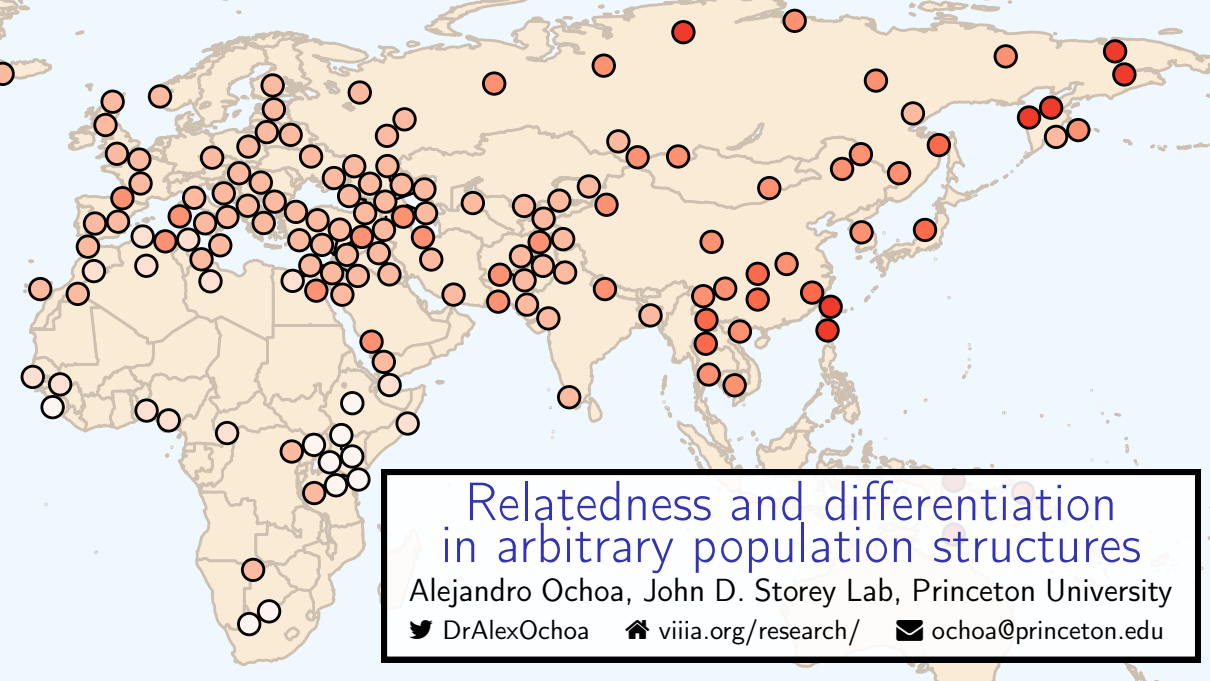


Relatedness and differentiation in arbitrary population structures

Alejandro Ochoa, John D. Storey Lab, Princeton University

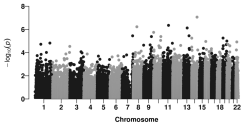
DrAlexOchoa viiia.org/research/ ochoa@princeton.edu

Why study relatedness?

Human genetics
is fascinating!

Why study relatedness?

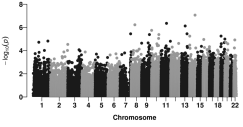
Human genetics is fascinating!



Pop. structure confounds association studies (GWAS)

Why study relatedness?

Human genetics is fascinating!

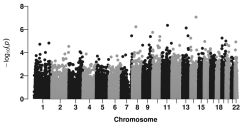


Pop. structure confounds association studies (GWAS)

Heritability of complex traits

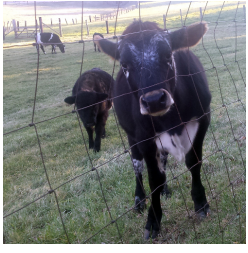
Why study relatedness?

Human genetics is fascinating!



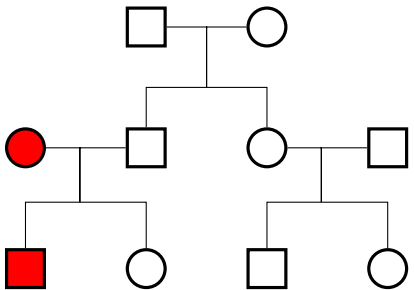
Pop. structure confounds association studies (GWAS)

Heritability of complex traits

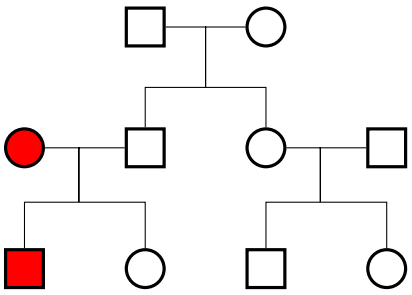


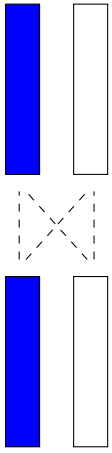
Animal and plant breeding

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

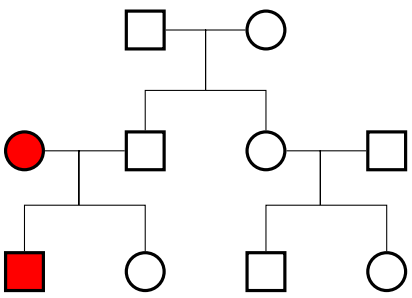


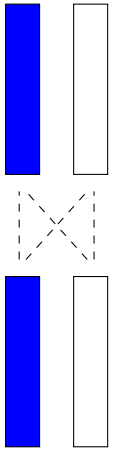
The kinship coefficient for parent-child: $\frac{1}{4}$

