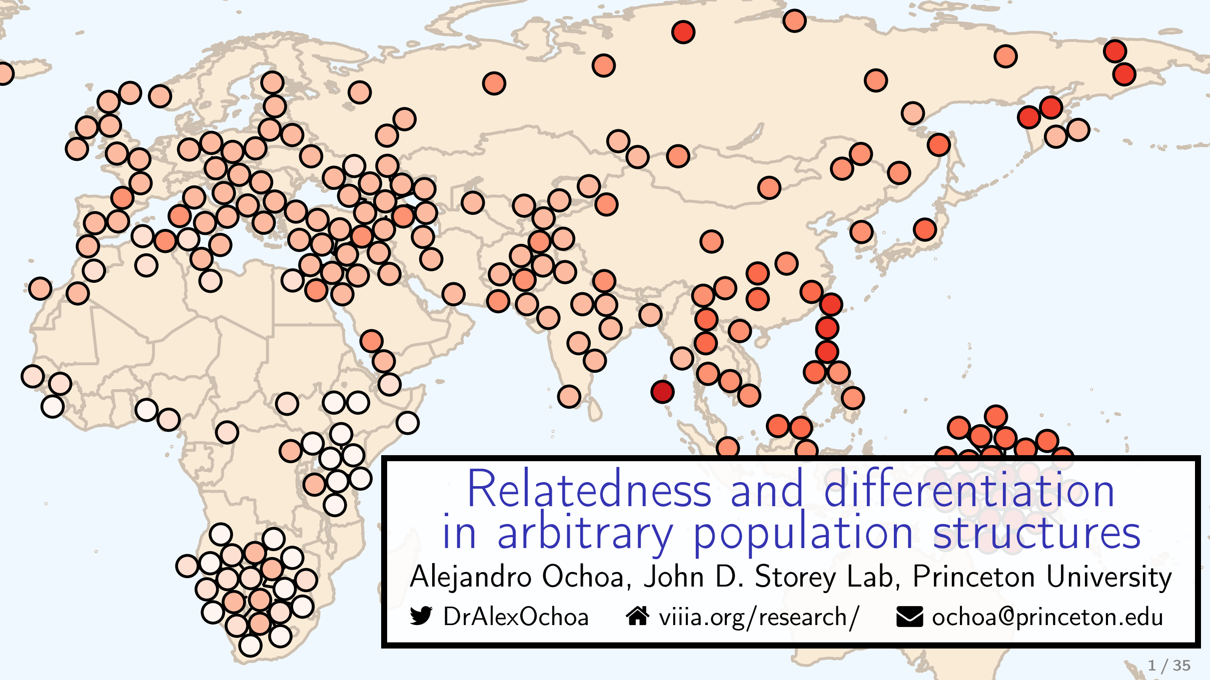


Relatedness and differentiation in arbitrary population structures

Alejandro Ochoa, John D. Storey Lab, Princeton University

DrAlexOchoa ☖ viiia.org/research/ ✉ ochoa@princeton.edu

My research areas / Contributions

1. Stratified False Discovery Rate (FDR)
    - Finding: per-stratum local FDR maximizes power controlling overall FDR
    - Improved power in protein domain prediction
    - Identified protein classes with problematic statistics

My research areas / Contributions

1. Stratified False Discovery Rate (FDR)
   - Finding: per-stratum local FDR maximizes power controlling overall FDR
   - Improved power in protein domain prediction
   - Identified protein classes with problematic statistics
2. Genetic and other factors controlling the placebo response
   - Collaboration with Otsuka Pharmaceutical
   - Mixed-effects modeling of longitudinal response in drug trials for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder
   - Genome-wide association study for individual-specific placebo response

My research areas / Contributions

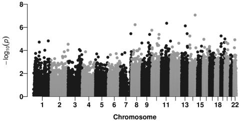
1. Stratified False Discovery Rate (FDR)
    - Finding: per-stratum local FDR maximizes power controlling overall FDR
    - Improved power in protein domain prediction
    - Identified protein classes with problematic statistics
2. Genetic and other factors controlling the placebo response
    - Collaboration with Otsuka Pharmaceutical
    - Mixed-effects modeling of longitudinal response in drug trials for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder
    - Genome-wide association study for individual-specific placebo response
3. **Kinship and F_{ST} for arbitrary population structures**
    - Motivation: world-wide human population structure
    - Generalized definitions and models
    - Novel bias calculations validated by simulations
    - Novel non-parametric estimator with greatly improved accuracy

Why study relatedness?

Human genetics
is fascinating!

Why study relatedness?

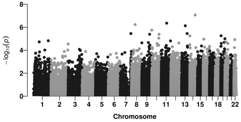
Human genetics is fascinating!



Pop. structure confounds association studies (GWAS)

Why study relatedness?

Human genetics is fascinating!

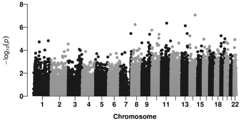


Pop. structure confounds association studies (GWAS)

Heritability of complex traits

Why study relatedness?

Human genetics is fascinating!

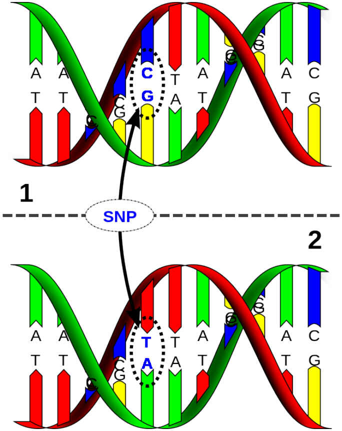


Pop. structure confounds association studies (GWAS)

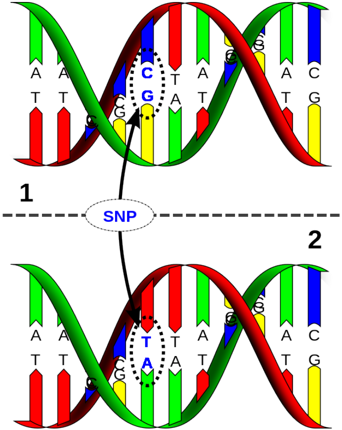
Heritability of complex traits

Animal and plant breeding

Single Nucleotide Polymorphism (SNP) data



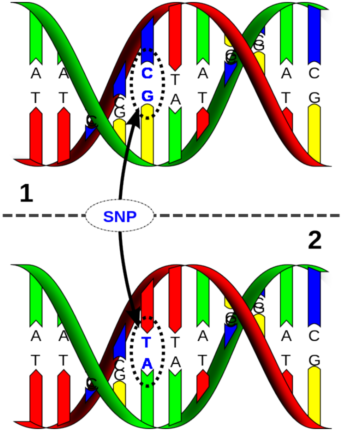
Single Nucleotide Polymorphism (SNP) data



⇒

Genotype	x_{ij}
CC	0
CT	1
TT	2

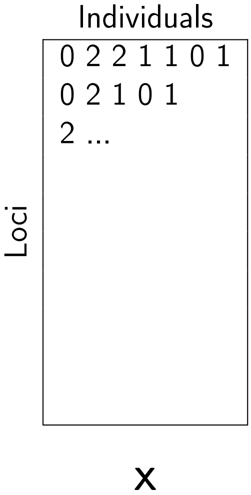
Single Nucleotide Polymorphism (SNP) data



⇒

Genotype	x_{ij}
CC	0
CT	1
TT	2

⇒



0 2 2 1 1 0 1
0 2 1 0 1
2 ...

