1}{n} F_{\mathrm{ST}}$.

Admixture models

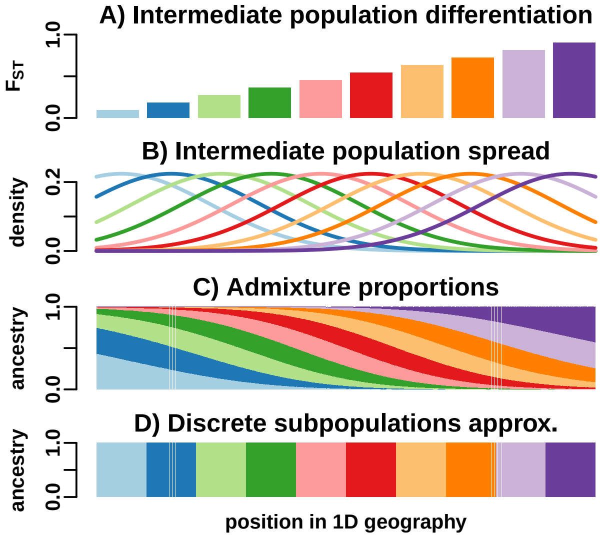
Draw alleles from a mixture of populations:

$$\pi_{ij} = \sum_{u=1}^{K} p_i^{S_u} q_{ju},$$

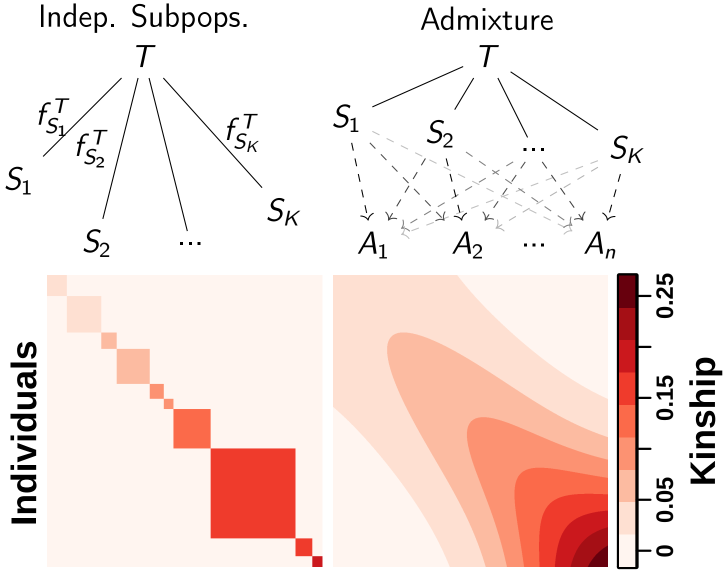
where q_{ju} is ancestry proportion, $p_i^{S_u}$ is AF in subpopulation S_u.
If subpopulations are independent and $f_{S_u}^{T}$ is F_{ST} of S_u relative to T, then

$$\theta_{jk}^{T} = \sum_{u=1}^{K} q_{ju} q_{ku} f_{S_u}^{T}, \qquad F_{\mathrm{ST}} = \sum_{j=1}^{n} \sum_{u=1}^{K} w_j q_{ju}^2 f_{S_u}^{T}.$$