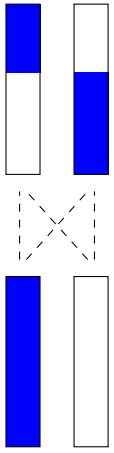




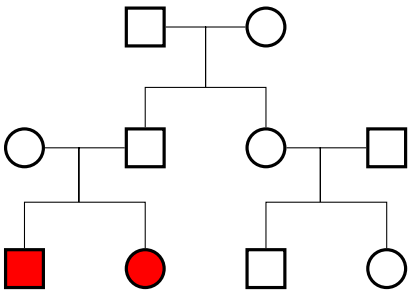
The kinship coefficient for parent-child: $\frac{1}{4}$



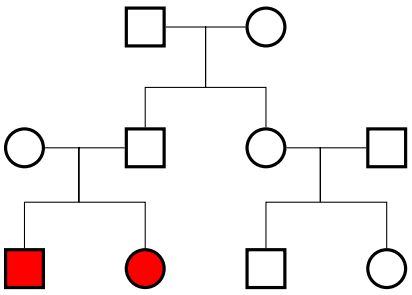


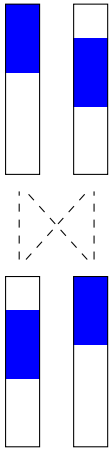


The kinship coefficient for siblings: $\frac{1}{4}$ on average

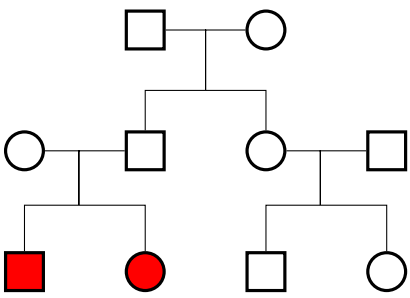


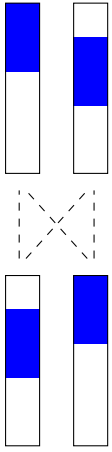
The kinship coefficient for siblings: $\frac{1}{4}$ on average

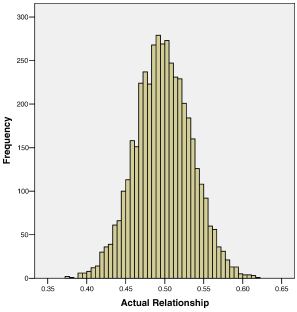




The kinship coefficient for siblings: $\frac{1}{4}$ on average

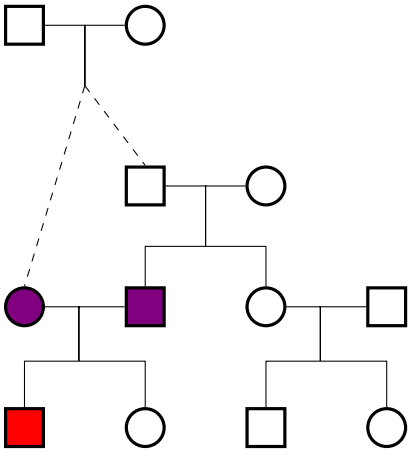




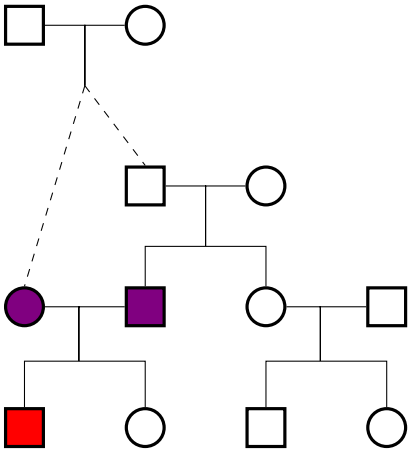


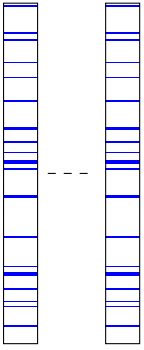
Visscher *et al.* (2006)

The inbreeding coefficient in populations

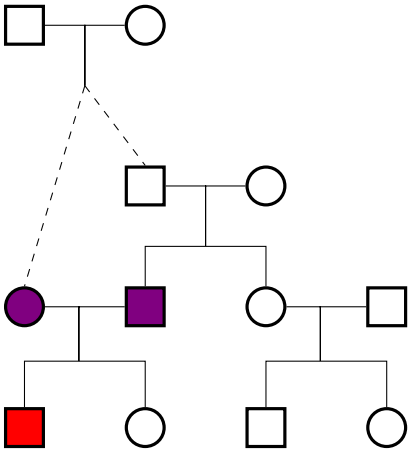


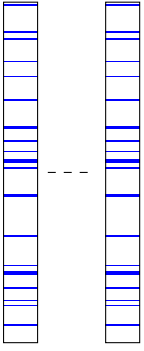
The inbreeding coefficient in populations