$\mathbf{X}$

Hardy-Weinberg Equillibrium (HWE): Binomial draws

x_{ij} = genotype at locus i for individual j.

p_i^T = frequency of reference allele at locus i, (ancestral) population T.

Hardy-Weinberg Equillibrium (HWE): Binomial draws

x_{ij} = genotype at locus i for individual j.

p_i^T = frequency of reference allele at locus i, (ancestral) population T.

Under HWE:

$$\begin{aligned}
\Pr(x_{ij} = 2 | p_i^T) &= \left(p_i^T\right)^2, \\
\Pr(x_{ij} = 1 | p_i^T) &= 2p_i^T \left(1 - p_i^T\right), \\
\Pr(x_{ij} = 0 | p_i^T) &= \left(1 - p_i^T\right)^2.
\end{aligned}$$

Hardy-Weinberg Equillibrium (HWE): Binomial draws

x_{ij} = genotype at locus i for individual j.

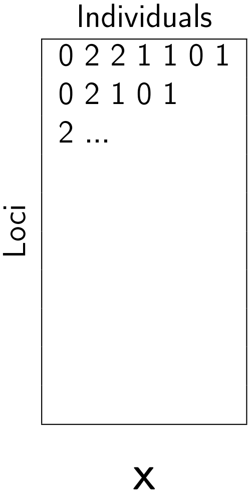
p_i^T = frequency of reference allele at locus i, (ancestral) population T.

Under HWE:

$$
\begin{aligned}
\Pr(x_{ij} = 2 | p_i^T) &= \left(p_i^T\right)^2 , \\
\Pr(x_{ij} = 1 | p_i^T) &= 2 p_i^T \left(1 - p_i^T\right) , \\
\Pr(x_{ij} = 0 | p_i^T) &= \left(1 - p_i^T\right)^2 .
\end{aligned}
$$

HWE not valid under population structure!

Goal: measure dependence structure of genotype matrix columns



Individuals

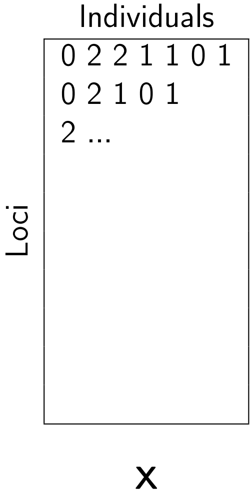
Loci

0 2 2 1 1 0 1
0 2 1 0 1
2 ...

X

High-dimensional binomial data

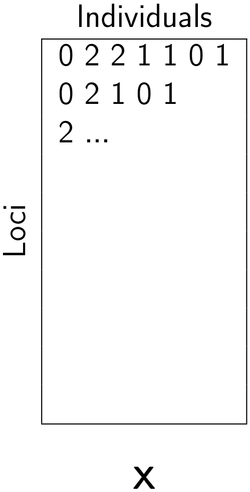
Goal: measure dependence structure of genotype matrix columns



High-dimensional binomial data

Population structure
⇒ dependence between individuals (columns)

Goal: measure dependence structure of genotype matrix columns



High-dimensional binomial data

Population structure
⇒ dependence between individuals (columns)

Linkage disequilibrium
⇒ dependence between loci (rows)

Model parameters

IBD(T): “Identical By Descent” for ancestral population T — shared coin flips

Model parameters

IBD(T): “Identical By Descent” for ancestral population T — shared coin flips

f_j^T: **Inbreeding coefficient**
Pr. that the two alleles at a random locus of individual j are IBD(T)

$$\mathrm{Var}(x_{ij}|T) = 2p_i^T\left(1 - p_i^T\right)\left(1 + f_j^T\right)$$

Model parameters

IBD(T): "Identical By Descent" for ancestral population T — shared coin flips

f_j^T: **Inbreeding coefficient**
Pr. that the two alleles at a random locus of individual j are IBD(T)

$$\mathrm{Var}(x_{ij}|T) = 2p_i^T \left(1 - p_i^T\right)\left(1 + f_j^T\right)$$

φ_{jk}^T: **Kinship coefficient**
Pr. that two alleles, one at random from each of individuals j and k, at one random locus are IBD(T)

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

Model parameters

IBD(T): "Identical By Descent" for ancestral population T — shared coin flips

f_j^T: **Inbreeding coefficient**
Pr. that the two alleles at a random locus of individual j are IBD(T)

$$\mathrm{Var}(x_{ij} | T) = 2p_i^T \left(1 - p_i^T\right) \left(1 + f_j^T\right)$$

φ_{jk}^T: **Kinship coefficient**
Pr. that two alleles, one at random from each of individuals j and k, at one random locus are IBD(T)

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

F_{ST}: **Fixation index**
Pr. that two random alleles in a subpopulation at a random locus are IBD(T)

Existing approaches

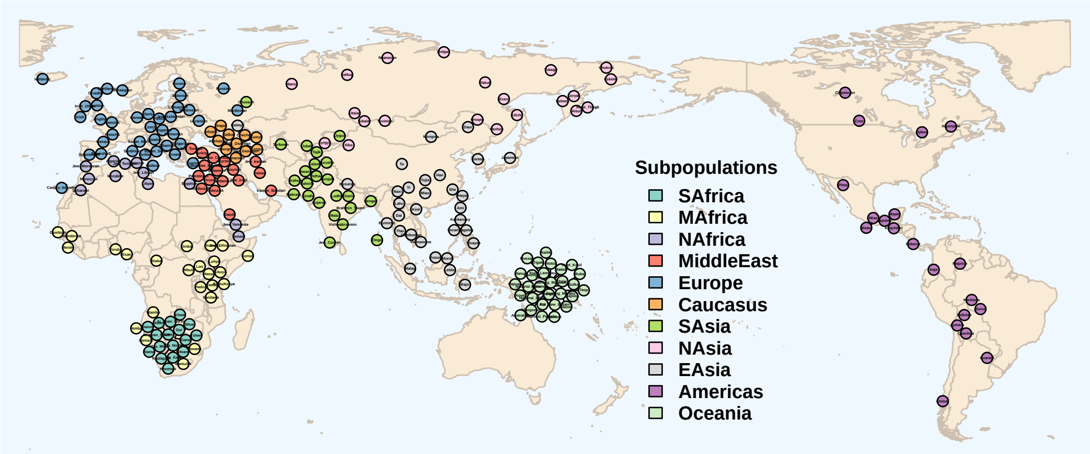
1. F_{ST} estimation
    - *For independent subpopulations only!*
    - Weir-Cockerham (WC) estimator (1984) — 15K citations!
    - "Hudson" pairwise estimator (2013) tweaks WC
    - BayeScan (2008) — 1.2K citations

