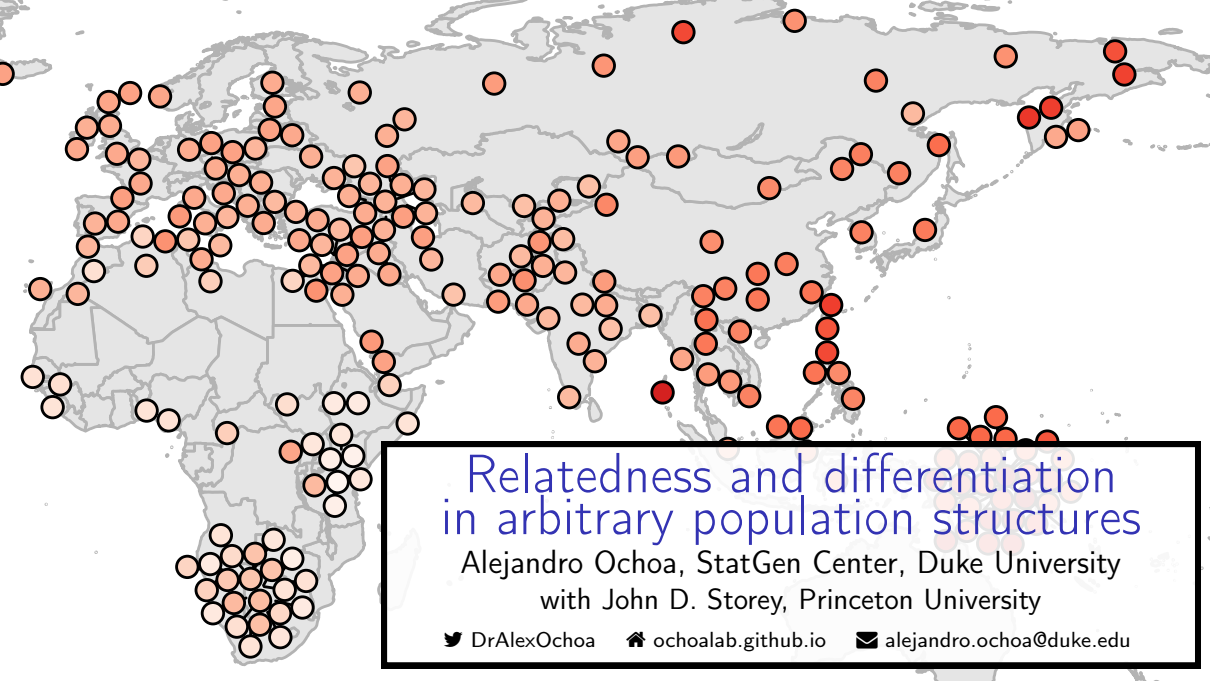


Relatedness and differentiation in arbitrary population structures

Alejandro Ochoa, StatGen Center, Duke University
with John D. Storey, Princeton University

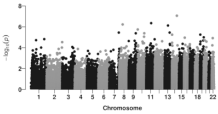
DrAlexOchoa ochoalab.github.io alejandro.ochoa@duke.edu

Why study relatedness?

Human genetics is fascinating!

Why study relatedness?

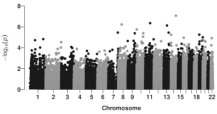
Human genetics is fascinating!



Genetic Association Studies confounded by relatedness

Why study relatedness?

Human genetics is fascinating!

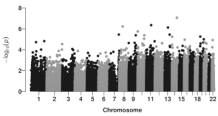


Genetic Association Studies confounded by relatedness

Heritability of complex traits

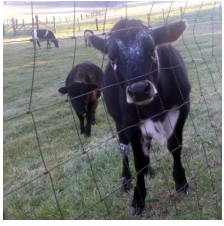
Why study relatedness?

Human genetics is fascinating!



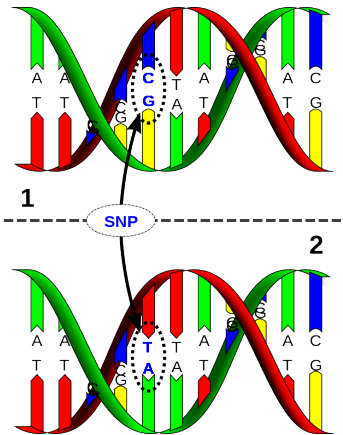
Genetic Association Studies confounded by relatedness

Heritability of complex traits

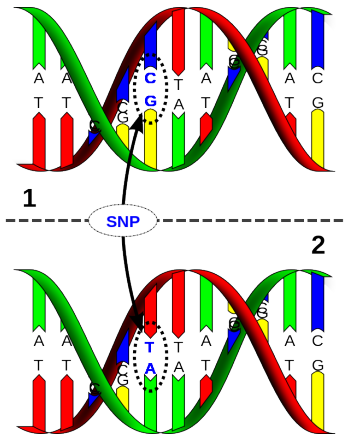


Animal and plant breeding

Single Nucleotide Polymorphism (SNP) data



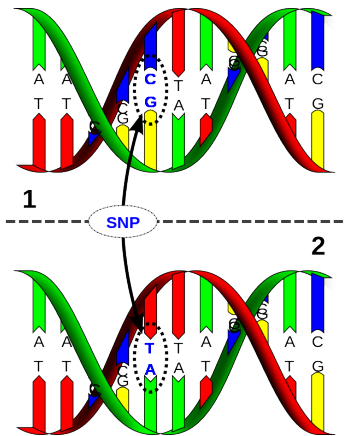
Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

Genotype	x_{ij}
CC	0
CT	1
TT	2

Single Nucleotide Polymorphism (SNP) data



$\Rightarrow$

Genotype	x_{ij}
CC	0
CT	1
TT	2

$\Rightarrow$

Loci

Individuals

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

X

Hardy-Weinberg Equillibrium (HWE): Binomial draws

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

p_i = frequency of reference allele at locus i.

Hardy-Weinberg Equillibrium (HWE): Binomial draws

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

p_i = frequency of reference allele at locus i.

Under HWE:

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

Hardy-Weinberg Equillibrium (HWE): Binomial draws

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

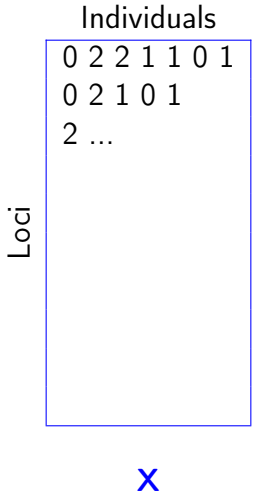
p_i = frequency of reference allele at locus i.

Under HWE:

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

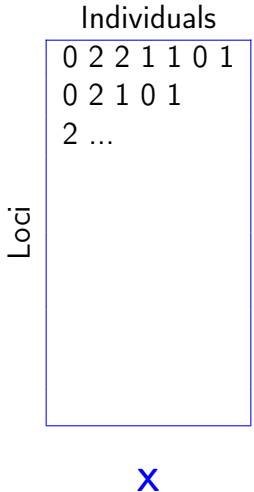
HWE not valid under population structure!