The inbreeding coefficient in populations





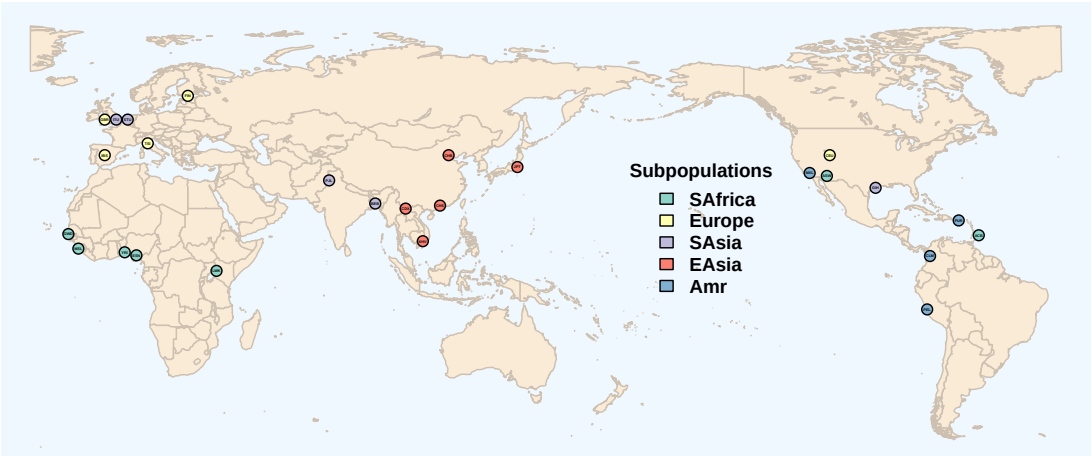
Measurements relative to a reference pop.:

Inbreeding $= 0$ in the local population

Inbreeding ≥ 0 relative to a distant ancestral population

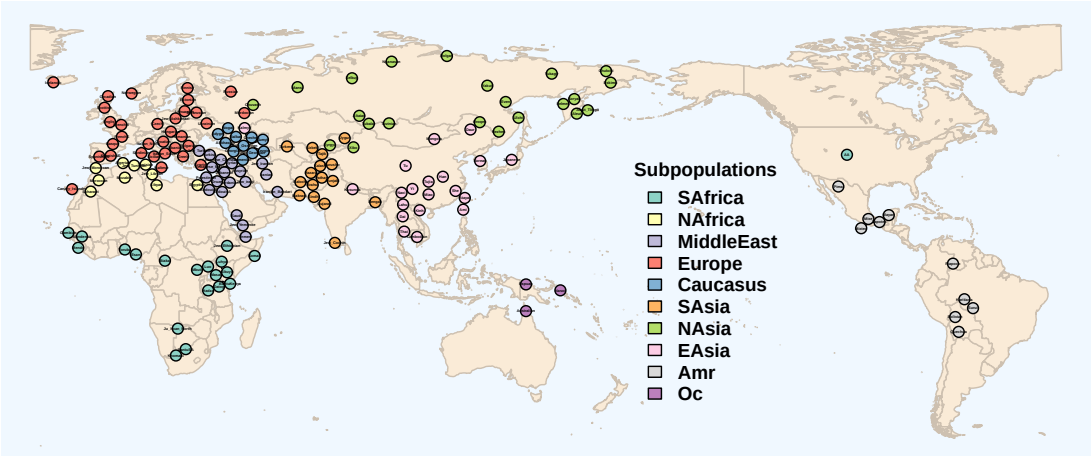
Better measured using covariance

Dataset: 1000 Genomes Project (2013)



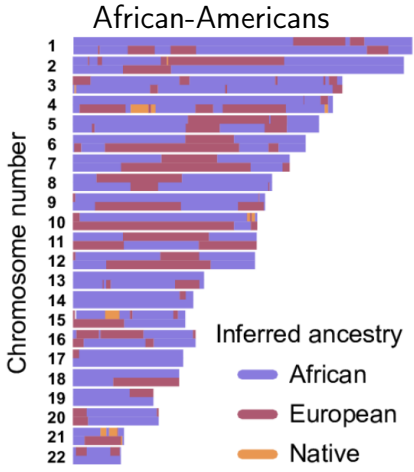
2,504 indivs. from 26 locs. — 20,417,698 loci (asc. in YRI) — WGS trios, etc.

Dataset: Human Origins (2016)

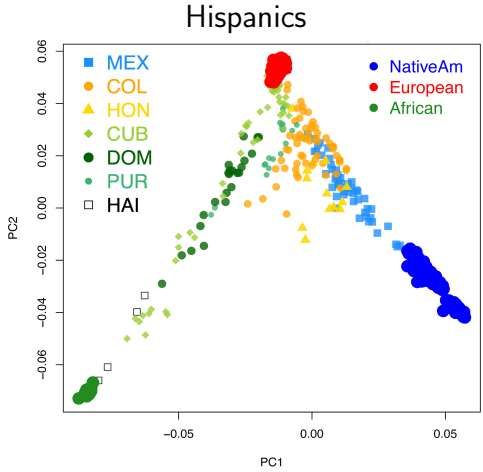


2,066 indivs. from 163 locs. — 595,911 loci — SNP chip

Recently admixed populations



Baharian, *et al.* (2016)



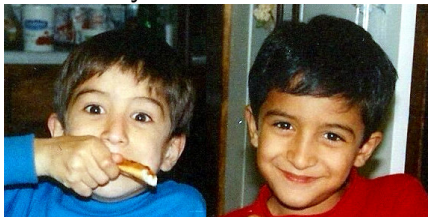
Moreno-Estrada, *et al.* (2013)

Admixed siblings from different populations?

