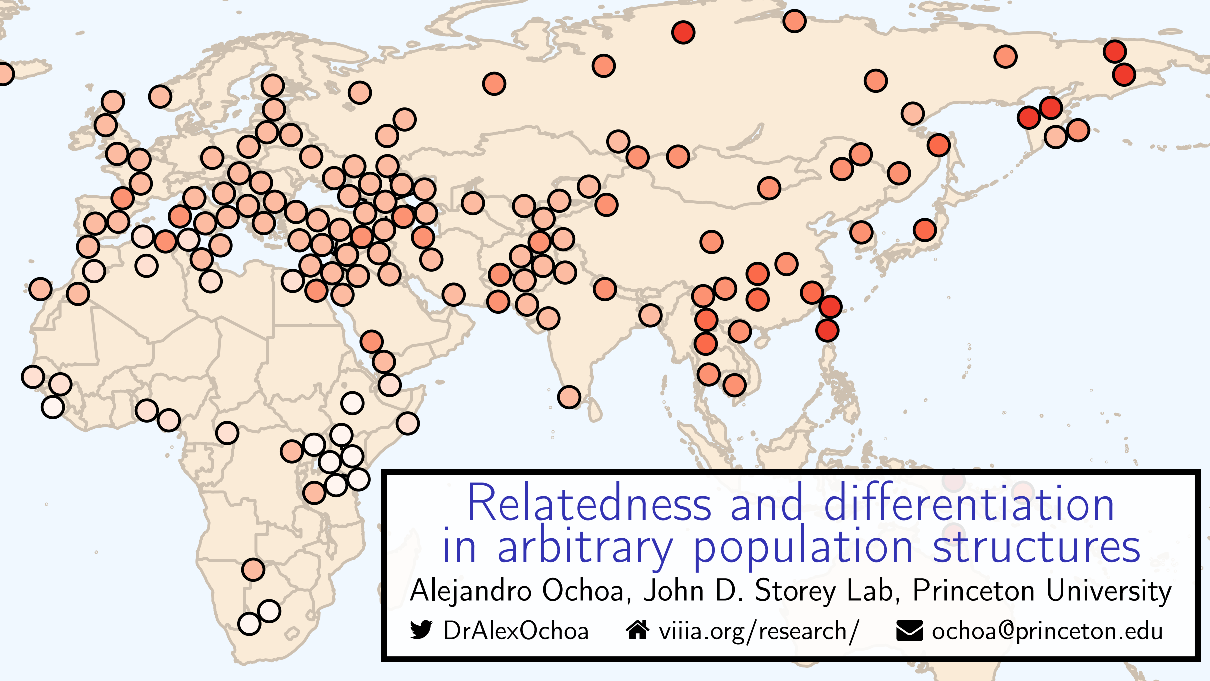
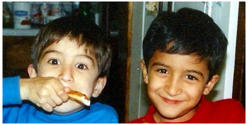


Relatedness and differentiation in arbitrary population structures

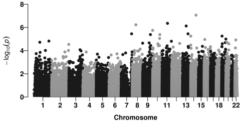
Alejandro Ochoa, John D. Storey Lab, Princeton University

🐦 DrAlexOchoa 🏠 viiia.org/research/ ✉ ochoa@princeton.edu

Why study relatedness?



Human genetics is fascinating!

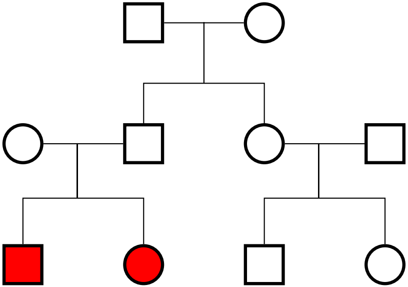


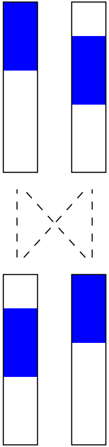
Search for disease-causing genetic variants