Our admixture simulation

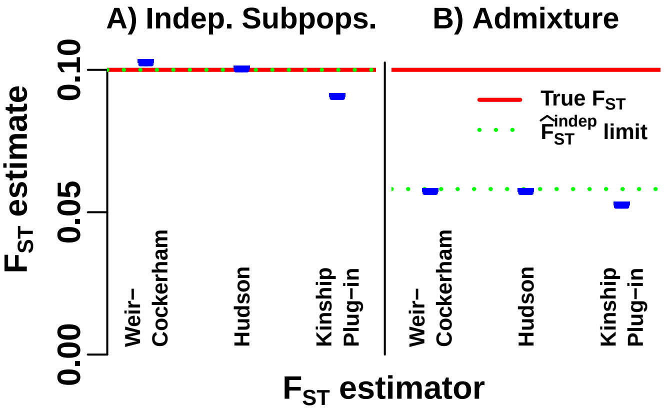


Comparison of population structures in simulation



Bias estimating the generalized F_{ST}

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

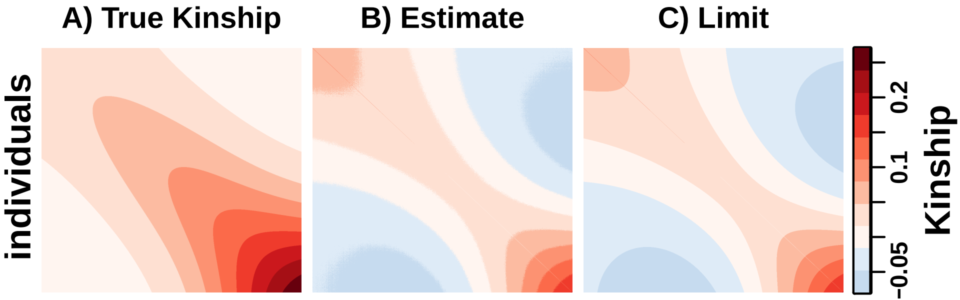


Bias estimating kinship coefficients

The popular kinship estimator from genotypes and its limit are

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

where $\bar{\varphi}_j^T$ and $\bar{\varphi}^T$ are weighted mean kinships. In our admixture simulation:



A new kinship estimator

Bias in new kinship estimator is parametrized by $\bar{\varphi}^T$:

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

$$\hat{\varphi}_{jk}^{T,\text{New}} = \frac{\sum_{i=1}^{m}(x_{ij} - 1)(x_{ik} - 1) - 1}{4\sum_{i=1}^{m} \hat{p}_i^T\left(1 - \hat{p}_i^T\right)} + 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \bar{\varphi}^T}{1 - \bar{\varphi}^T}.$$

Remaining bias in $\hat{\varphi}_{jk}^{T,\text{New}}$ comes from estimating $p_i^T\left(1 - p_i^T\right)$ with $\hat{p}_i^T\left(1 - \hat{p}_i^T\right)$.

A new kinship estimator

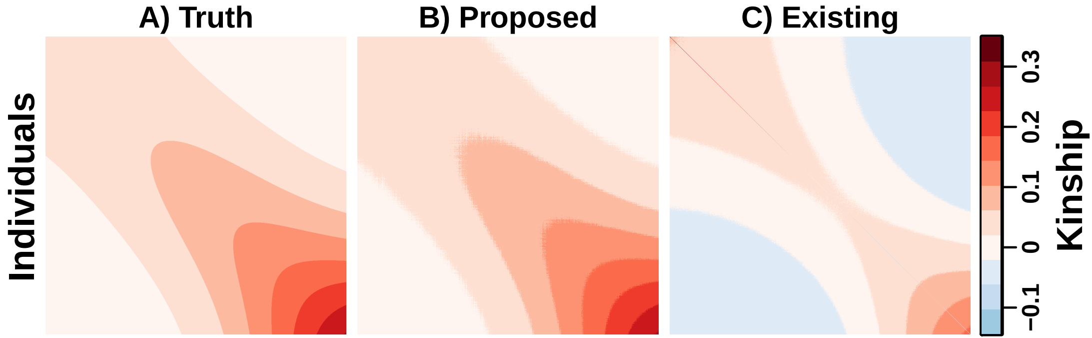
Limit of proposed estimate:

$$\hat{\varphi}_{jk}^{T,\text{New}} = \frac{\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1)-1}{4\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)}+1 \quad \xrightarrow[m\to\infty]{\text{a.s.}} \quad \frac{\varphi_{jk}^T-\bar{\varphi}^T}{1-\bar{\varphi}^T},$$

