$F_{ST}$ 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!

Heritability of complex traits

Pop. structure confounds association studies (GWAS)

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.

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

$$F_{ST} = \bar{f}_T^S = \theta^T \text{ or } \bar{\theta}^T.$$ 

Initially estimated from pedigrees.

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

$$F_{ST} = \frac{\text{Var}(p^s_T) T}{p^T_i(1 - p^T_i)}.$$

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

- Shared across loci $i$.
- No $\mu$ or selection.

Goals:

- Varies per locus $i$.
- Measures $\mu$ 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

$T$

$T$

$L_{jk}$

kinship or coancestry

inbreeding

$L_j$

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

$\left(1 - F_{IT}\right) = \left(1 - F_{IS}\right) \left(1 - F_{ST}\right)$
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) ⇒ 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?

Solution: treat every individual as its own population!

Moreno-Estrada, et al. (2013)
SNP data

Example: Genotype

\[ x_{ij} \]

0 1 2

Genome Wide SNP Data

Individuals

SNPs

0 2 2 1 1 0 1
0 2 1 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”.

\[ p_i = 0.25 \]
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 $f_j^T > 0$. 

\[ p_i = 0.25, \quad f_j = 0.5 \]
Kinship coefficients $\varphi^T_{jk}$

$\varphi^T_{jk}$: 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

<table>
<thead>
<tr>
<th>$j, k$ relation</th>
<th>$\varphi^T_{jk}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>self</td>
<td>1/2</td>
</tr>
<tr>
<td>child</td>
<td>1/4</td>
</tr>
<tr>
<td>sibling</td>
<td>1/4</td>
</tr>
<tr>
<td>half sibling</td>
<td>1/8</td>
</tr>
<tr>
<td>uncle or nephew</td>
<td>1/8</td>
</tr>
<tr>
<td>first cousins</td>
<td>1/16</td>
</tr>
<tr>
<td>second cousins</td>
<td>1/64</td>
</tr>
<tr>
<td>unrelated</td>
<td>0</td>
</tr>
</tbody>
</table>
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 $p^T_i$, with covariances modulated by the kinship coefficients:

$$E[x_{ij}|T] = 2p^T_i,$$
$$\text{Var}(x_{ij}|T) = 2p^T_i (1 - p^T_i) (1 + f^T_j),$$
$$\text{Cov}(x_{ij}, x_{ik}|T) = 4p^T_i (1 - p^T_i) \varphi^T_{jk}.$$

Note that $\varphi^T_{jj} = \frac{1}{2} (1 + f^T_j)$.

(Wright 1921, Malécot 1948, Wright 1951, Jacquard 1970).
Individual-level analogs of $F_{\text{IT}}$, $F_{\text{IS}}$, $F_{\text{ST}}$

“Total” coef., analogous to $F_{\text{IT}}$:

$f_j^T$ and $\varphi_{jk}^T$ are relative to $T$.

“Local” coef., analogous to $F_{\text{IS}}$:

$f_j^L$ is relative to $L_j$,

$\varphi_{jk}^L$ is relative to $L_{jk}$.

“Structural” coef., analogous to $F_{\text{ST}}$:

\[
\begin{align*}
  f_{L_j}^T &= \frac{f_j^T - f_j^L}{1 - f_j^L}, \\
  f_{L_{jk}}^T &= \frac{\varphi_{jk}^T - \varphi_{jk}^L}{1 - \varphi_{jk}^L}.
\end{align*}
\]
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 \( F_{ST} \) 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* $\pi_{ij}$, modulated by coancestry coefficients $\theta_{jk}^T$:

$$E[\pi_{ij} | T] = p_i^T,$$
$$\text{Cov}(\pi_{ij}, \pi_{ik} | T) = p_i^T (1 - p_i^T) \theta_{jk}^T,$$
$$x_{ij} | \pi_{ij} \sim \text{Binomial}(2, \pi_{ij}).$$