Goal: measure dependence structure of genotype matrix columns



High-dimensional binomial data

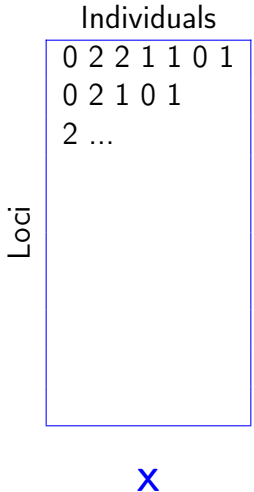
Goal: measure dependence structure of genotype matrix columns



High-dimensional binomial data

Relatedness / Population structure
$\Rightarrow$ dependence between individuals (columns)

Goal: measure dependence structure of genotype matrix columns

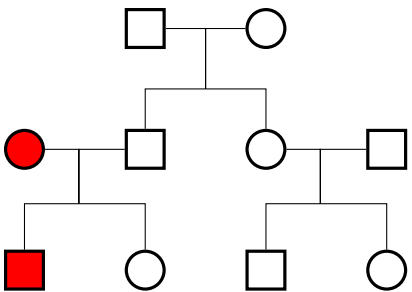


High-dimensional binomial data

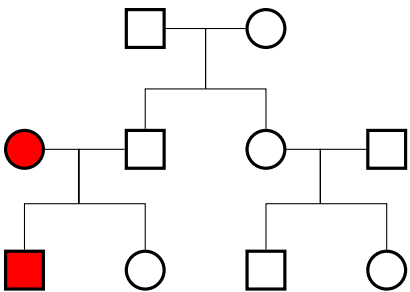
Relatedness / Population structure
$\Rightarrow$ dependence between individuals (columns)

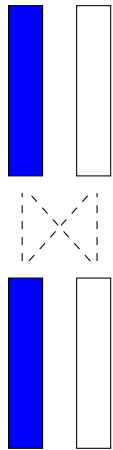
Linkage disequilibrium
$\Rightarrow$ dependence between loci (rows)

The kinship coefficient for parent-child: $\frac{1}{4}$

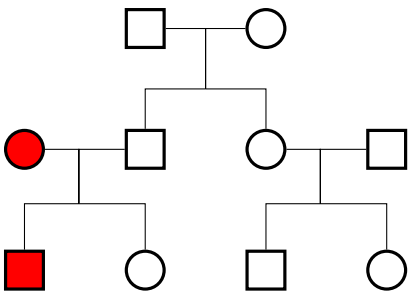


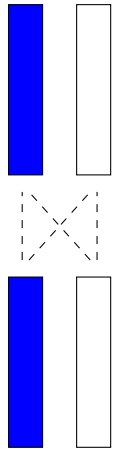
The kinship coefficient for parent-child: $\frac{1}{4}$

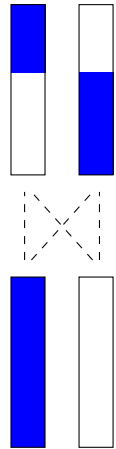




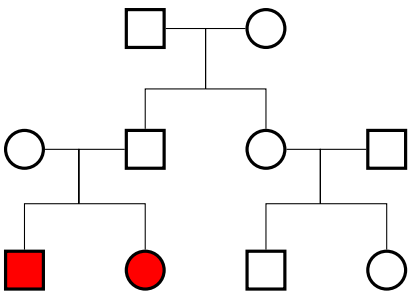
The kinship coefficient for parent-child: $\frac{1}{4}$



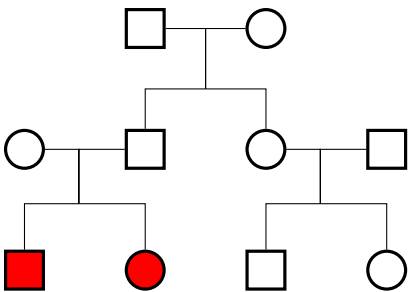


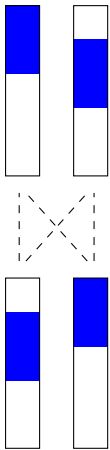


The kinship coefficient for siblings: $\frac{1}{4}$ on average

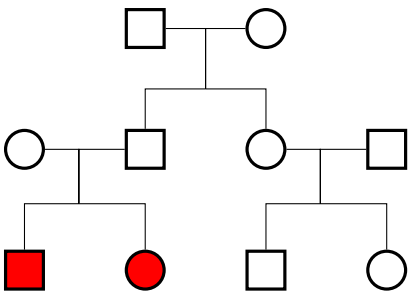


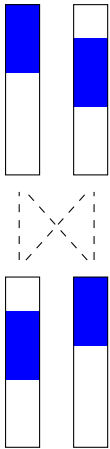
The kinship coefficient for siblings: $\frac{1}{4}$ on average

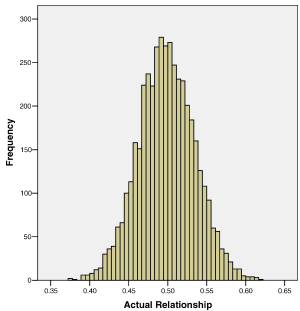




The kinship coefficient for siblings: $\frac{1}{4}$ on average

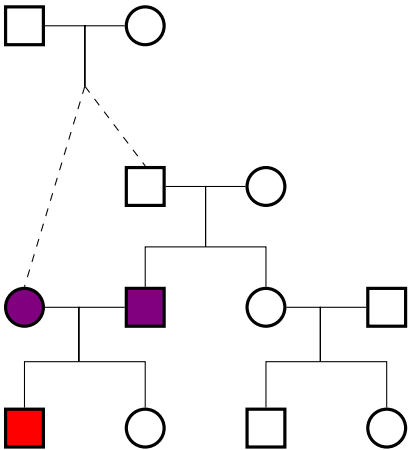




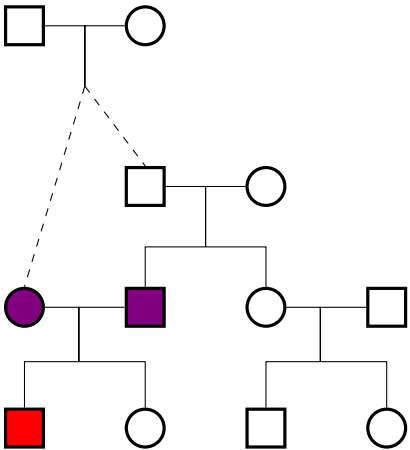


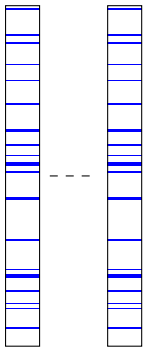
Visscher *et al.* (2006)

The inbreeding coefficient in populations

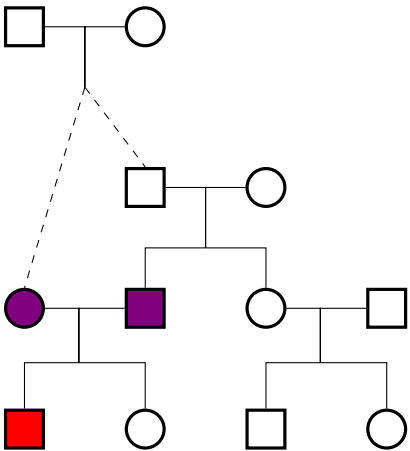


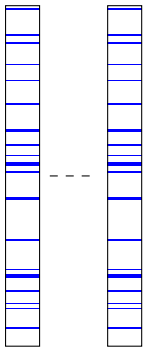
The inbreeding coefficient in populations





The inbreeding coefficient in populations





Measurements relative to a reference pop.:

Inbreeding $= 0$ in the local population

Inbreeding ≥ 0 relative to a distant ancestral population

Better measured using covariance

Model parameters

IBD: "Identical By Descent" (given implicit ancestral pop. T) — shared coin flips

Model parameters

IBD: "Identical By Descent" (given implicit ancestral pop. T) — shared coin flips

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

$$\mathrm{Var}(x_{ij}) = 2p_i\,(1 - p_i)\,(1 + f_j)$$

Model parameters

IBD: "Identical By Descent" (given implicit ancestral pop. T) — shared coin flips

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

$$\mathrm{Var}(x_{ij}) = 2p_i\,(1 - p_i)\,(1 + f_j)$$

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

$$\mathrm{Cov}(x_{ij}, x_{ik}) = 4p_i\,(1 - p_i)\,\varphi_{jk}$$

Model parameters

IBD: "Identical By Descent" (given implicit ancestral pop. T) — shared coin flips

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

$$\mathrm{Var}(x_{ij}) = 2p_i\,(1 - p_i)\,(1 + f_j)$$

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

$$\mathrm{Cov}(x_{ij}, x_{ik}) = 4p_i\,(1 - p_i)\,\varphi_{jk}$$

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

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 genetic association 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)

Theoretical results: **new kinship estimator!**

$x_{ij} \in \{0, 1, 2\}$: Genotype at locus i of individual j.