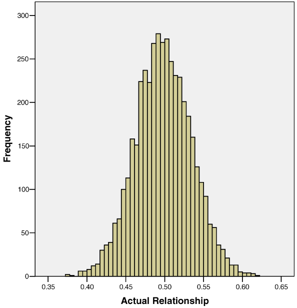
Heritability of complex traits

Animal and plant breeding

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

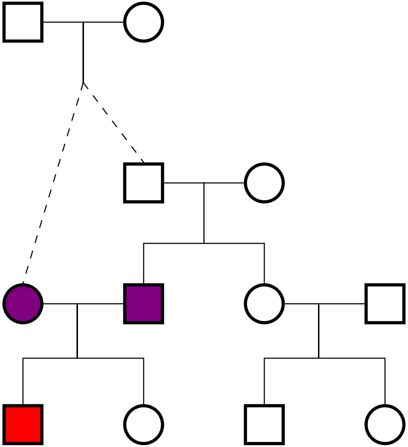


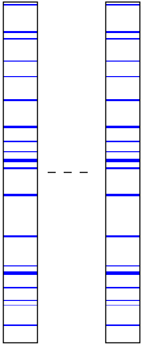




Visscher *et al.* (2006)

The inbreeding coefficient in populations





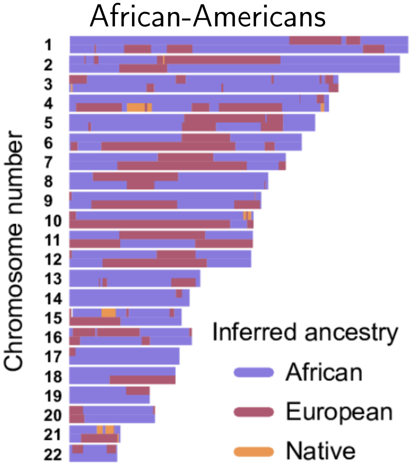
Measurements relative to a reference pop.:

Inbreeding = 0 in the local population

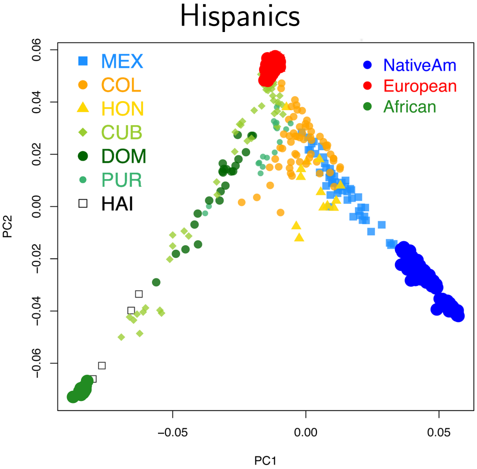
Inbreeding ≥ 0 relative to a distant ancestral population

Better measured using covariance

Recently admixed populations



Baharian *et al.* (2016)



Moreno-Estrada *et al.* (2013)

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

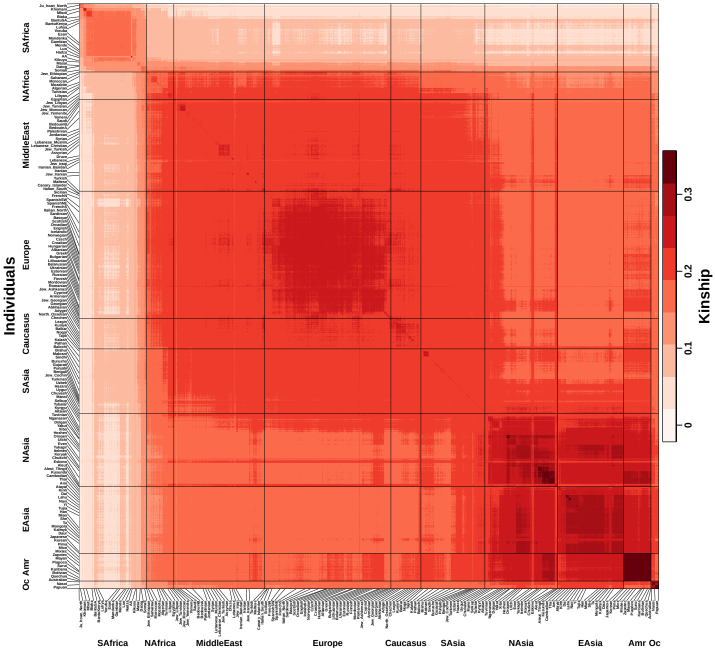
Statistical model:

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

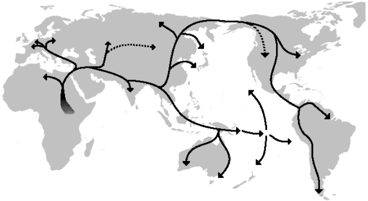
(Wright 1921, 1951; Malécot 1948; Jacquard 1970).