Lucy and Maria, UK

Admixed siblings from different populations?

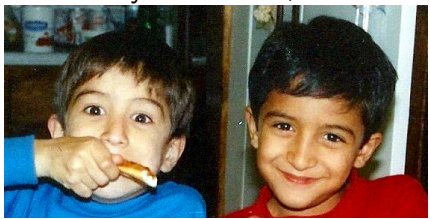
Lucy and Maria, UK



Ochoa brothers, MX

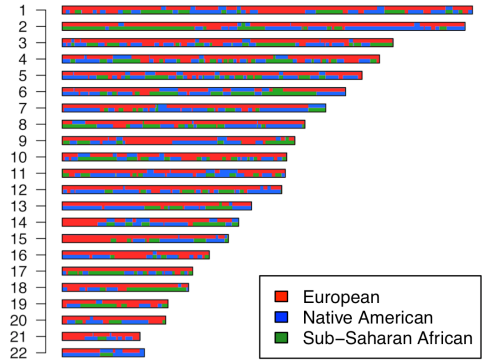
Admixed siblings from different populations?

Lucy and Maria, UK



Ochoa brothers, MX

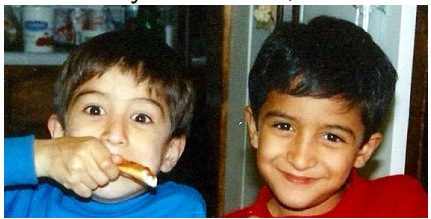
High Admixture LD:



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

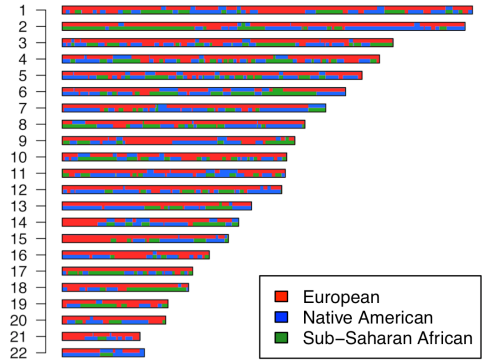
Admixed siblings from different populations?

Lucy and Maria, UK



Ochoa brothers, MX

High Admixture LD:



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

Solution: treat every individual as its own population!

F_{ST} in the independent subpopulation model

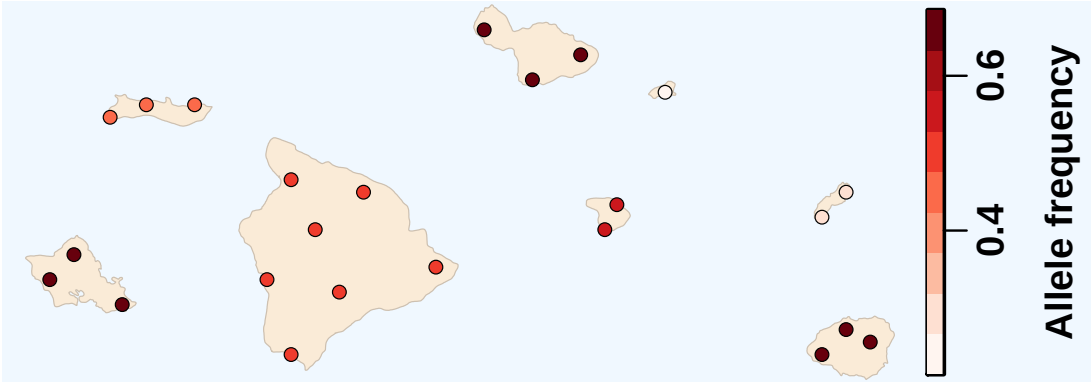


Illustration.