Theoretical results: **new kinship estimator!**

$x_{ij} \in \{0, 1, 2\}$: Genotype at locus i of individual j. Model:

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

Theoretical results: **new kinship estimator!**

$x_{ij} \in \{0, 1, 2\}$: Genotype at locus i of individual j. Model:

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

Standard estimator is **biased**:

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

Theoretical results: **new kinship estimator!**

$x_{ij} \in \{0, 1, 2\}$: Genotype at locus i of individual j. Model:

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

Standard estimator is **biased**:

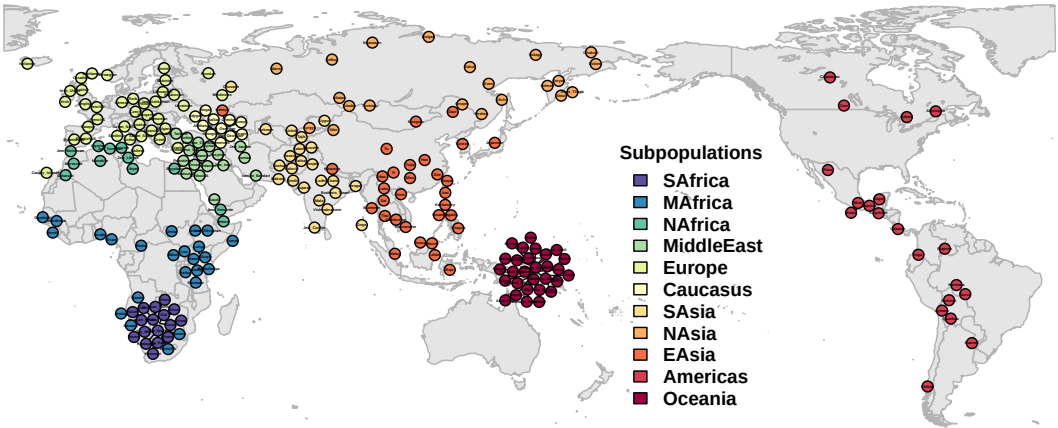
$$\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}, \qquad \hat{\varphi}_{jk}^{\text{std}} = \frac{\sum_{i=1}^{m}\left(x_{ij} - 2\hat{p}_i\right)\left(x_{ik} - 2\hat{p}_i\right)}{4\sum_{i=1}^{m}\hat{p}_i\,(1 - \hat{p}_i)} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k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

`popkin`: first unbiased kinship estimator! — R package on CRAN

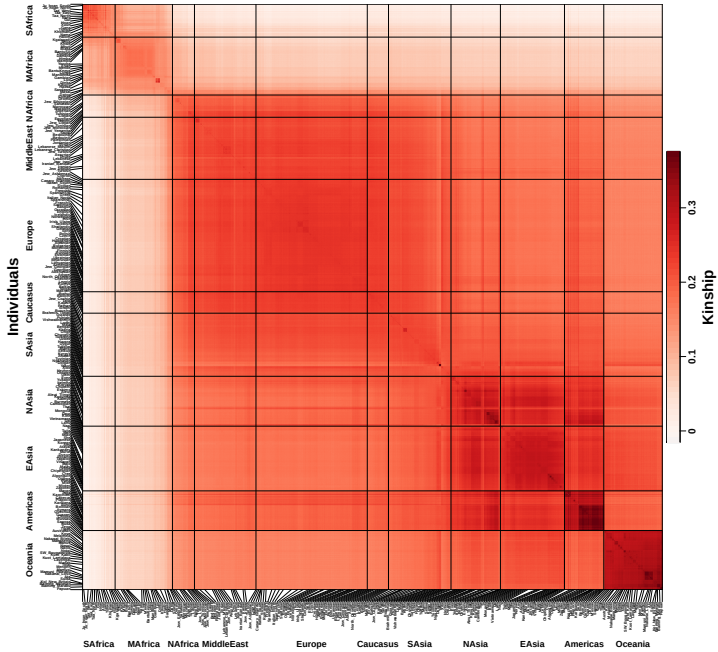
$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - 1)(x_{ik} - 1) - 1, \qquad A_{\min} = \min_{u \neq v} \frac{1}{|S_u||S_v|}\sum_{j \in S_u}\sum_{k \in S_v} A_{jk},$$

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

Dataset: Human Origins

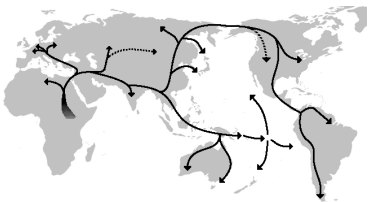


2,922 indivs. from 244 locs. — 593,124 loci — SNP chip
Lazaridis *et al.* (2014), (2016); Skoglund *et al.* (2016)



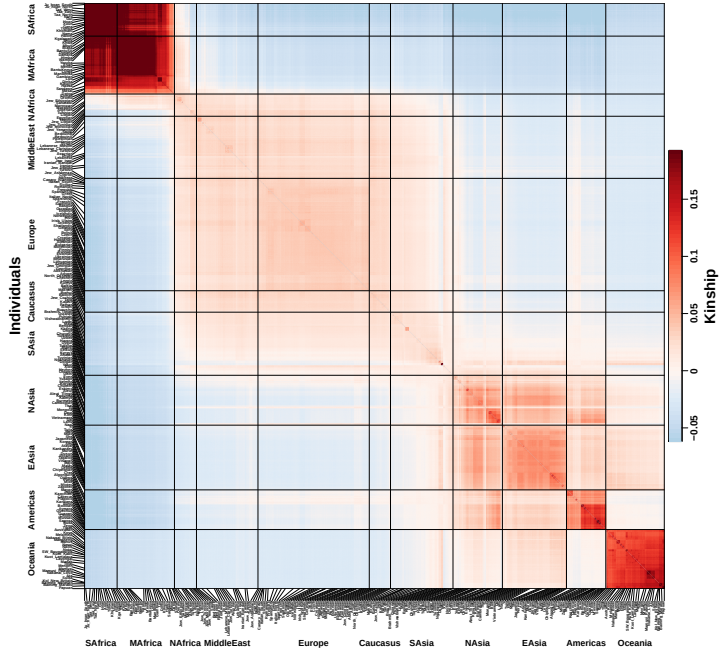
Our new kinship estimates

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



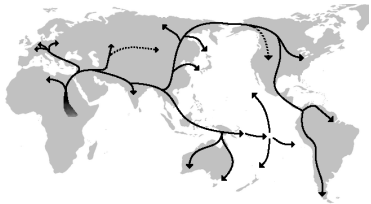
Edited from Ephert [CC BY-SA 3.0], via Wikimedia Commons