Existing approaches

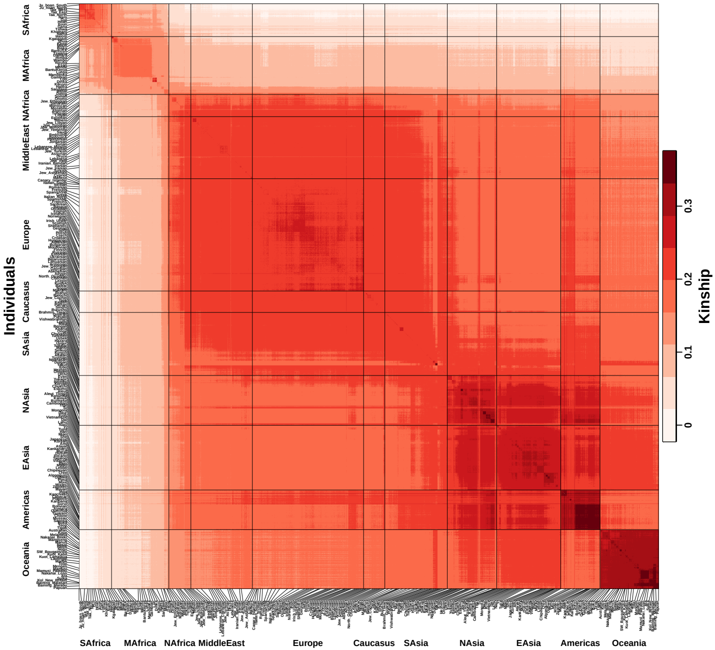
1. F_{ST} estimation
   - *For independent subpopulations only!*
   - Weir-Cockerham (WC) estimator (1984) — 15K citations!
   - "Hudson" pairwise estimator (2013) tweaks WC
   - BayeScan (2008) — 1.2K citations

2. Kinship estimation
   - "Standard" kinship estimator (1950s)
     - Used by most modern GWAS approaches that control for population structure (PCA, LMM, adj. χ^2; top paper 6K citations)
     - GCTA heritability estimation (2 papers: 4K citations)
   - Our novel finding: accuracy requires unstructured population (a minority of closely-related individuals)

Dataset: Human Origins

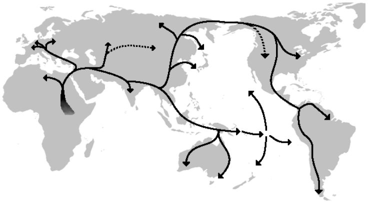


2,922 indivs. from 244 locs. — 593,124 loci — SNP chip
Lazaridis *et al.* (2014), (2016); Skoglund *et al.* (2016)



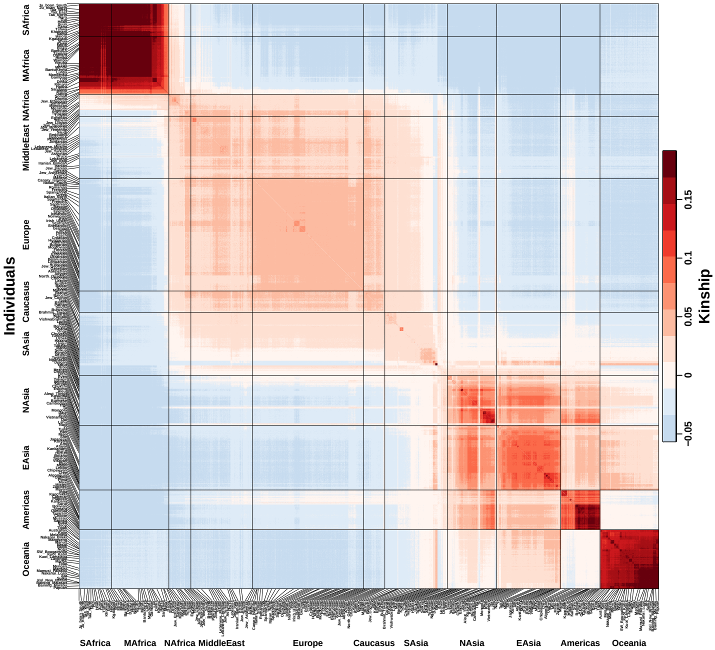
Our new kinship estimates

Genotypes from "Human Origins" (Lazaridis *et al.* 2014, 2016; Skoglund *et al.* 2016)



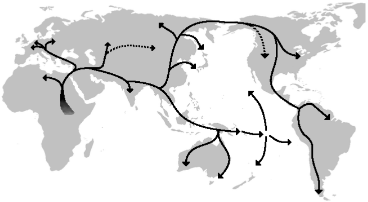
Edited from Ephert [CC BY-SA 3.0], via Wikimedia Commons

*Inbreeding coeffs. on diagonal



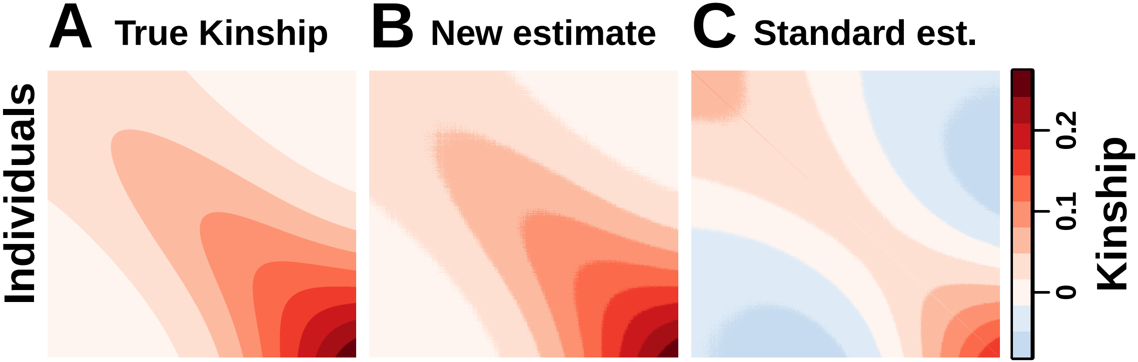
Standard kinship estimates

Genotypes from "Human Origins" (Lazaridis *et al.* 2014, 2016; Skoglund *et al.* 2016)