F_{ST} in the independent subpopulation model

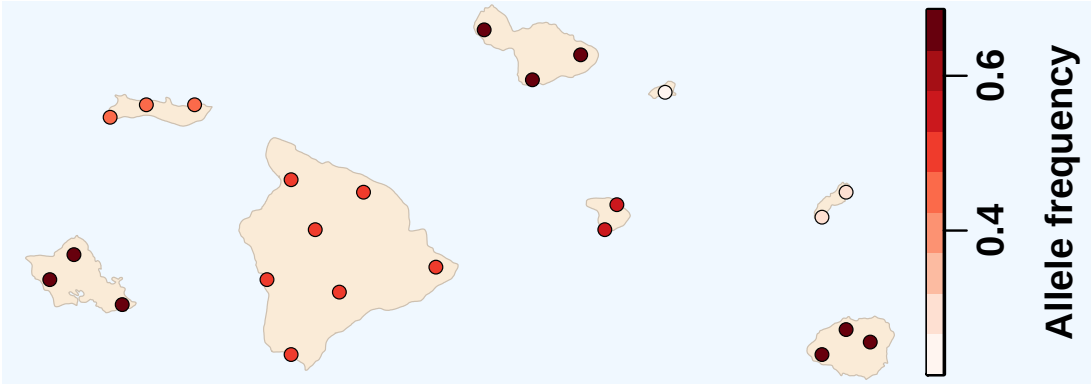
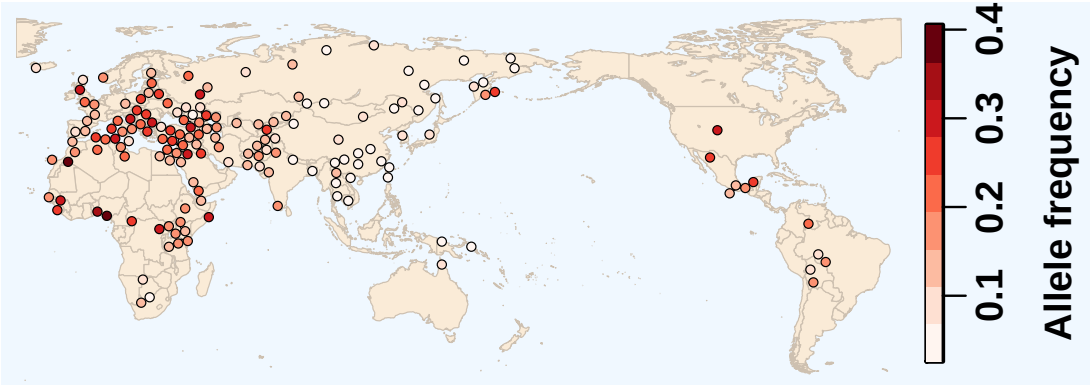


Illustration.

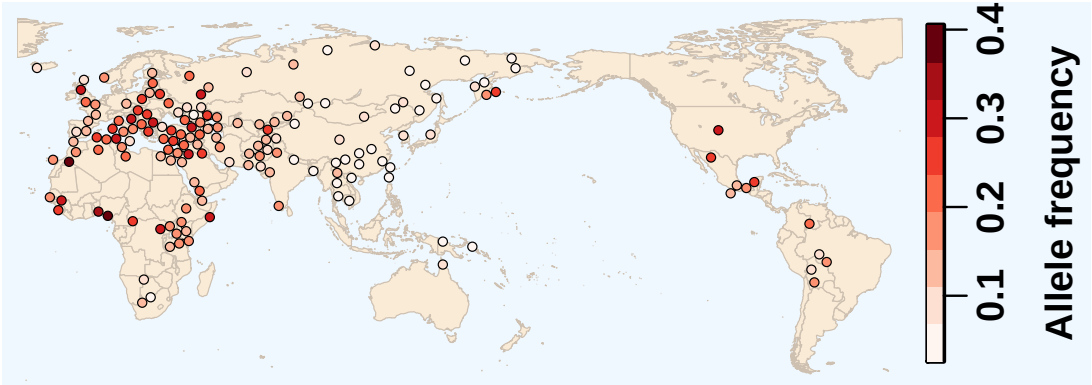
$$F_{ST} = \frac{\text{Var}\left(p_i^S \middle| T\right)}{p_i^T\left(1 - p_i^T\right)}.$$

F_{ST} measures population structure / differentiation



Median diff. SNP in Human Origins (rs7626601; given MAF $\geq$ 10%).

F_{ST} measures population structure / differentiation

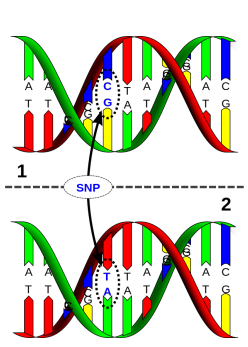


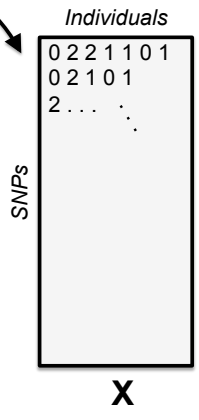
Median diff. SNP in Human Origins (rs7626601; given MAF $\geq 10\%$).

$\hat{F}_{ST}^{WC} \approx 0.0712$ using Weir-Cockerham estimator and $K = 163$.

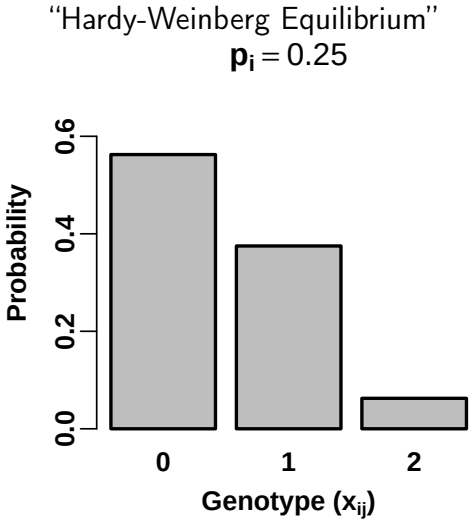
Single Nucleotide Polymorphism (SNP) data