*Inbreeding coeffs. on diagonal



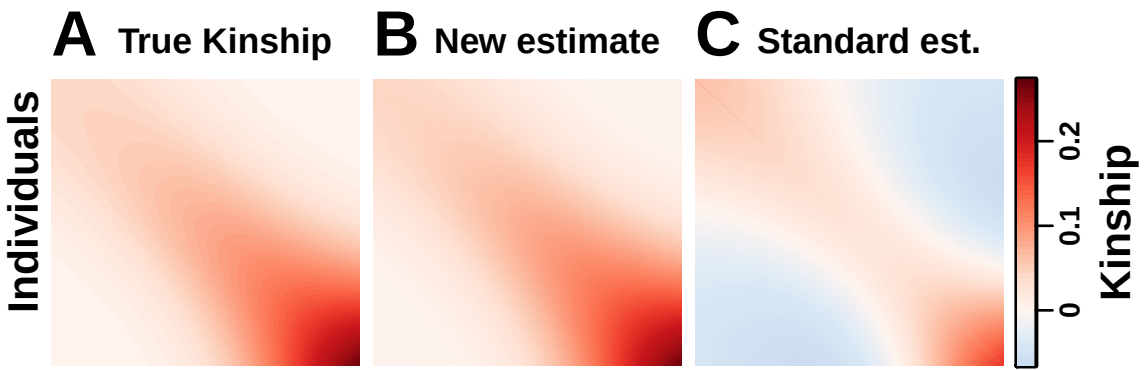
Standard kinship estimates

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

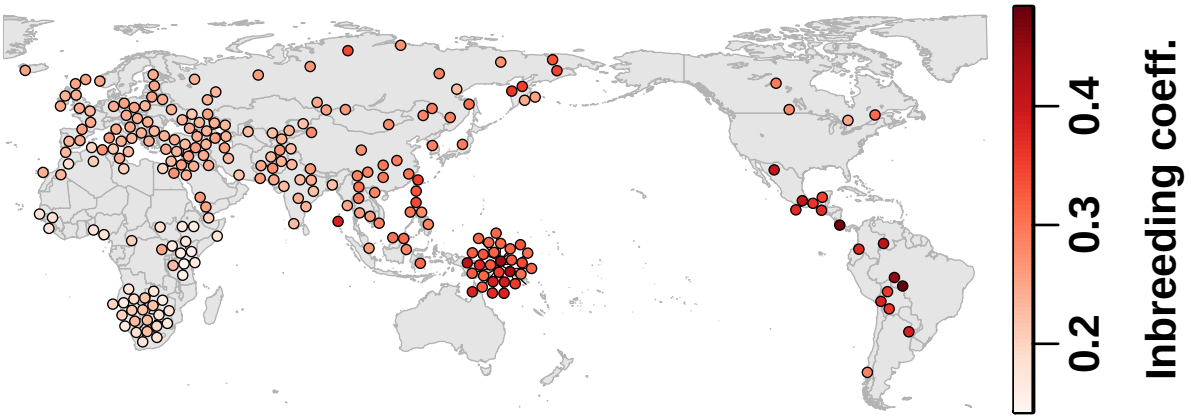


Edited from Ephert [CC BY-SA 3.0], via Wikimedia Commons

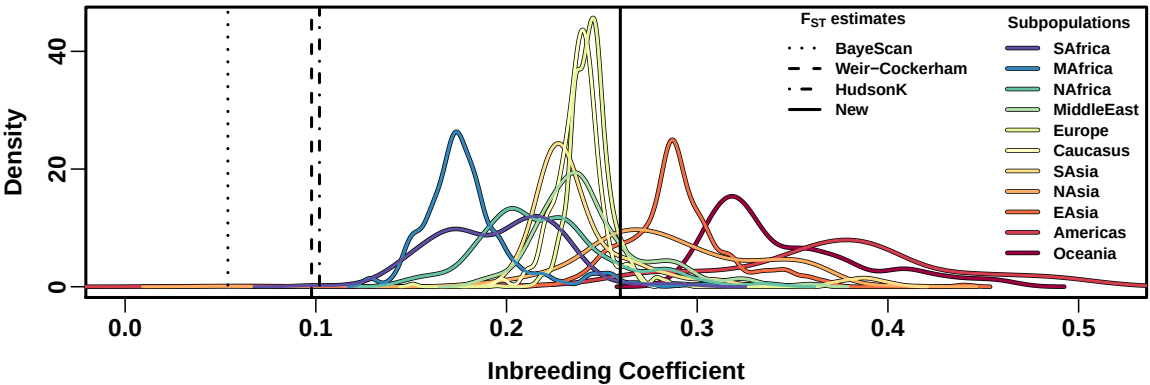
Only our new estimator is accurate in simulations



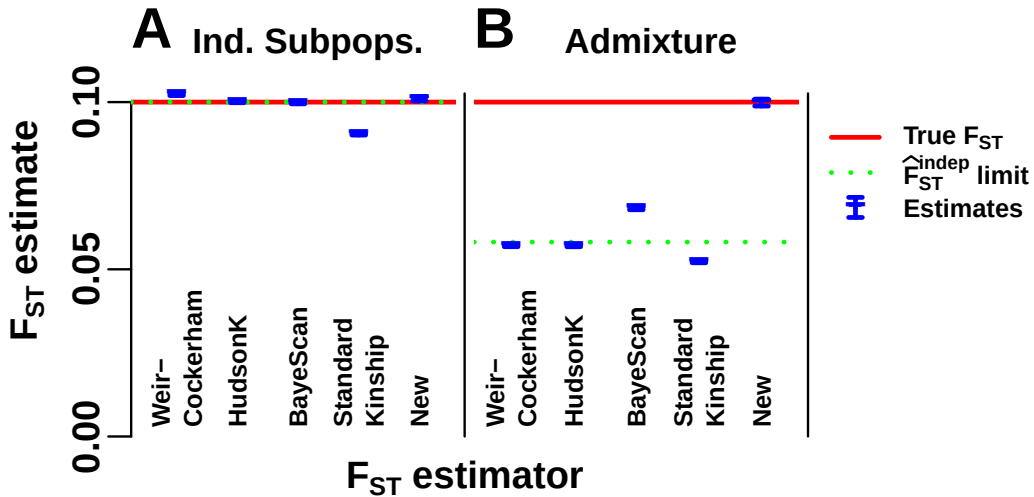
Population-level inbreeding increases with distance from Africa



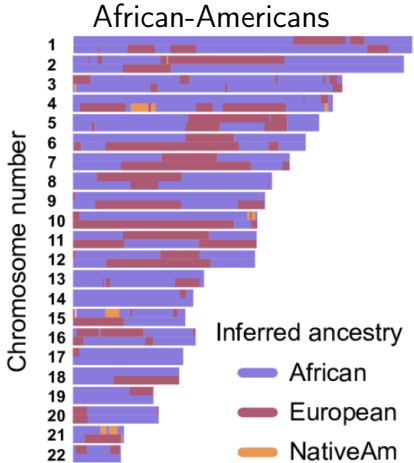
Differentiation (F_{ST}) previously underestimated



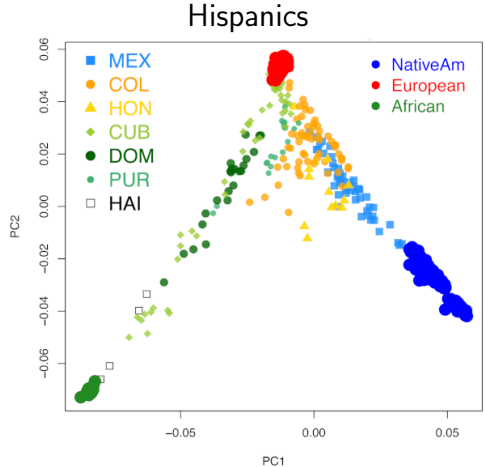
Only our new method estimates generalized F_{ST} accurately



Recently-admixed populations



Baharian *et al.* (2016)



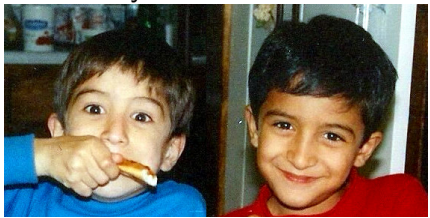
Moreno-Estrada *et al.* (2013)

Admixed siblings from different subpopulations?