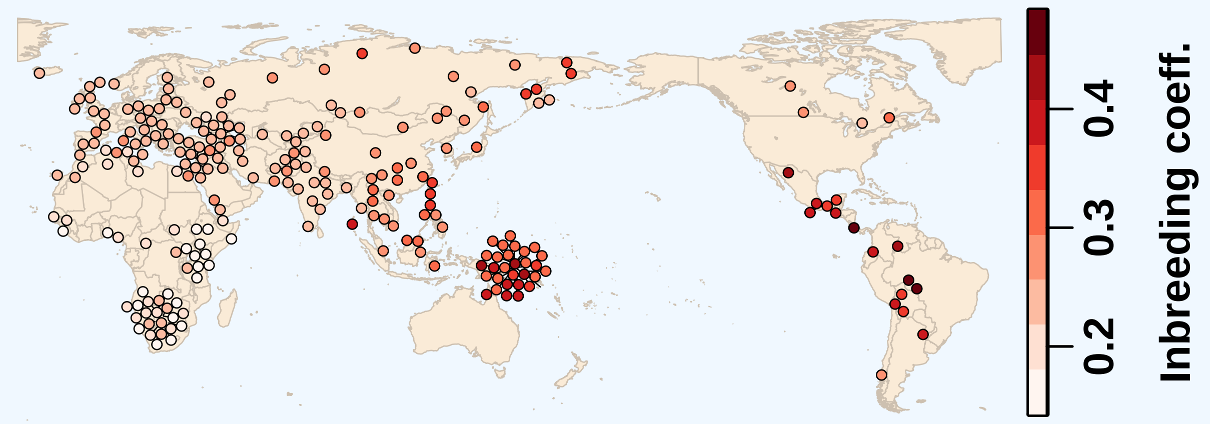


Edited from Ephert [CC BY-SA 3.0], via Wikimedia Commons

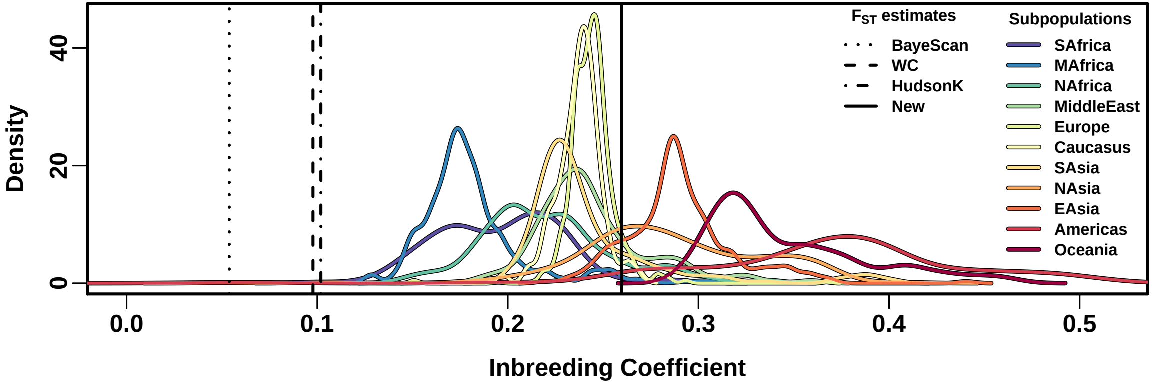
Only our new estimator is accurate in simulations



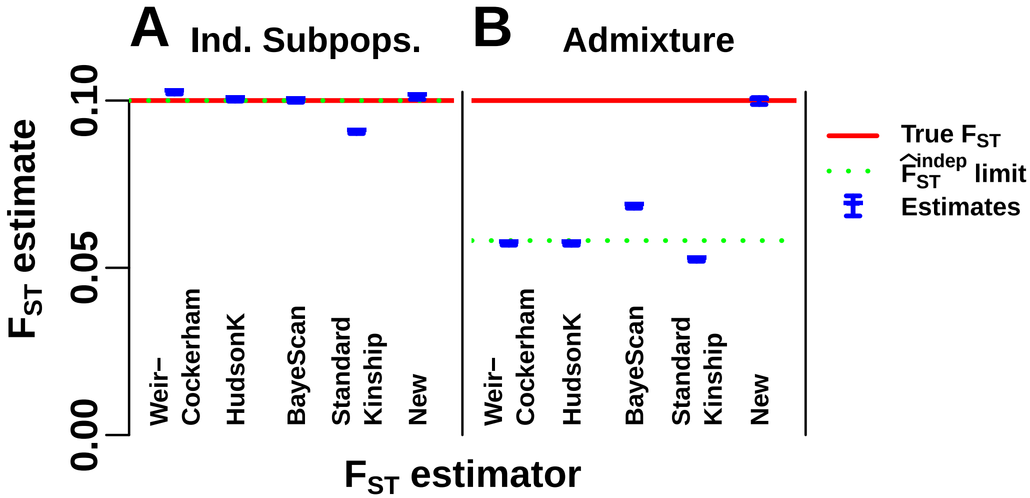
Population-level inbreeding increases with distance from Africa



Differentiation (F_{ST}) previously underestimated

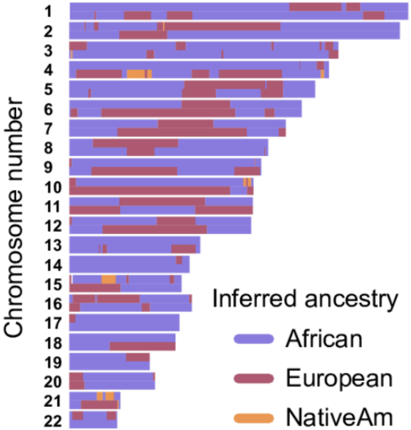


Only our new method estimates generalized F_{ST} accurately



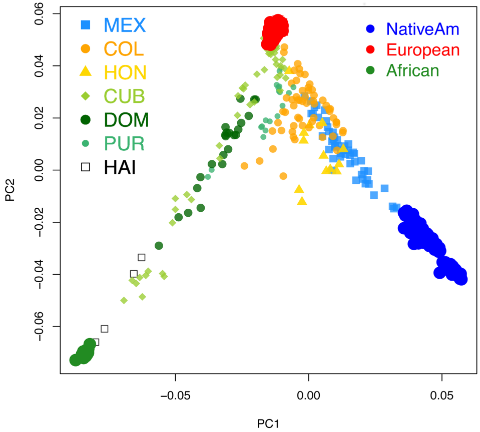
Recently-admixed populations

African-Americans



Baharian *et al.* (2016)

Hispanics



Moreno-Estrada *et al.* (2013)

Admixed siblings from different populations?

Lucy and Maria, UK

Admixed siblings from different populations?