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



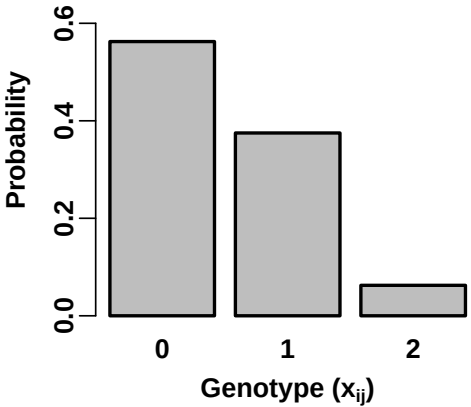


Increased homozygosity in a structured population

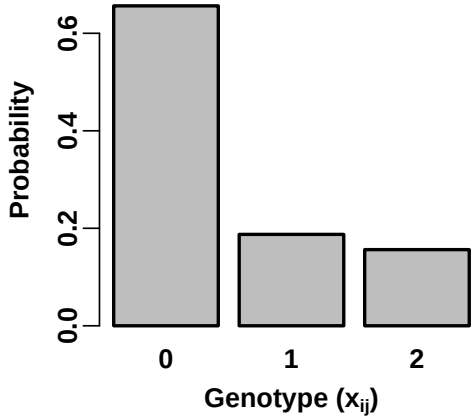


Increased homozygosity in a structured population

"Hardy-Weinberg Equilibrium"
$\mathbf{p_i} = 0.25$



A structured population
$\mathbf{p_i} = 0.25, \quad \mathbf{f_j} = 0.5$



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

Statistical model:

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

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

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

Statistical model:

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

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

We developed a new kinship estimation framework that works for arbitrary population structures!

Wright's F_{ST}

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

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

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

The generalized F_{ST}

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

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

The 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:

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

The generalized F_{ST}

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

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

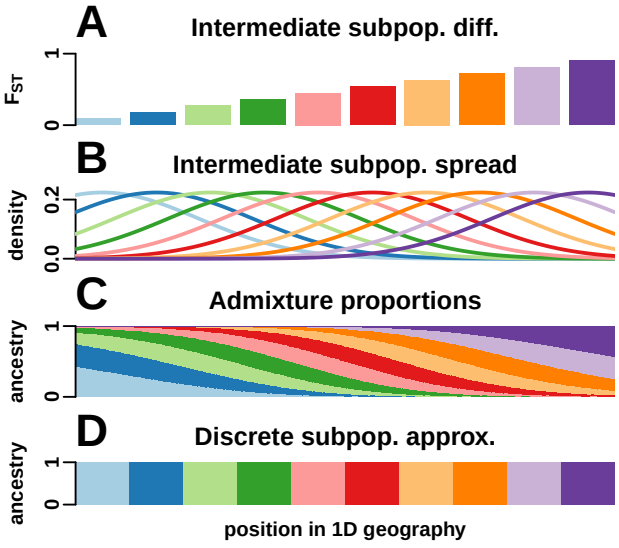
Generalized F_{ST}: applicable to arbitrary population structures, equals previous definition for non-overlapping subpopulations:

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

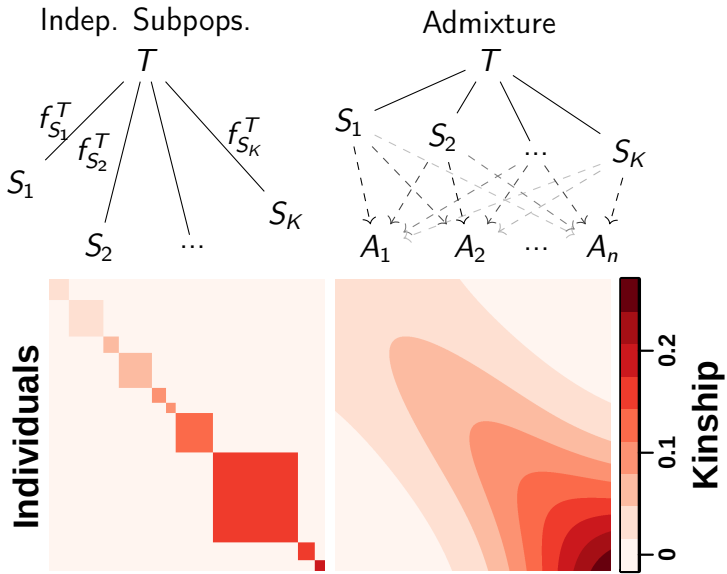
Mean heterozygosity in a structured population:

$$\bar{H}_i = 2p_i^T\left(1 - p_i^T\right)\left(1 - F_{ST}\right).$$

Our admixture simulation



Comparison of population structures in simulation



Bias in standard kinship estimator

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

Standard kinship estimator:

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

Estimator has a distorted bias (varies for every pair of individuals j, k):

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

New estimator: two steps

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

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

New estimator: two steps

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.