Lucy and Maria, UK

Admixed siblings from different subpopulations?

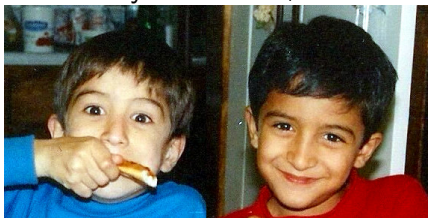
Lucy and Maria, UK



Ochoa brothers, MX

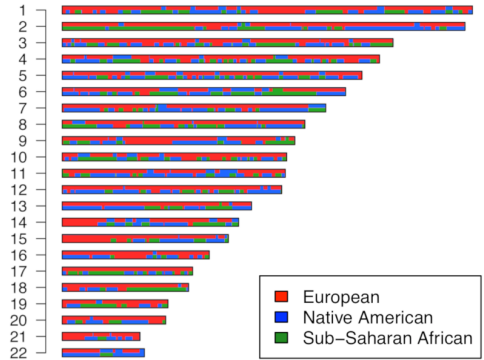
Admixed siblings from different subpopulations?

Lucy and Maria, UK



Ochoa brothers, MX

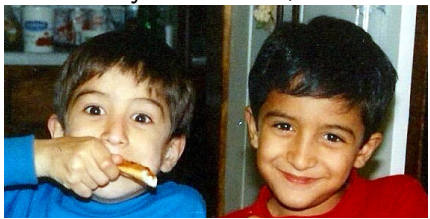
High Admixture LD:



Moreno-Estrada *et al.* (2013)

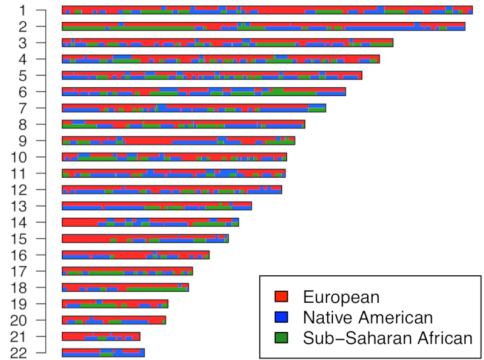
Admixed siblings from different subpopulations?

Lucy and Maria, UK



Ochoa brothers, MX

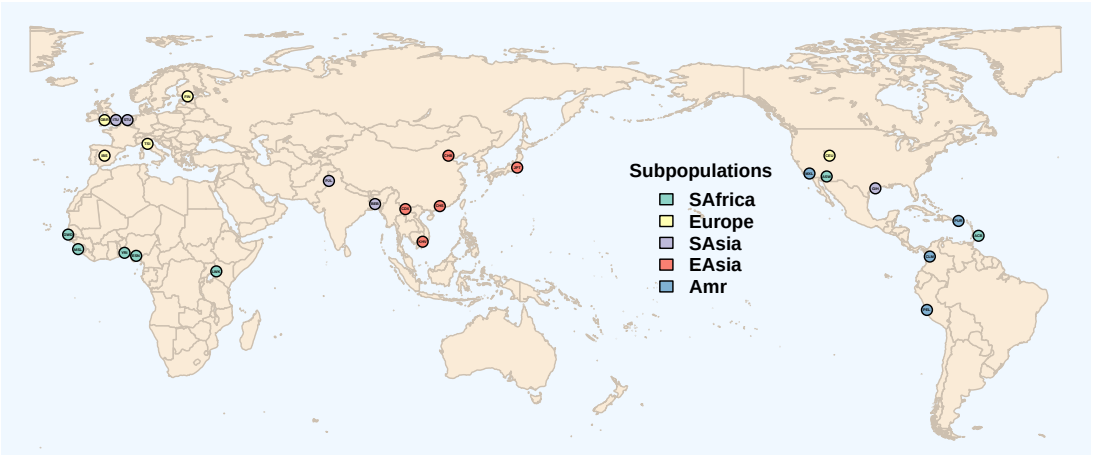
High Admixture LD:



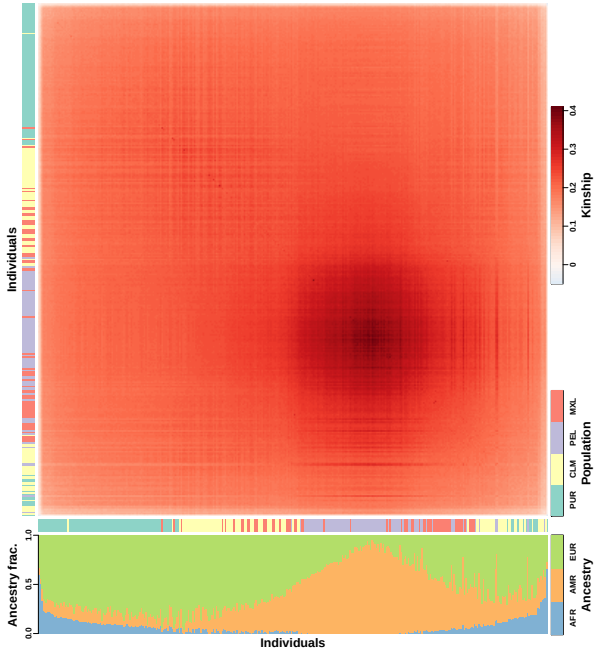
Moreno-Estrada *et al.* (2013)

Solution: treat every individual as their own subpopulation!

Dataset: 1000 Genomes Project (2013)



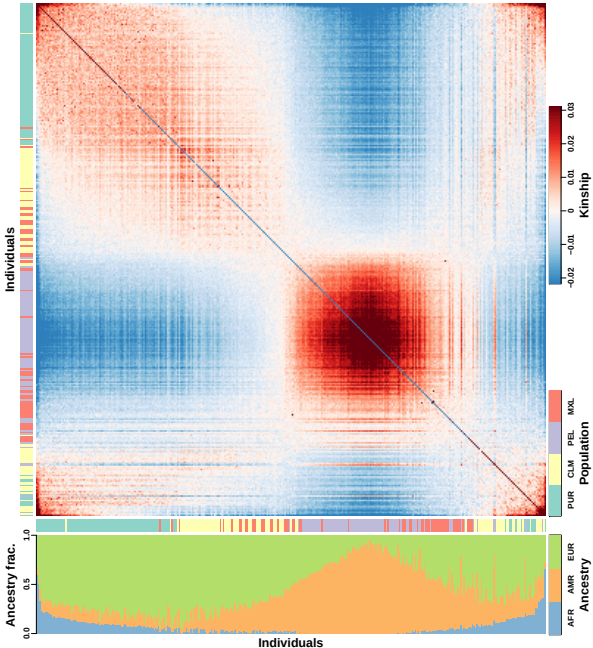
2,504 indivs. from 26 locs. — 20,417,698 loci (asc. in YRI) — WGS trios, etc.



Kinship driven by admixture in Hispanics

Our new kinship estimates

Genotypes from the 1000 Genomes Project (2013)



Standard kinship estimates

Hispanics in 1000 Genomes

Genotypes from the 1000 Genomes Project (2013)

R popkin implementation: fast and low memory usage!

Estimator:

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \qquad \hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{A_{\min}}.$$

R popkin implementation: fast and low memory usage!