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

$$\theta_{jk}^T = \begin{cases} 
\varphi_{jk}^T & \text{if } j \neq k, \\
2\varphi_{jj}^T - 1 & \text{if } j = k.
\end{cases}$$

$$F_{ST} = \sum_{j=1}^{n} w_j \theta_{jj}^T$$
**$F_{ST}$ estimation under independent subpopulations**

Weir-Cockerham and Hudson $F_{ST}$ estimators with $\pi_{ij}$ simplify to

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

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

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

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

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

$$\mathbb{E} \left[ \frac{1}{m} \sum_{i=1}^{m} \hat{p}_i^T (1 - \hat{p}_i^T) \right] = p(1 - p)^T \left( 1 - \frac{F_{ST}}{n} \right).$$
**$F_{ST}$ estimation under arbitrary coancestry**

Weir-Cockerham and Hudson $F_{ST}$ estimators with $\pi_{ij}$ simplify to

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

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

$$\hat{F}_{ST}^{\text{indep}} = \frac{\sum_{i=1}^{m} \hat{\sigma}_i^2}{\sum_{i=1}^{m} \hat{p}_i^T (1 - \hat{p}_i^T) + \frac{1}{n} \hat{\sigma}_i^2} \quad \text{a.s.}$$

Under the general coancestry model, the system is underdetermined:

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

In independent subpopulations

$$\bar{\theta}^T = \frac{1}{n} F_{ST}.$$ 

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

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

Draw alleles from a mixture of populations:

\[ \pi_{ij} = \sum_{u=1}^{K} p_{S_u}^{i} q_{ju}, \]

where \( q_{ju} \) is ancestry proportion, \( p_{S_u}^{i} \) 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

\[ \theta_{jk}^{T} = \sum_{u=1}^{K} q_{ju} q_{ku} f_{S_u}^{T}, \quad F_{ST} = \sum_{j=1}^{n} \sum_{u=1}^{K} w_j q_{ju}^2 f_{S_u}^{T}. \]
Our admixture simulation

A) Intermediate population differentiation

B) Intermediate population spread

density

C) Admixture proportions

D) Discrete subpopulations approx.

position in 1D geography
Comparison of population structures in simulation

Indep. Subpops.  
\[ T \]
\[ f_{S_1}^T \]
\[ f_{S_2}^T \]
\[ S_1 \]
\[ f_{S_K}^T \]
\[ S_K \]

Admixture  
\[ T \]
\[ S_1 \]  \[ S_2 \]  \[ \ldots \]  \[ S_K \]
\[ \]
\[ A_1 \]  \[ A_2 \]  \[ \ldots \]  \[ A_n \]

Kinship

Individuals

0 0.05 0.15 0.25

Kinship
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:

![Graph showing bias in $F_{ST}$ estimation.

A) Indep. Subpops.  
- Weir-Cockerham  
- Hudson  
- Kinship Plug-in

B) Admixture  
- True $F_{ST}$  
- $F_{ST}$ limit  
- $F_{ST}$ estimate

<table>
<thead>
<tr>
<th>$F_{ST}$ estimator</th>
<th>A) Indep. Subpops.</th>
<th>B) Admixture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weir-Cockerham</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hudson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kinship Plug-in</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- | True $F_{ST}$ | $F_{ST}$ limit | $F_{ST}$ estimate |
Bias estimating kinship coefficients

The popular kinship estimator from genotypes and its limit are

\[
\hat{\phi}_{jk}^T = \frac{\sum_{i=1}^{m} (x_{ij} - 2\hat{p}_i^T) (x_{ik} - 2\hat{p}_i^T)}{4 \sum_{i=1}^{m} \hat{p}_i^T (1 - \hat{p}_i^T)}
\]

\[
\text{a.s.} \quad \frac{\phi_{jk}^T - \bar{\phi}_j^T - \bar{\phi}_k^T + \bar{\phi}^T}{1 - \bar{\phi}^T} \quad \rightarrow \quad m \rightarrow \infty
\]

where \( \phi_j^T \) and \( \phi^T \) are weighted mean kinships. In our admixture simulation:

A) True Kinship  B) Estimate  C) Limit
A new kinship estimator

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

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

$$
\hat{\phi}_{jk}^{T,\text{New}} = \sum_{i=1}^{m} (x_{ij} - 1)(x_{ik} - 1) - 1 \quad \text{a.s.} \quad \frac{\phi_{jk}^T - \bar{\phi}^T}{1 - \bar{\phi}^T}, \quad m \to \infty
$$

Remaining bias in $\hat{\phi}_{jk}^{T,\text{New}}$ comes from estimating $p_i^T (1 - p_i^T)$ with $\hat{p}_i^T (1 - \hat{p}_i^T)$. 
A new kinship estimator

Limit of proposed estimate:

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

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

\[
\min_{j,k} \hat{\varphi}_{jk,\text{New}}^T \quad \xrightarrow{\text{a.s.}} \quad \frac{-\bar{\varphi}^T}{1 - \bar{\varphi}^T},
\]
Performance of proposed estimator

A) Truth

B) Proposed

C) Existing

Kinship

Individuals
Population-level and Individual-level distances in 1000 Genomes

A) Distant populations

YRI
ynchronal populations

YRI

CEU

CHB

B) Hispanic populations

PEL

MXL

CLM

PUR

C) European populations

FIN

GBR

IBS

TSI

D) Pairwise comparisons

0.0

0.1

0.2

E) Pairwise comparisons

0.00

0.05

0.10

F) Pairwise comparisons

0.000

0.005

0.010

pairwise F_{ST}
Revised $F_{ST}$ estimates in 1000 Genomes

A) Independent Pop Model

B) 1000 Genomes

C) Differentiation

WC
Hudson
New (indiv)
New (mean)
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

![Graph showing PANSS scores over weeks for different studies.](image)

![Diagram of protein sequence analysis with labeled domains.](image)
Future work: Selection tests

\(x_i\): genotype vector at SNP \(i\),
\(\hat{\Phi}^T\): kinship matrix estimate,
\(\hat{\rho}_i^T\): ancestral allele frequency estimate,

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

\[
z_i^2 = \frac{(x_i - 2\hat{\rho}_i^T 1)^T (\hat{\Phi}^T)^{-1} (x_i - 2\hat{\rho}_i^T 1)}{4\hat{\rho}_i^T (1 - \hat{\rho}_i^T)}.
\]

Complements other info such as selective sweeps.
Future work: Admixture LD

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

More general local kinship estimation?

Moreno-Estrada, et al. (2013)
Future work: Kinship in Recent Mutations

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

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

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

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

Also recall:

$$1 - \varphi_{jk}^T = (1 - \varphi_{jk}^S) (1 - f_s^T).$$

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
Suppose the kinship matrix $\Phi^T_i = (\varphi^T_{ijk})$ varies per locus $i$:

$$\text{Cov} (x_{ij}, x_{ik} | T) = 4p_i^T (1 - p_i^T) \varphi^T_{ijk}.$$ 

This $\Phi^T_i$ replaces the global kinship $\Phi^T$ used in LMM and adjusted $\chi^2$ 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 $i$:

$$\text{Cov} (x_{ij}, x_{ik} \mid T) = 4p_i^T (1 - p_i^T) \varphi_{ijk}^T.$$ 

Let $y = (y_j)$ 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 $C$:

$$\text{Cov}(y \mid T) = \sigma^2 \left( h^2 2\Phi^T + (1 - h^2)I \right),$$

where

$$\Phi^T = \sum_{i \in C} w_i \Phi_i^T, \quad w_i \propto \beta_i^2 p_i^T (1 - p_i^T).$$