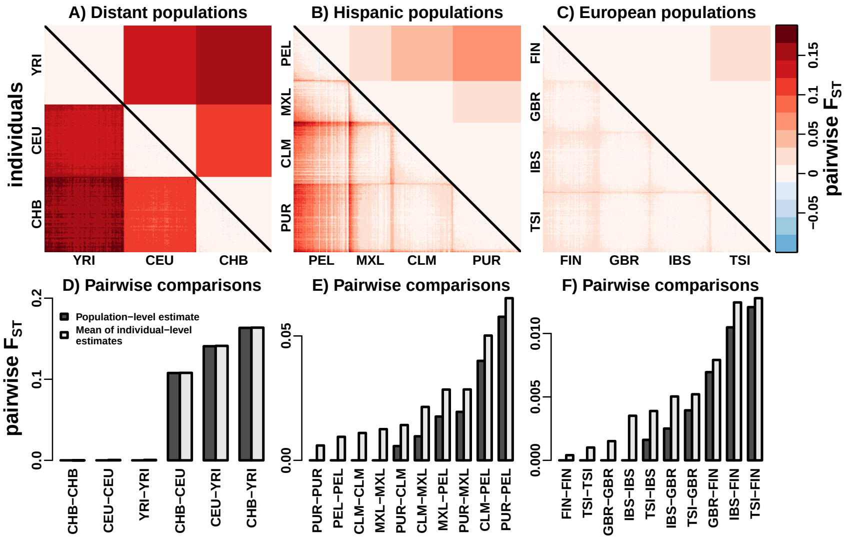
If $\min_{j,k}\varphi_{jk}^T = 0$, then

$$\min_{j,k}\hat{\varphi}_{jk}^{T,\text{New}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{-\bar{\varphi}^T}{1-\bar{\varphi}^T}$$

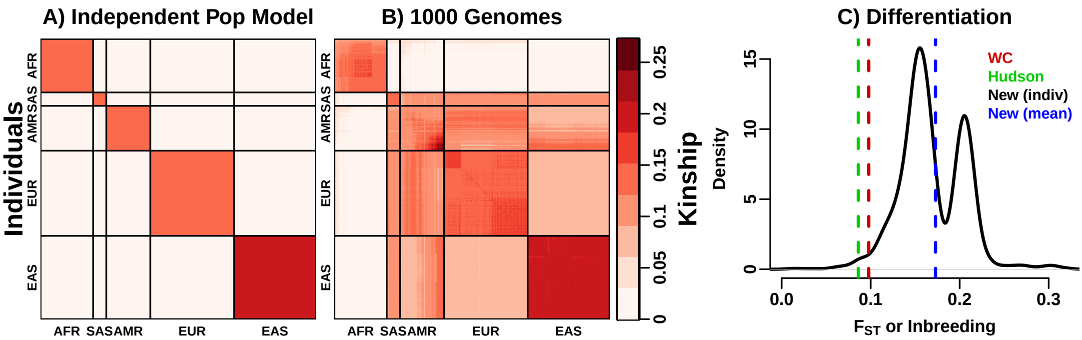
Performance of proposed estimator



Population-level and Individual-level distances in 1000 Genomes



Revised F_{ST} estimates in 1000 Genomes



We have...

...generalized F_{ST} using parameters for arbitrary structure in terms of individuals.

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

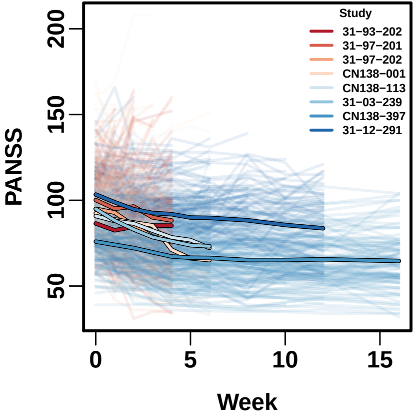
...characterized bias of common estimators when assumptions are broken.

...used an admixture simulation to illustrate biases.

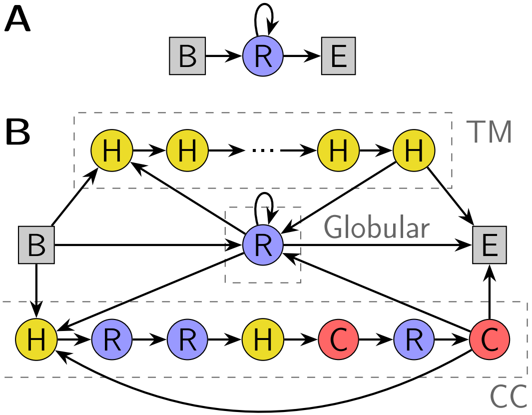
...developed new estimators of F_{ST} and kinship/coancestry.

Other work from Dr. Ochoa

Modeling the placebo response in psychiatric drug trials
Collaboration with Otsuka Pharma.