Estimator:

$$A_{jk} = \frac{1}{m} \sum_{i=1}^{m} (x_{ij} - 1)(x_{ik} - 1) - 1, \qquad \hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{A_{\min}}.$$

Fastest: matrix product (R vectorizes)

$$\mathbf{A} = \frac{1}{m} (\mathbf{X} - \mathbf{1})^{\intercal} (\mathbf{X} - \mathbf{1}) - \mathbf{1}.$$

R popkin implementation: fast and low memory usage!

Estimator:

$$A_{jk} = \frac{1}{m} \sum_{i=1}^{m} (x_{ij} - 1)(x_{ik} - 1) - 1, \qquad \hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{A_{\min}}.$$

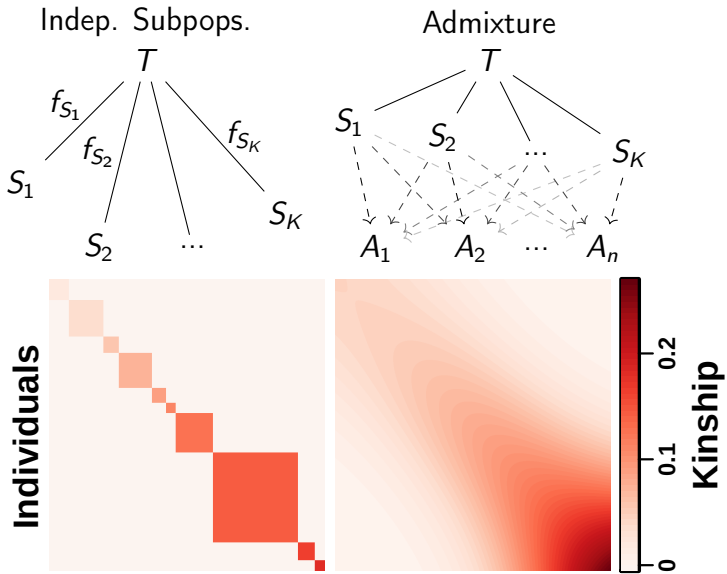
Fastest: matrix product (R vectorizes)

$$\mathbf{A} = \frac{1}{m} (\mathbf{X} - \mathbf{1})^{\intercal} (\mathbf{X} - \mathbf{1}) - \mathbf{1}.$$

Problem: R consumes too much memory.
Solution:

- Compute $m\mathbf{A}$ in parts (it's a running sum), max memory is controlled
- Further problems: missing genotypes, excessive matrix copying (solved using `RcppEigen`)

Comparison of population structures in simulation



F_{ST} in the independent subpopulation model

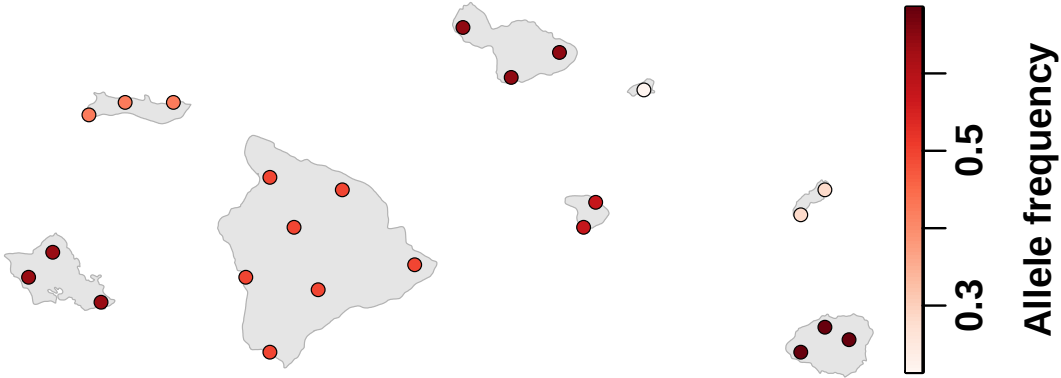


Illustration.

F_{ST} in the independent subpopulation model

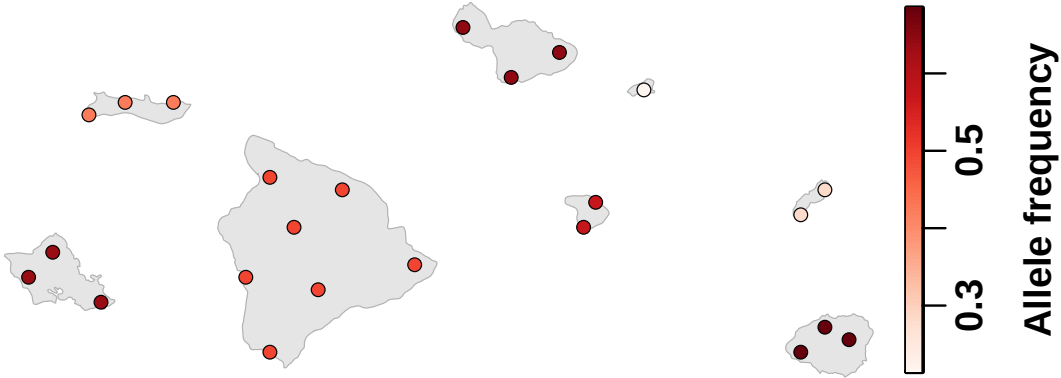


Illustration.

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

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

Wright's F_{ST}

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

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

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

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

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

Our generalized F_{ST}

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

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

Our generalized F_{ST}

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

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

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

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

Our generalized F_{ST}

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

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

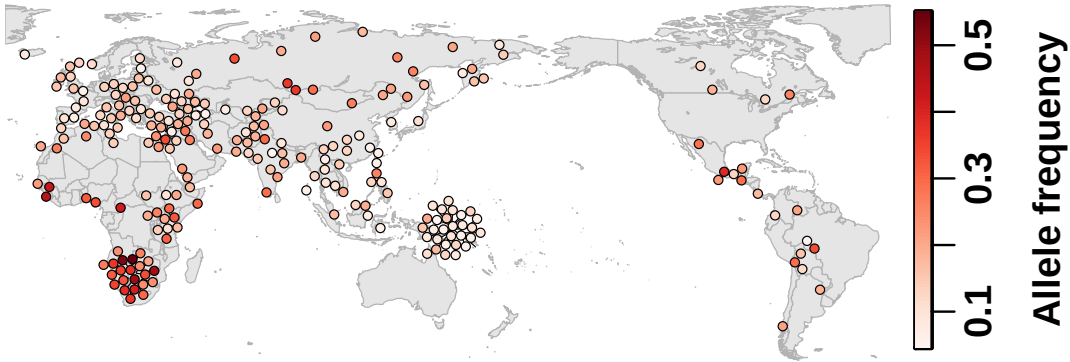
Generalized F_{ST}: applicable to arbitrary population structures, equals previous definition for subpopulations:

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

Mean heterozygosity in a structured population:

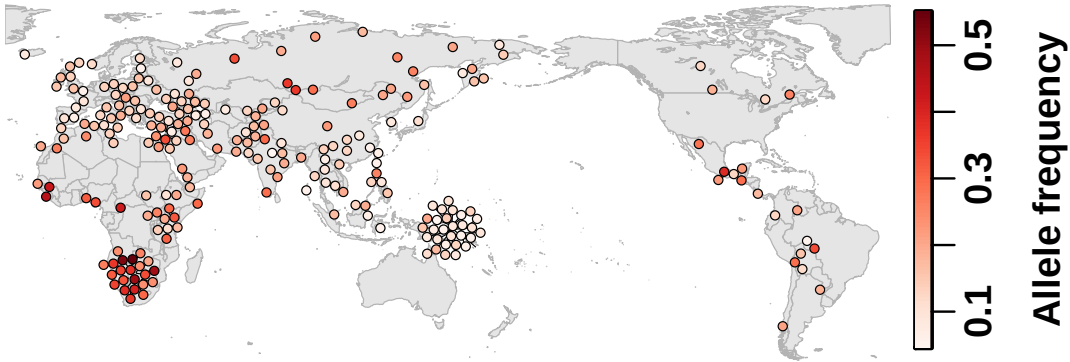
$$\bar{H}_i = \frac{1}{n}\sum_{j=1}^{n} \Pr(x_{ij} = 1) = 2p_i\left(1 - p_i\right)\left(1 - F_{ST}\right).$$