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

New estimator: two steps

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

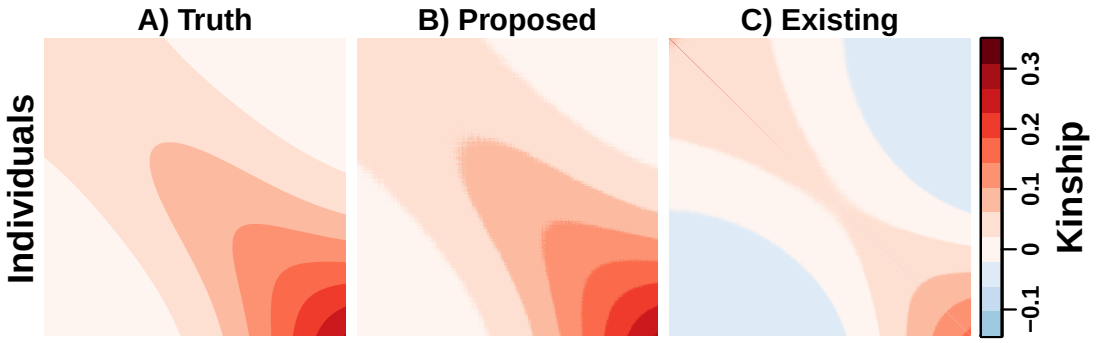
$$\hat{\varphi}_{jk}^{T,\text{preadj}} = \frac{\sum_{i=1}^{m}(x_{ij}-1)(x_{ik}-1)-1}{4\sum_{i=1}^{m}\hat{p}_i^T\left(1-\hat{p}_i^T\right)} + 1 \xrightarrow[m\to\infty]{\text{a.s.}} \frac{\varphi_{jk}^T - \bar{\varphi}^T}{1-\bar{\varphi}^T},$$

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

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

$$\hat{f}_j^{T,\text{new}} = 2\hat{\varphi}_{jj}^{T,\text{new}} - 1 \xrightarrow[m\to\infty]{\text{a.s.}} f_j^T, \qquad \hat{F}_{\text{ST}}^{\text{new}} = \sum_{j=1}^{n} w_j \hat{f}_j^{T,\text{new}} \xrightarrow[m\to\infty]{\text{a.s.}} F_{\text{ST}}$$

Performance of proposed estimator



Bias in F_{ST} estimators for independent subpopulations

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

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

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

Bias in F_{ST} estimators for independent subpopulations

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

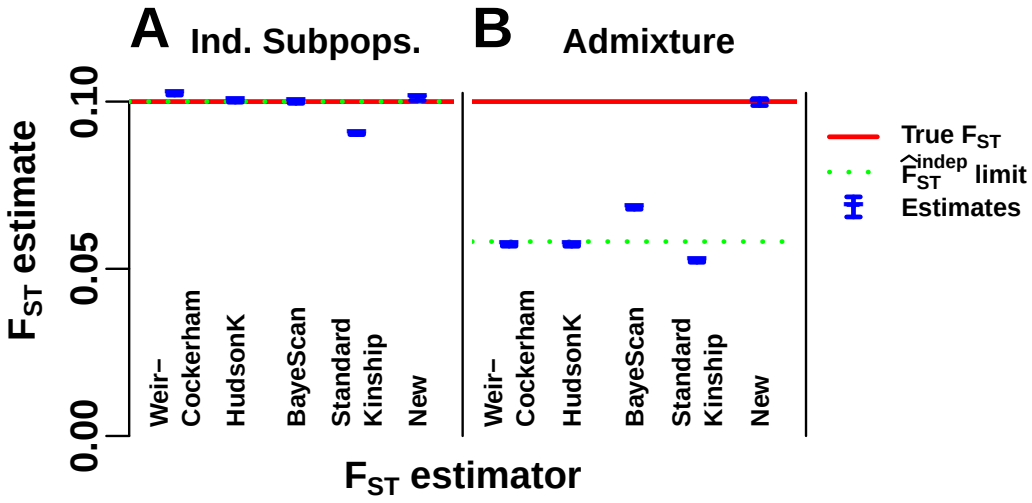
$$\hat{F}_{ST}^{indep} = \frac{\sum_{i=1}^{m} \hat{\sigma}_i^2}{\sum_{i=1}^{m} \hat{p}_i^T \left(1 - \hat{p}_i^T\right) + \frac{1}{n}\hat{\sigma}_i^2},$$

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

Estimator is biased in dependent subpopulations:

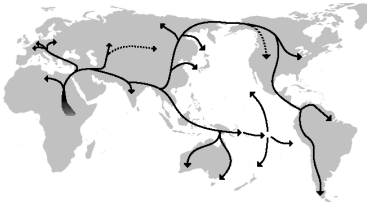
$$\hat{F}_{ST}^{indep} \xrightarrow[m\to\infty]{\text{a.s.}} \frac{F_{ST} - \frac{1}{n-1}\left(n\bar{\theta}^T - F_{ST}\right)}{1 - \frac{1}{n-1}\left(n\bar{\theta}^T - F_{ST}\right)}.$$

Bias estimating the generalized F_{ST}



New kinship estimates

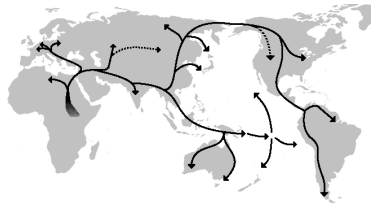
Genotypes from "Human Origins" (Lazaridis et al. 2014, 2016)