Protein sequence analysis
Improving sequence homology stats



Future work: Selection tests

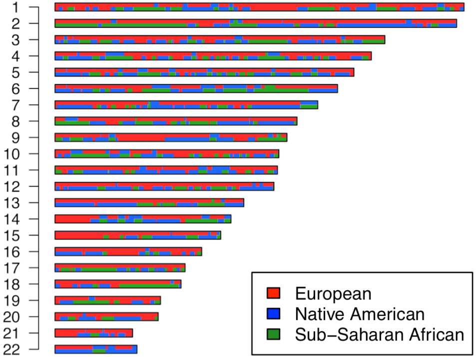
$\mathbf{x}_i$: genotype vector at SNP i,
$\hat{\Phi}^T$: kinship matrix estimate,
$\hat{p}_i^T$: ancestral allele frequency estimate,

Then this generalized z-score measures deviation of this SNP from the neutral genetic structure:

$$z_i^2 = \frac{\left(\mathbf{x}_i - 2\hat{p}_i^T \mathbf{1}\right)^\intercal \left(\hat{\Phi}^T\right)^{-1} \left(\mathbf{x}_i - 2\hat{p}_i^T \mathbf{1}\right)}{4\hat{p}_i^T \left(1 - \hat{p}_i^T\right)}.$$

Complements other info such as selective sweeps.

Future work: Admixture LD



Moreno-Estrada, *et al.* (2013)

Simple extension:
The kinship matrix varies per locus depending on population assignments.

More general local kinship estimation?

Future work: Kinship in Recent Mutations

Recall the following only holds for neutral SNPs polymorphic in T:

$$\mathrm{E}[x_{ij}|T] = 2p_i^T,$$
$$\mathrm{Cov}(x_{ij}, x_{ik}|T) = 4p_i^T \left(1 - p_i^T\right) \varphi_{jk}^T.$$

A SNP that arose from recent mutation in S instead has $p_i^T = 0$ or 1 and:

$$\mathrm{E}[x_{ij}|S] = 2p_i^S,$$
$$\mathrm{Cov}(x_{ij}, x_{ik}|S) = 4p_i^S \left(1 - p_i^S\right) \varphi_{jk}^S.$$

Also recall:

$$\left(1 - \varphi_{jk}^T\right) = \left(1 - \varphi_{jk}^S\right) \left(1 - f_S^T\right).$$

Recent mutations require special treatment in GWAS/herit. studies!

Acknowledgments

John D. Storey
Andrew Bass
Irineo Cabreros
Chee Chen
Wei Hao
Emily Nelson
Riley Skeen-Gaar

Neo Christopher Chung
Institute of Informatics
University of Warsaw

Funding:
National Institutes of Health
Otsuka Pharmaceuticals

PRINCETON UNIVERSITY

Lewis-Sigler Institute for Integrative Genomics

CENTER FOR STATISTICS AND MACHINE LEARNING

Future work: Variable kinship in GWAS

Suppose the kinship matrix $\Phi_i^T = (\varphi_{ijk}^T)$ varies per locus i:

$$\mathrm{Cov}\left(x_{ij}, x_{ik} \middle| T\right) = 4 p_i^T \left(1 - p_i^T\right) \varphi_{ijk}^T.$$

This Φ_i^T replaces the global kinship Φ^T used in LMM and adjusted χ^2 GWAS, varying given local admixture or the recent mutation model.

Future work: Variable kinship in heritability estimation

Suppose the kinship matrix $\Phi_i^T = (\varphi_{ijk}^T)$ varies per locus i:

$$\mathrm{Cov}\left(x_{ij}, x_{ik} | T\right) = 4 p_i^T \left(1 - p_i^T\right) \varphi_{ijk}^T .$$

Let $\mathbf{y} = (y_j)$ be a trait controlled by additive genetic effects as

$$y_j = \mu + \sum_{i \in C} \beta_i x_{ij} + \epsilon_j ,$$

The trait's covariance structure is now given by the mean kinship at causal loci C:

$$\mathrm{Cov}(\mathbf{y} | T) = \sigma^2 \left(h^2 2 \bar{\Phi}^T + (1 - h^2) \mathbf{I} \right) , \quad \text{where}$$
$$\bar{\Phi}^T = \sum_{i \in C} w_i \Phi_i^T , \quad w_i \propto \beta_i^2 p_i^T (1 - p_i^T) .$$