F_{ST} measures population structure / differentiation



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

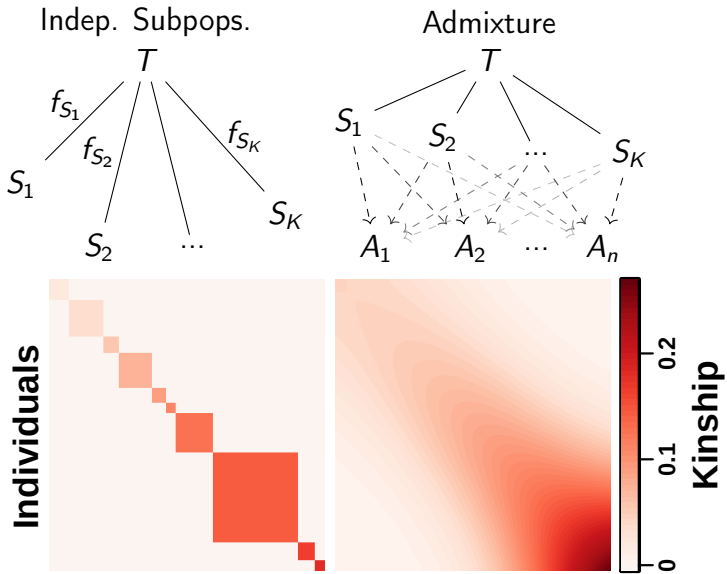
F_{ST} measures population structure / differentiation



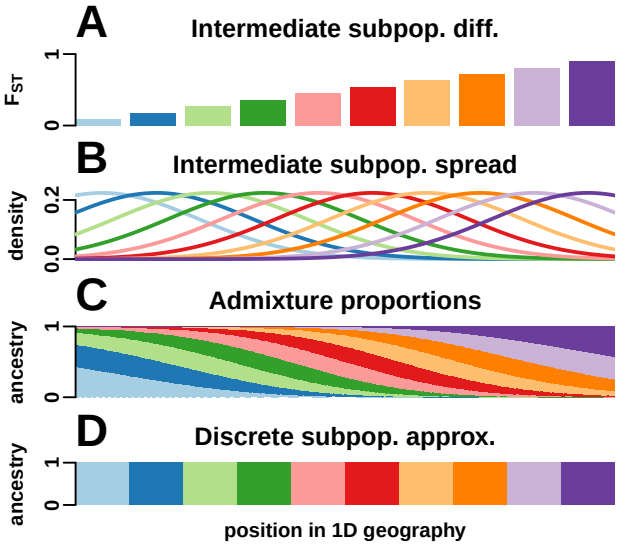
Median diff. SNP in Human Origins (rs2650044; given MAF $\geq 10\%$).

$\hat{F}_{ST}^{WC} \approx 0.0961$ 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
i	locus index
j, k	individual indexes
p_i	ref allele frequency
x_{ij}	genotype (num ref alleles)
φ_{jk}	kinship of j, k
f_j	inbreeding of j

Statistical model:

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

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

Problem: common estimators not consistent under structure

Estimate of ancestral allele frequency:

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

Variance asymptotically > 0 under population structure:

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

Problem: common estimators not consistent under structure

Estimate of ancestral allele frequency:

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

Variance asymptotically > 0 under population structure:

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

(For indep. individuals: $\bar{\varphi} = \frac{1}{2n}$.
$\Rightarrow n_{\mathrm{eff}} \approx 6$ indep. haplotypes in Human Origins!)

Problem: common estimators not consistent under structure

Estimate of ancestral allele frequency:

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

Variance asymptotically > 0 under population structure:

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

(For indep. individuals: $\bar{\varphi} = \frac{1}{2n}$.
$\Rightarrow n_{\mathrm{eff}} \approx 6$ indep. haplotypes in Human Origins!)

Naive estimators that use $\hat{p}_i$ (next) are not consistent!

Bias in standard kinship estimator

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

Bias in standard kinship estimator

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

Bias varies by j, k:

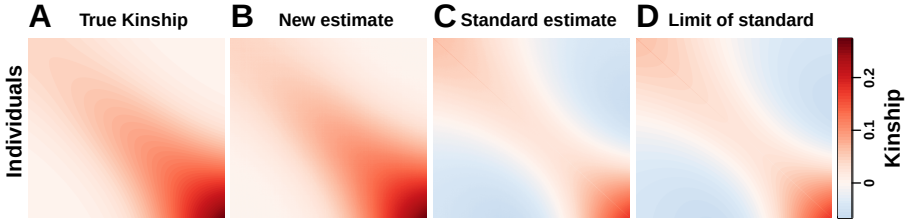
$$\hat{\varphi}_{jk}^{\text{std}} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1 - \bar{\varphi}}.$$

Bias in standard kinship estimator

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

Bias varies by j, k:

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



Our new estimator (R package `popkin` on CRAN)

Step 1: estimates kinship scaled by nuisance v (function of all p_i)

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \qquad \mathrm{E}\left[A_{jk}\right] = (\varphi_{jk}-1)v$$

Our new estimator (R package `popkin` on CRAN)

Step 1: estimates kinship scaled by nuisance v (function of all p_i)

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \qquad \mathrm{E}\left[A_{jk}\right] = (\varphi_{jk}-1)v$$

Step 2: Estimate minimum, unbias "step 1" estimates

$$\text{If} \quad \mathrm{E}[A_{\min}] = -v, \qquad \text{then} \qquad \hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{A_{\min}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}.$$

Our new estimator (R package `popkin` on CRAN)

Step 1: estimates kinship scaled by nuisance v (function of all p_i)

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \qquad \mathrm{E}\left[A_{jk}\right] = (\varphi_{jk}-1)v$$

Step 2: Estimate minimum, unbias "step 1" estimates

$$\text{If} \quad \mathrm{E}[A_{\min}] = -v, \qquad \text{then} \qquad \hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{A_{\min}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}.$$

Practical estimator: average within most extreme subpopulations

$$A_{\min} = \min_{u \neq v} \frac{1}{|S_u||S_v|} \sum_{j \in S_u} \sum_{k \in S_v} A_{jk}.$$

Our new estimator (R package `popkin` on CRAN)

Step 1: estimates kinship scaled by nuisance v (function of all p_i)

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1) - 1, \qquad \mathrm{E}\left[A_{jk}\right] = (\varphi_{jk}-1)v$$

Step 2: Estimate minimum, unbias "step 1" estimates

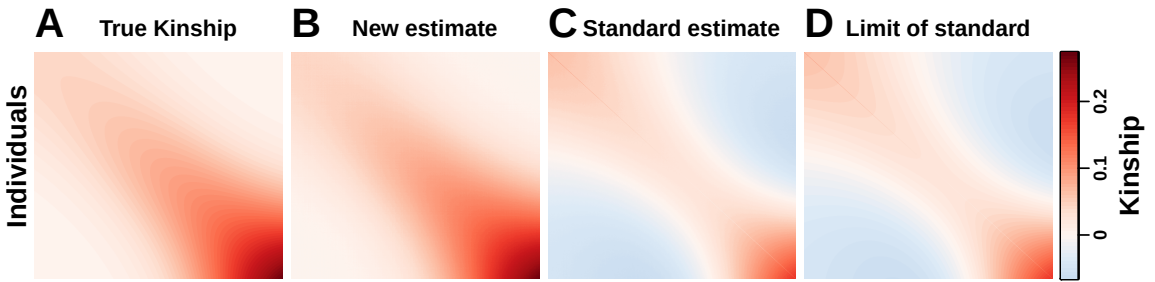
$$\text{If} \quad \mathrm{E}[A_{\min}] = -v, \qquad \text{then} \qquad \hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{A_{\min}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}.$$