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

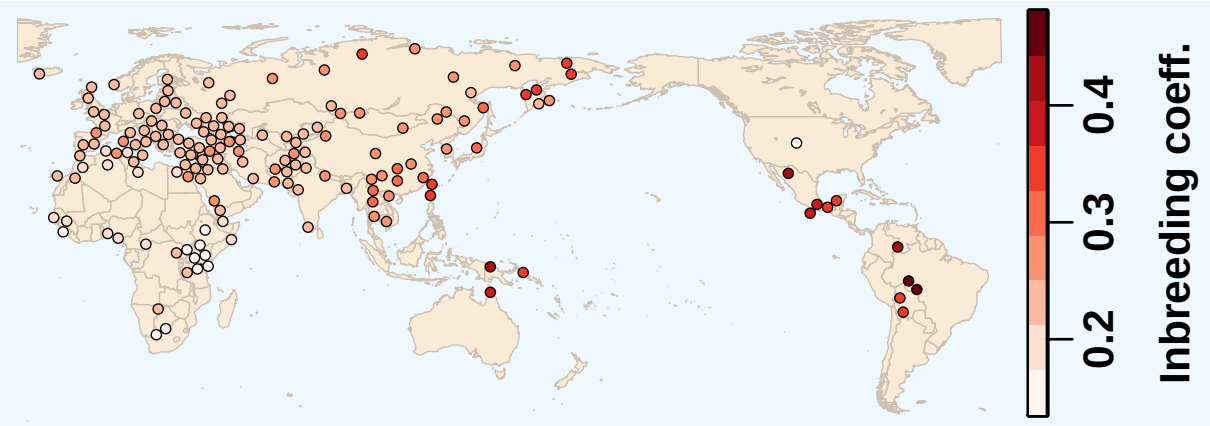
Standard kinship estimates

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

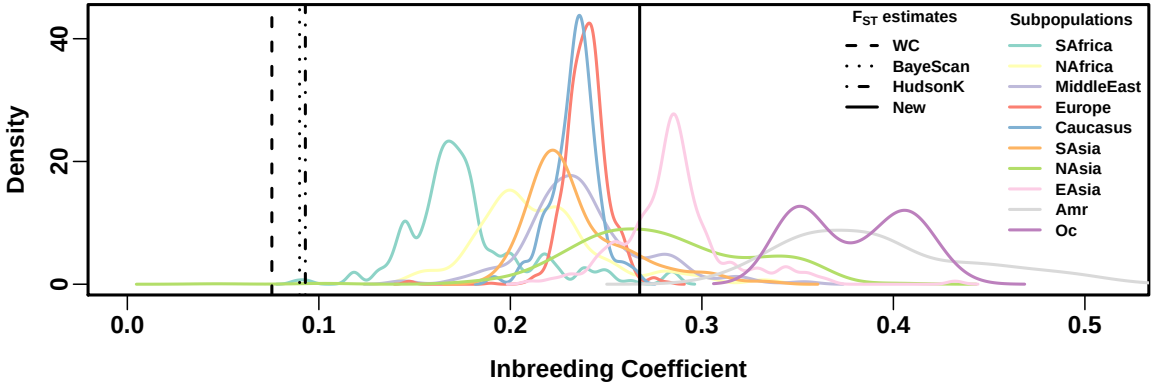


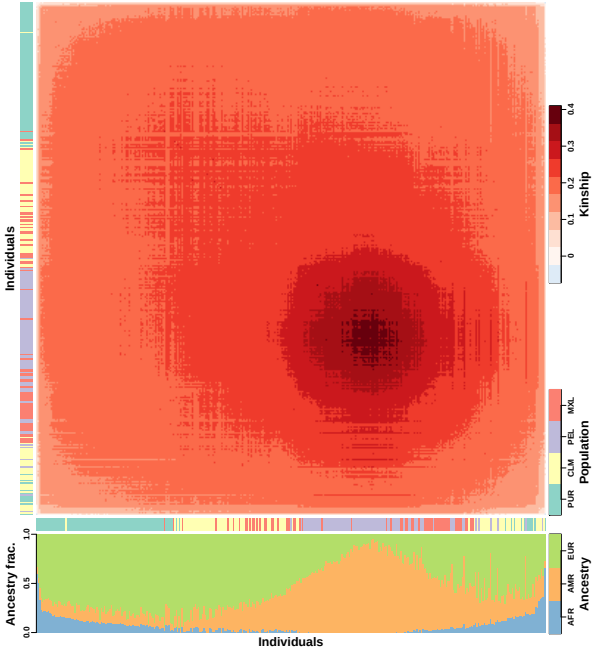
Edited from Ephert [CC BY-SA 3.0], via Wikimedia Commons

Population-level inbreeding



Differentiation (F_{ST}) previously underestimated





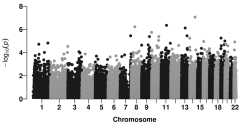
Kinship driven by admixture in Hispanics

New kinship estimates

Genotypes from the 1000 Genomes Project (2013)

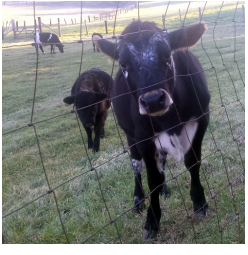
Improved relatedness has repercussions across genetics!

Easy to measure routinely



Search for disease-causing genetic variants

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 Pharmaceuticals