Lucy and Maria, UK

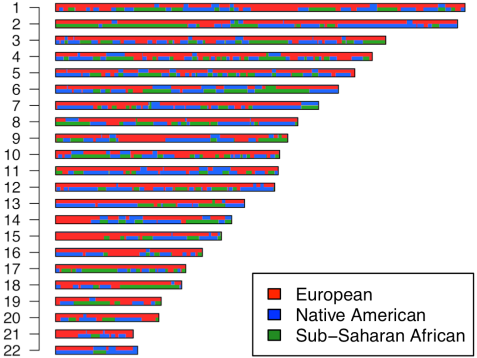
Ochoa brothers, MX

Admixed siblings from different populations?

Lucy and Maria, UK

Ochoa brothers, MX

High Admixture LD:



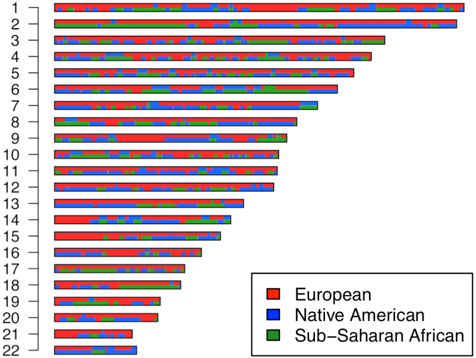
Moreno-Estrada *et al.* (2013)

Admixed siblings from different populations?

Lucy and Maria, UK

Ochoa brothers, MX

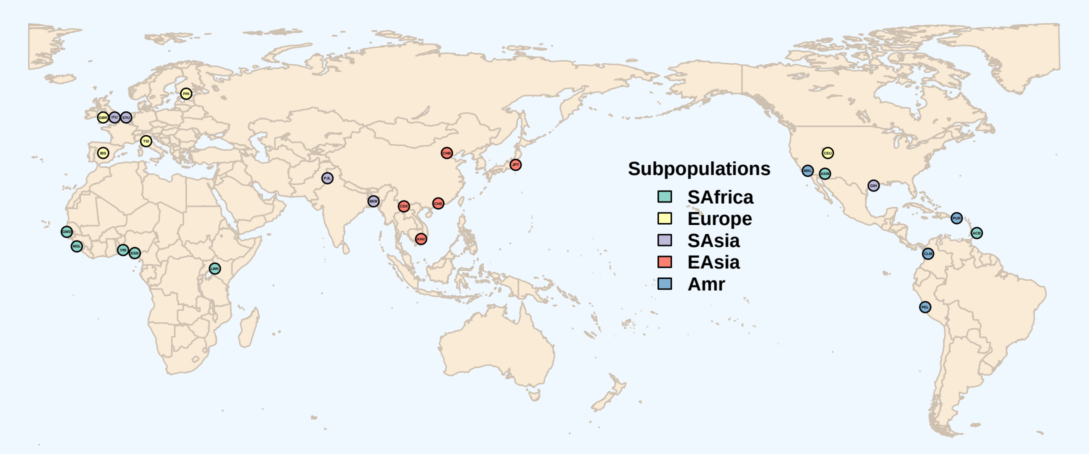
High Admixture LD:



Moreno-Estrada *et al.* (2013)

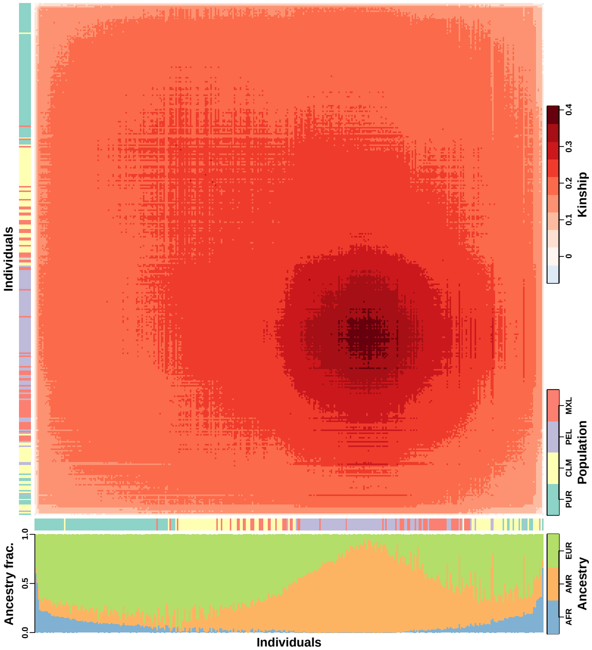
Solution: treat every individual as its own population!

Dataset: 1000 Genomes Project (2013)



2,504 indivs. from 26 locs. — 20,417,698 loci (asc. in YRI) — WGS trios, etc.

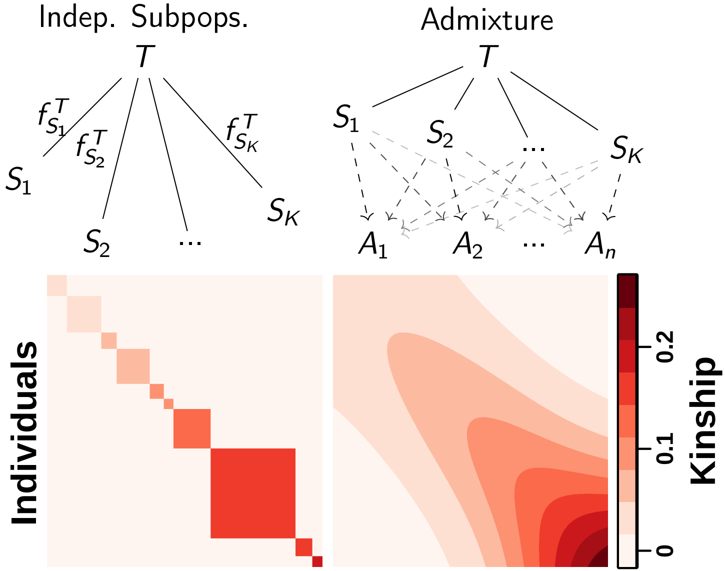
Kinship driven by admixture in Hispanics



Our new kinship estimates

Genotypes from the 1000 Genomes Project (2013)

Comparison of population structures in simulation



F_{ST} in the independent subpopulation model

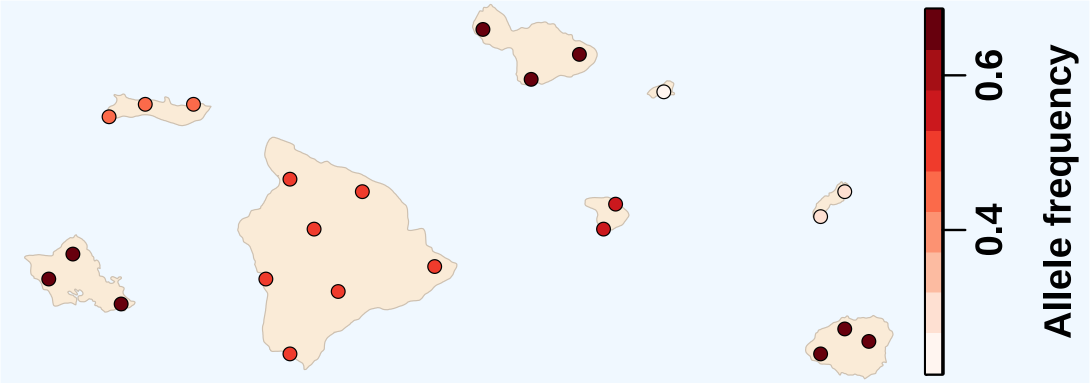


Illustration.