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

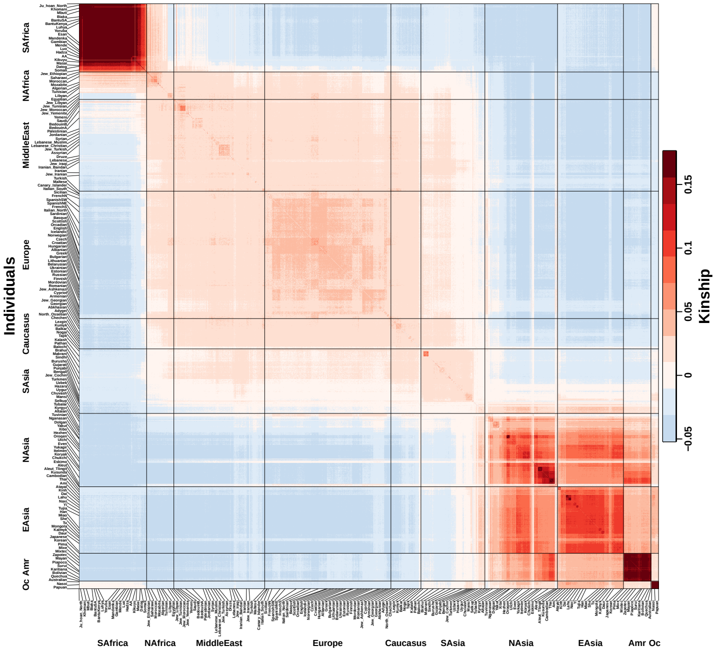
New kinship estimates



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

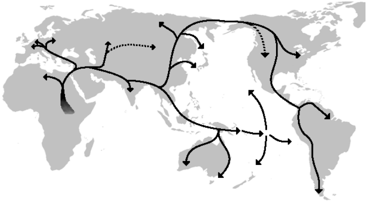


Edited from Ephert [CC BY-SA 3.0], via Wikimedia Commons



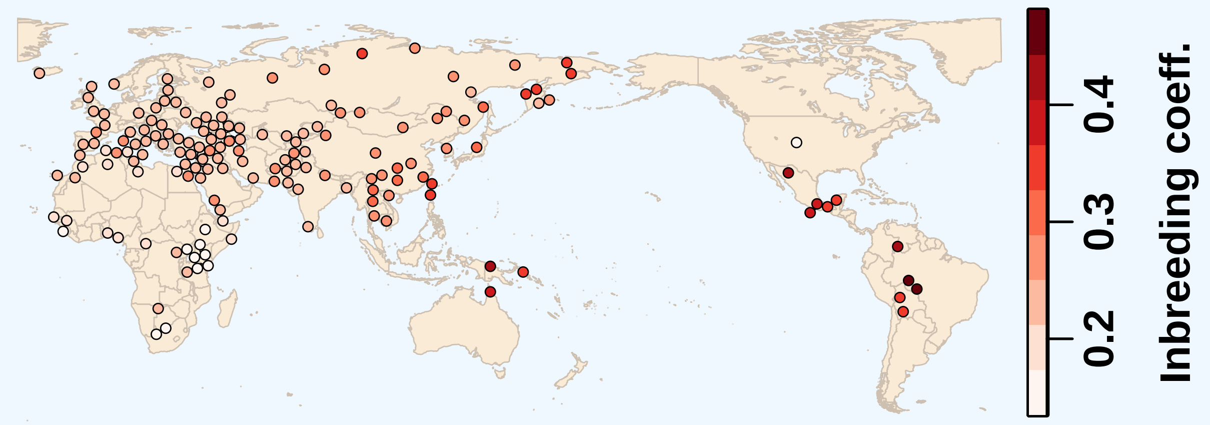
Standard kinship estimates

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

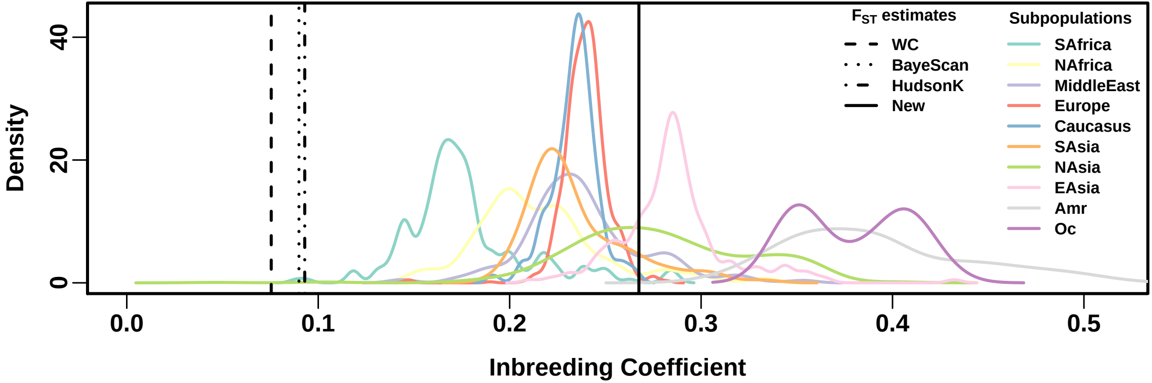


Edited from Ephert [CC BY-SA 3.0], via Wikimedia Commons

Population-level inbreeding

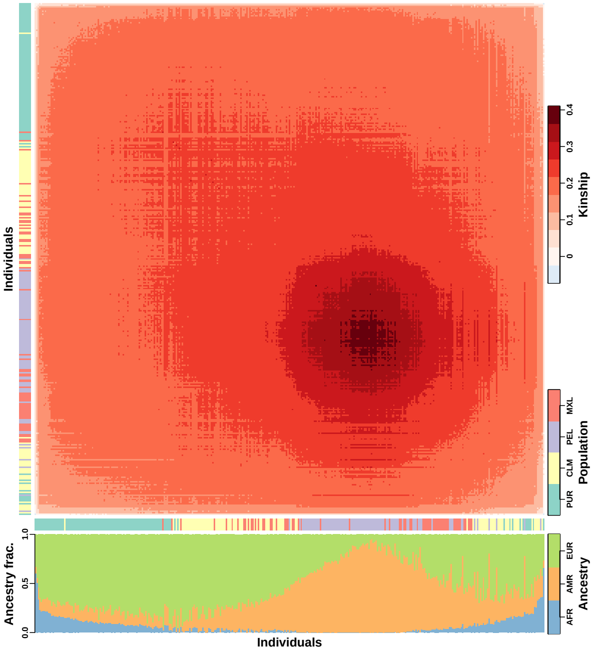


Differentiation (F_{ST}) previously underestimated



Generalized $F_{\mathbf{ST}}$: reduction in mean heterozygosity from ancestral population.
(Prev. F_{ST}: proportion of variation between **independent** subpopulations).

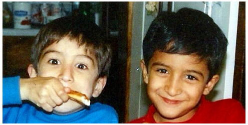
Kinship driven by admixture in Hispanics



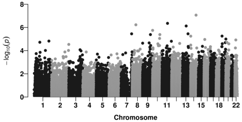
New kinship estimates

Genotypes from the 1000 Genomes Project (2012)

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

PRINCETON UNIVERSITY

Lewis-Sigler Institute for Integrative Genomics

Wright's F_{ST}

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:

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

The generalized F_{ST}

Need the new concept of 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:

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

Bias in F_{ST} estimators for independent subpopulations

Previous estimators are biased for n dependent subpopulations even when each subpopulation is infintely large (known AFs π_{ij}):

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

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

Bias in standard kinship estimator

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

$$\hat{p}_i^T = \frac{1}{2} \sum_{j=1}^n w_j x_{ij}, \quad \bar{\varphi}_j^T = \sum_{k'=1}^n w_{k'} \varphi_{jk'}^T, \quad \bar{\varphi}^T = \sum_{j'=1}^n \sum_{k'=1}^n w_{j'} w_{k'} \varphi_{j'k'}^T$$

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

Standard ancestral variance estimate also downwardly biased:

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

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.

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

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