Practical estimator: average within most extreme subpopulations

$$A_{\min} = \min_{u\neq v} \frac{1}{|S_u||S_v|} \sum_{j\in S_u}\sum_{k\in S_v} A_{jk}.$$

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

Performance of new estimator



Bias in F_{ST} estimators for independent subpopulations

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

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

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

Bias in F_{ST} estimators for independent subpopulations

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

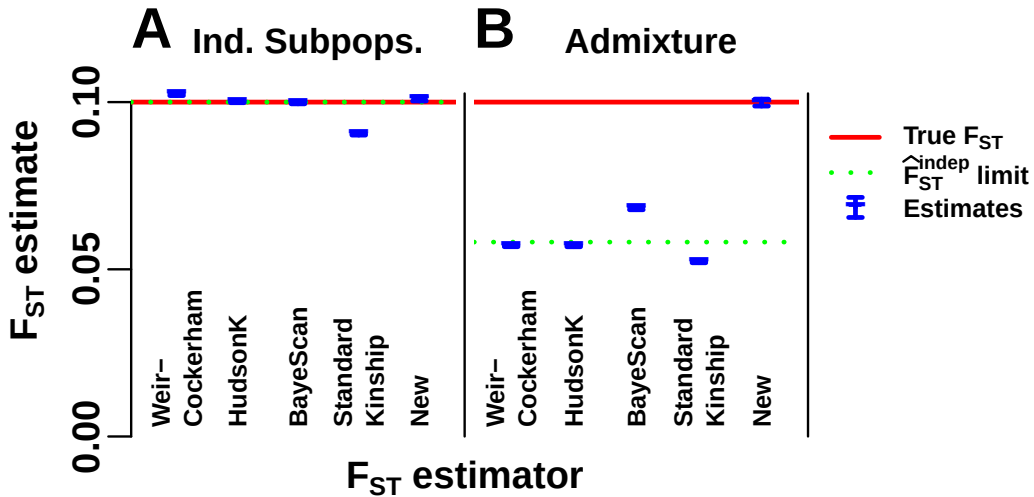
$$\hat{F}_{ST}^{indep} = \frac{\sum_{i=1}^{m} \hat{\sigma}_i^2}{\sum_{i=1}^{m} \hat{p}_i (1 - \hat{p}_i) + \frac{1}{n}\hat{\sigma}_i^2},$$

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

Estimator is biased in dependent subpopulations:

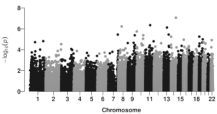
$$\hat{F}_{ST}^{indep} \xrightarrow[m\to\infty]{a.s.} \frac{F_{ST} - \frac{1}{n-1}\left(n\bar{\theta} - F_{ST}\right)}{1 - \frac{1}{n-1}\left(n\bar{\theta} - F_{ST}\right)}.$$

Only our new method estimates generalized F_{ST} accurately



The future: improved kinship has repercussions across genetics!

Accurate and efficient estimation, Ancestry



Association studies, Selection tests

Heritability of complex traits

Animal and plant breeding

Acknowledgments

Princeton University
John D. Storey
Wei Hao

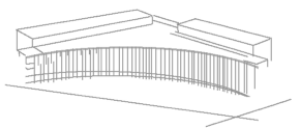
University of Warsaw
Neo Christopher Chung

Funding
National Institutes of Health
Otsuka Pharmaceutical

Duke StatGen
Bioinformatics and Biostatistics

Duke Center for Statistical Genetics and Genomics

PRINCETON UNIVERSITY



Lewis-Sigler Institute for Integrative Genomics

Method-of-moments derivation of new estimator (1/3)

Compute raw moments:

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

For symmetry, raw moments of $2 - x_{ij}$ (counting other allele):

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

Resist the temptation to solve for p_i!

Method-of-moments derivation of new estimator (2/3)

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

Let's average second moments. First note:

$$\frac{1}{2}\left(x_{ij}x_{ik} + (2-x_{ij})(2-x_{ik})\right) = (1-x_{ij})(1-x_{ik}) + 1,$$
$$\frac{1}{2}\left(p_i^2 + (1-p_i)^2\right) = \frac{1}{2} - p_i\,(1-p_i).$$

Therefore, the symmetric estimator is

$$\mathrm{E}\left[(1-x_{ij})(1-x_{ik}) + 1\right] = 2 + 4p_i\,(1-p_i)\,(\varphi_{jk}-1) \quad \Rightarrow$$
$$\mathrm{E}\left[(1-x_{ij})(1-x_{ik}) - 1\right] = 4p_i\,(1-p_i)\,(\varphi_{jk}-1).$$

Method-of-moments derivation of new estimator (3/3)

$$\mathrm{E}\left[(1 - x_{ij})(1 - x_{ik}) - 1\right] = 4p_i\left(1 - p_i\right)\left(\varphi_{jk} - 1\right).$$

Average across loci to reduce variance

$$A_{jk} = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - 1)(x_{ik} - 1) - 1,$$

$$\mathrm{E}\left[A_{jk}\right] = \left(\varphi_{jk} - 1\right)v,$$

$$v = \frac{4}{m}\sum_{i=1}^{m} p_i\left(1 - p_i\right).$$

A good estimate of the minimum value yields φ_{jk}:

$$\text{If} \quad \mathrm{E}[A_{\min}] = -v, \qquad \text{then} \qquad \hat{\varphi}_{jk}^{\text{new}} = 1 - \frac{A_{jk}}{A_{\min}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}.$$

Derivation of kinship model

Here treat $x_{ij}, x_{ik} \in \{0, 1\}$ as haplotypes. Joint distribution:

	Independent	Full IBD	Partial IBD
$\Pr(x_{ij} = 1, x_{ik} = 1)$	p_i^2	p_i	$(1 - \varphi_{jk})\, p_i^2 + \varphi_{jk} p_i$
$\Pr(x_{ij} = 1, x_{ik} = 0)$	$p_i\,(1 - p_i)^2$	0	$(1 - \varphi_{jk})\, p_i\,(1 - p_i)^2$
$\Pr(x_{ij} = 0, x_{ik} = 1)$	$p_i\,(1 - p_i)^2$	0	$(1 - \varphi_{jk})\, p_i\,(1 - p_i)^2$
$\Pr(x_{ij} = 0, x_{ik} = 0)$	$(1 - p_i)^2$	$1 - p_i$	$(1 - \varphi_{jk})\,(1 - p_i)^2 + \varphi_{jk}\,(1 - p_i)$

Kinship model: mixture of (Independent, Full IBD), weights $(1 - \varphi_{jk}, \varphi_{jk})$.
This follows:

$$\mathrm{E}[x_{ij}] = \mathrm{E}[x_{ik}] = p_i, \quad \mathrm{Cov}(x_{ij}, x_{ik}) = p_i\,(1 - p_i)\,\varphi_{jk}.$$

Admixture models

q_{ju}: ancestry proportion
$p_i^{S_u}$: AF in subpopulation S_u
f_{S_u}: F_{ST} of S_u

Draw alleles from a mixture of populations:

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

If subpopulations are independent,

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