F_{ST} in the independent subpopulation model

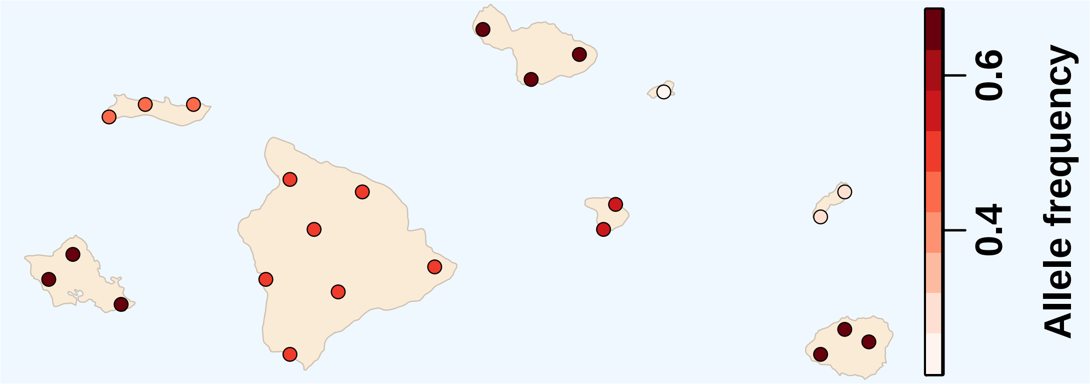


Illustration.

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

Here F_{ST} = proportion of variance explained by pop. structure

Wright's F_{ST}

T = Total, S = Subpopulation, I = Individual.

Total inbreeding: $$F_{\mathrm{IT}} = \frac{1}{|S|} \sum_{j \in S} f_j^T,$$

Local inbreeding: $$F_{\mathrm{IS}} = \frac{1}{|S|} \sum_{j \in S} f_j^S,$$

Structural inbreeding: $$F_{\mathrm{ST}} = \frac{F_{\mathrm{IT}} - F_{\mathrm{IS}}}{1 - F_{\mathrm{IS}}}.$$

Our generalized F_{ST}

Need new “local” subpopulations L_j (separates total from local inbreeding):

$$\left(1 - f_j^T\right) = \left(1 - f_j^{L_j}\right)\left(1 - f_{L_j}^T\right).$$

Our generalized F_{ST}

Need new "local" subpopulations L_j (separates total from local inbreeding):

$$\left(1 - f_j^T\right) = \left(1 - f_j^{L_j}\right)\left(1 - f_{L_j}^T\right).$$

Generalized F_{ST}: applicable to arbitrary population structures, equals previous definition for non-overlapping subpopulations:

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

Our generalized F_{ST}

Need new "local" subpopulations L_j (separates total from local inbreeding):

$$\left(1 - f_j^T\right) = \left(1 - f_j^{L_j}\right)\left(1 - f_{L_j}^T\right).$$

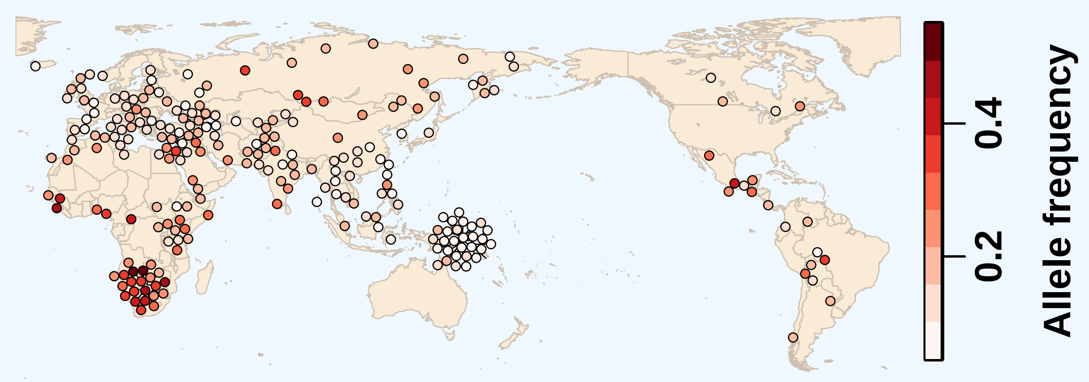
Generalized F_{ST}: applicable to arbitrary population structures, equals previous definition for non-overlapping subpopulations:

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

Mean heterozygosity in a structured population:

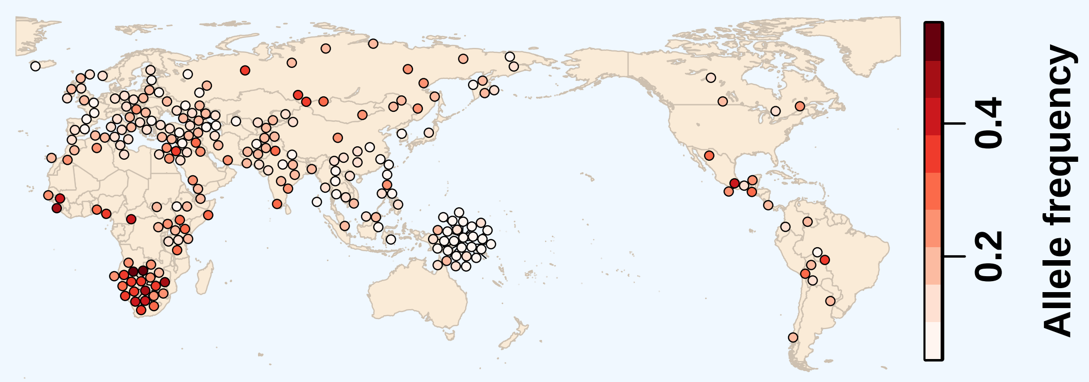
$$\bar{H}_i = \frac{1}{n} \sum_{j=1}^{n} \Pr(x_{ij} = 1 | T) = 2 p_i^T \left(1 - p_i^T\right)\left(1 - F_{\mathrm{ST}}\right).$$

F_{ST} measures population structure / differentiation



Median diff. SNP in Human Origins (rs2650044; given MAF $\geq$ 10%).

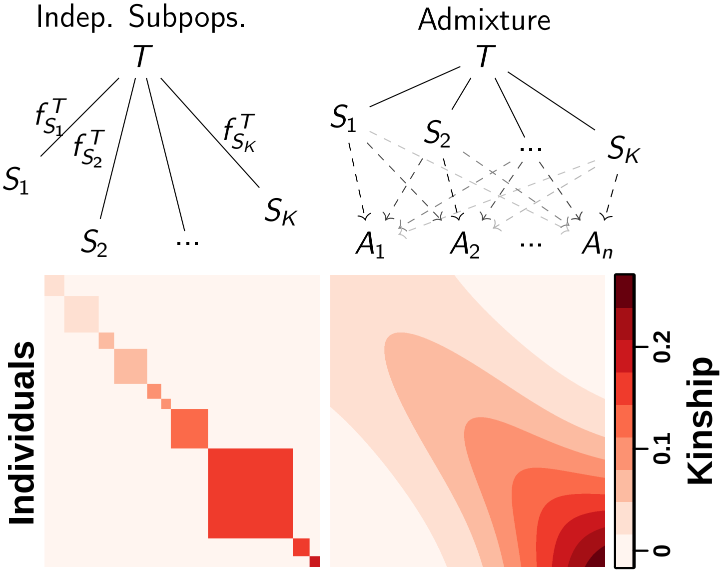
F_{ST} measures population structure / differentiation



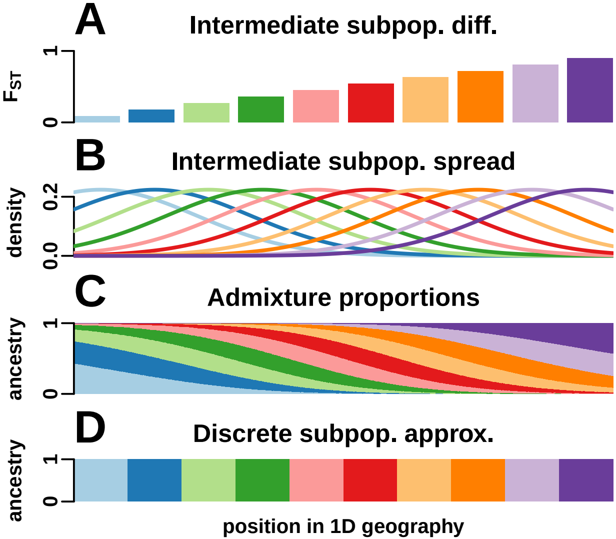
Median diff. SNP in Human Origins (rs2650044; given MAF $\geq$ 10%).

$\hat{F}_{\mathrm{ST}}^{\mathrm{WC}} \approx 0.0961$ using Weir-Cockerham estimator and $K = 244$.

Comparison of population structures in simulation



Our admixture simulation (R package ‘bnpsd’ on CRAN)



Kinship model for genotypes

symbol	meaning
T	ref ancestral population
i	locus index
j, k	individual indexes
p_i^T	ref allele frequency
x_{ij}	genotype (num ref alleles)
φ_{jk}^T	kinship of j, k
f_j^T	inbreeding of j

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

Statistical model:

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

Problem: common estimators not consistent under structure

Estimate of ancestral allele frequency:

$$\hat{p}_i^T = \frac{1}{2} \sum_{j=1}^{n} w_j x_{ij}$$

Variance asymptotically non-zero under population structure:

$$\mathrm{Var}\left(\hat{p}_i^T \middle| T \right) = p_i^T \left(1 - p_i^T \right) \bar{\varphi}^T$$

Problem: common estimators not consistent under structure

Estimate of ancestral allele frequency:

$$\hat{p}_i^T = \frac{1}{2} \sum_{j=1}^{n} w_j x_{ij}$$

Variance asymptotically non-zero under population structure:

$$\mathrm{Var}\left(\hat{p}_i^T \middle| T \right) = p_i^T \left(1 - p_i^T \right) \bar{\varphi}^T$$

(for independent individuals $\bar{\varphi}^T = \frac{1}{2n}(1 + F_{\mathrm{ST}})$. $\Rightarrow n_{\mathrm{Eff}} \approx 4$ in Human Origins!)

Problem: common estimators not consistent under structure

Estimate of ancestral allele frequency:

$$\hat{p}_i^T = \frac{1}{2} \sum_{j=1}^{n} w_j x_{ij}$$

Variance asymptotically non-zero under population structure:

$$\mathrm{Var}\left(\hat{p}_i^T \middle| T\right) = p_i^T \left(1 - p_i^T\right) \bar{\varphi}^T$$

(for independent individuals $\bar{\varphi}^T = \frac{1}{2n}(1 + F_{\mathrm{ST}})$. $\Rightarrow n_{\mathrm{Eff}} \approx 4$ in Human Origins!)

Therefore, naive estimators that use $\hat{p}_i^T$ (next) are not consistent!

Bias in standard kinship estimator

$$\hat{\varphi}_{jk}^{T,\text{std}} = \frac{\sum_{i=1}^{m} \left(x_{ij} - 2\hat{p}_i^T \right) \left(x_{ik} - 2\hat{p}_i^T \right)}{4 \sum_{i=1}^{m} \hat{p}_i^T \left(1 - \hat{p}_i^T \right)}, \qquad \hat{p}_i^T = \frac{1}{2} \sum_{j=1}^{n} w_j x_{ij}.$$

Bias in standard kinship estimator

$$\hat{\varphi}_{jk}^{T,\text{std}} = \frac{\sum_{i=1}^{m} \left(x_{ij} - 2\hat{p}_i^T\right)\left(x_{ik} - 2\hat{p}_i^T\right)}{4\sum_{i=1}^{m} \hat{p}_i^T \left(1 - \hat{p}_i^T\right)}, \qquad \hat{p}_i^T = \frac{1}{2}\sum_{j=1}^{n} w_j x_{ij}.$$

Bias varies by j, k:

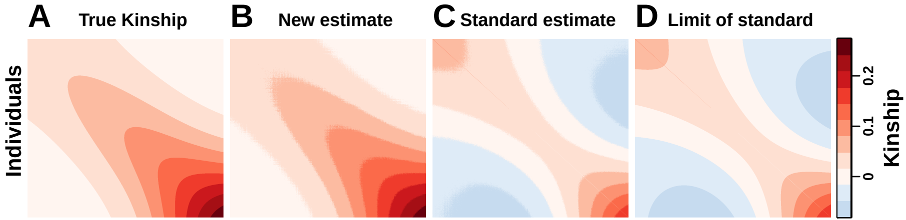
$$\hat{\varphi}_{jk}^{T,\text{std}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \bar{\varphi}_j^T - \bar{\varphi}_k^T + \bar{\varphi}^T}{1 - \bar{\varphi}^T}.$$

Bias in standard kinship estimator

$$\hat{\varphi}_{jk}^{T,\text{std}} = \frac{\sum_{i=1}^{m} \left(x_{ij} - 2\hat{p}_i^T \right) \left(x_{ik} - 2\hat{p}_i^T \right)}{4 \sum_{i=1}^{m} \hat{p}_i^T \left(1 - \hat{p}_i^T \right)}, \qquad \hat{p}_i^T = \frac{1}{2} \sum_{j=1}^{n} w_j x_{ij}.$$

Bias varies by j, k:

$$\hat{\varphi}_{jk}^{T,\text{std}} \xrightarrow[m \to \infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \bar{\varphi}_j^T - \bar{\varphi}_k^T + \bar{\varphi}^T}{1 - \bar{\varphi}^T}.$$



Our new estimator (R package ‘popkin’ on CRAN)

Step 1: “pre-adjusted” kinship estimator with uniform bias.

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

Our new estimator (R package ‘popkin’ on CRAN)

Step 1: “pre-adjusted” kinship estimator with uniform bias.

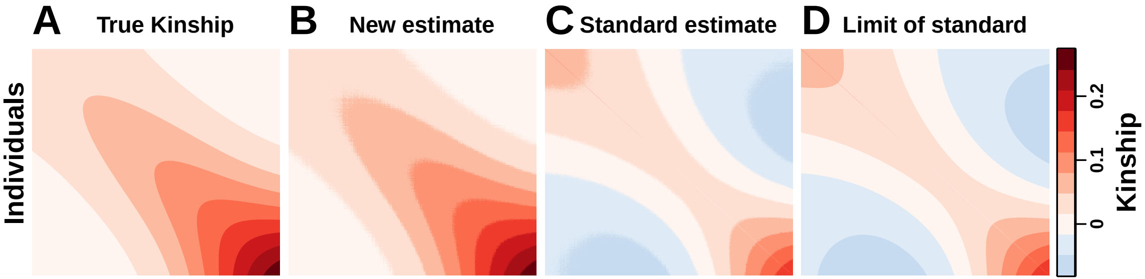
$$\hat{\varphi}_{jk}^{T,\text{preadj}} = \frac{\sum_{i=1}^{m} (x_{ij} - 1)(x_{ik} - 1) - 1}{4 \sum_{i=1}^{m} \hat{p}_i^T \left(1 - \hat{p}_i^T \right)} + 1 \xrightarrow[m \to \infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \bar{\varphi}^T}{1 - \bar{\varphi}^T},$$

Step 2: Estimate minimum kinship, use to unbias “step 1” estimates.

$$\hat{\varphi}_{\text{min}}^{T,\text{preadj}} \xrightarrow[m \to \infty]{\text{a.s.}} -\frac{\bar{\varphi}^T}{1 - \bar{\varphi}^T}, \qquad \hat{\varphi}_{jk}^{T,\text{new}} = \frac{\hat{\varphi}_{jk}^{T,\text{preadj}} - \hat{\varphi}_{\text{min}}^{T,\text{preadj}}}{1 - \hat{\varphi}_{\text{min}}^{T,\text{preadj}}} \xrightarrow[m \to \infty]{\text{a.s.}} \varphi_{jk}^T.$$

This yields consistent $\hat{f}_j^{T,\text{new}}$, $\hat{F}_{\text{ST}}^{\text{new}}$ estimators!

Performance of new estimator



Bias in F_{ST} estimators for independent subpopulations

Previous estimator for n subpopulations, simplified for known AFs (π_{ij}):

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

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

Bias in F_{ST} estimators for independent subpopulations

Previous estimator for n subpopulations, simplified for known AFs (π_{ij}):

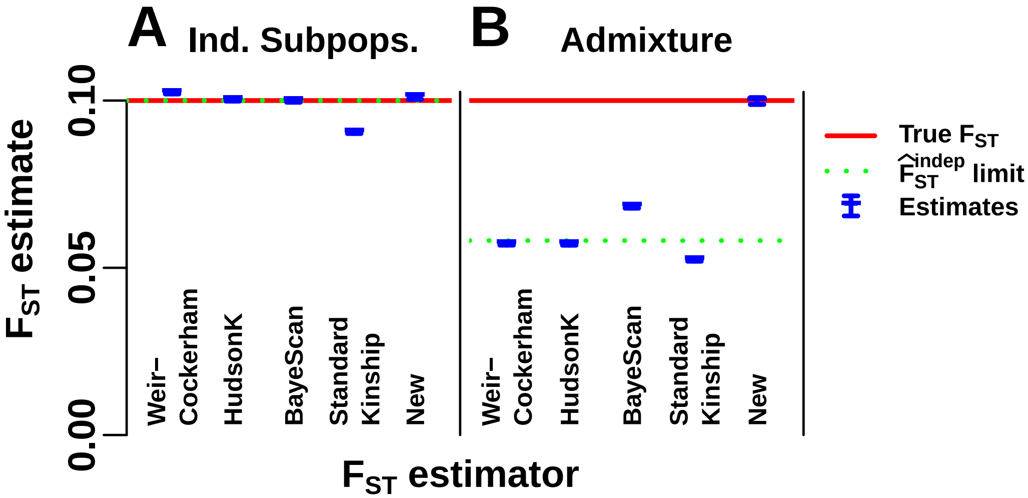
$$\hat{F}_{\text{ST}}^{\text{indep}} = \frac{\sum_{i=1}^{m} \hat{\sigma}_i^2}{\sum_{i=1}^{m} \hat{p}_i^T \left(1 - \hat{p}_i^T\right) + \frac{1}{n}\hat{\sigma}_i^2},$$

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

Estimator is biased in dependent subpopulations:

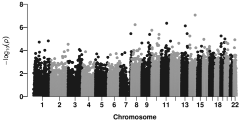
$$\hat{F}_{\text{ST}}^{\text{indep}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{\text{ST}} - \frac{1}{n-1}\left(n\bar{\theta}^T - F_{\text{ST}}\right)}{1 - \frac{1}{n-1}\left(n\bar{\theta}^T - F_{\text{ST}}\right)}.$$

Only our new method estimates generalized F_{ST} accurately



The future: improved kinship has repercussions across genetics!

Accurate and efficient estimation, admixture modeling



Association studies, selection tests

Bias in heritability of complex traits

Animal and plant breeding

Acknowledgments

John D. Storey
Andrew Bass
Irineo Cabreros
Wei Hao
Riley Skeen-Gaar

Neo Christopher Chung
University of Warsaw

Funding:
National Institutes of Health
Otsuka Pharmaceutical

PRINCETON UNIVERSITY

Lewis-Sigler Institute for Integrative Genomics