

# Relatedness and differentiation in arbitrary population structures

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

🐦 DrAlexOchoa

🏠 [ochoalab.github.io](https://ochoalab.github.io)

✉️ [alejandro.ochoa@duke.edu](mailto:alejandro.ochoa@duke.edu)

# New kinship estimator for arbitrary population structures

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

## New kinship estimator for arbitrary population structures

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

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

# New kinship estimator for arbitrary population structures

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

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

Standard estimator is **biased**:

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

# New kinship estimator for arbitrary population structures

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

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

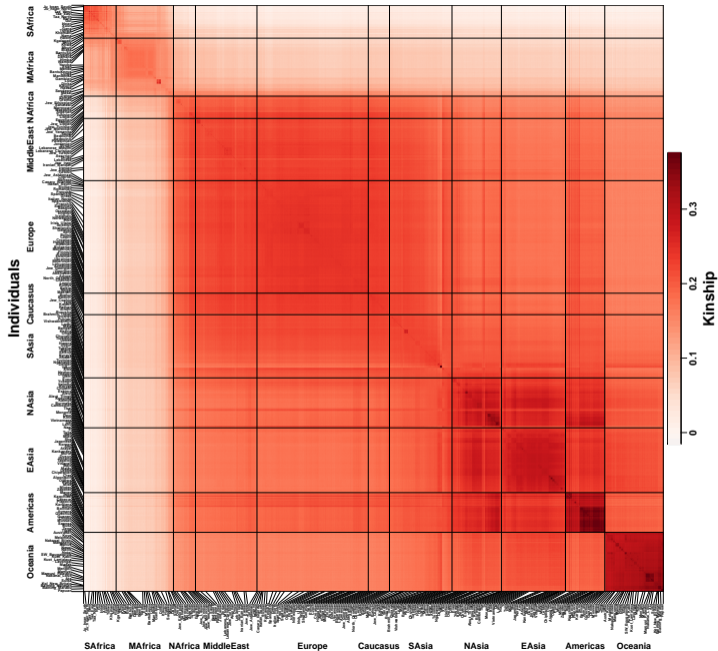
Standard estimator is **biased**:

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

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

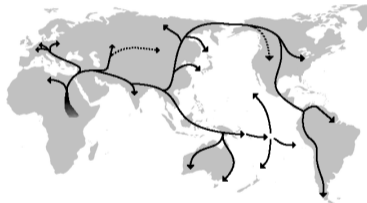
$$A_{jk} = \frac{1}{m} \sum_{i=1}^m (x_{ij} - 1)(x_{ik} - 1) - 1, \quad A_{\min} = \min_{u \neq v} \frac{1}{|S_u||S_v|} \sum_{j \in S_u} \sum_{k \in S_v} A_{jk},$$

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



# Our new kinship estimates

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

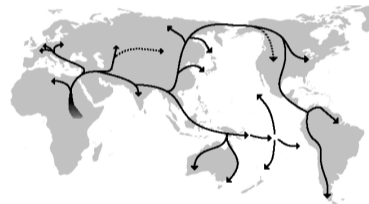
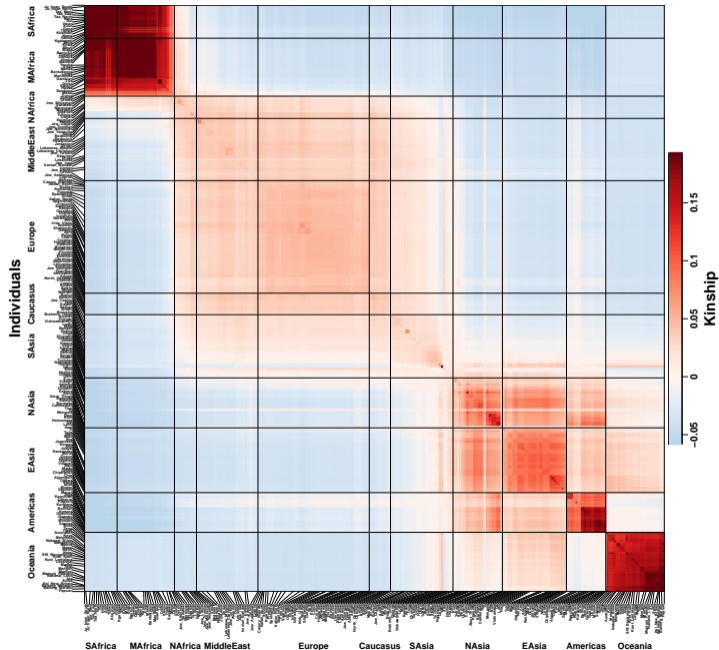


Edited from Ephert [CC BY-SA 3.0], via  
 Wikimedia Commons

\*Inbreeding coeffs. on diagonal

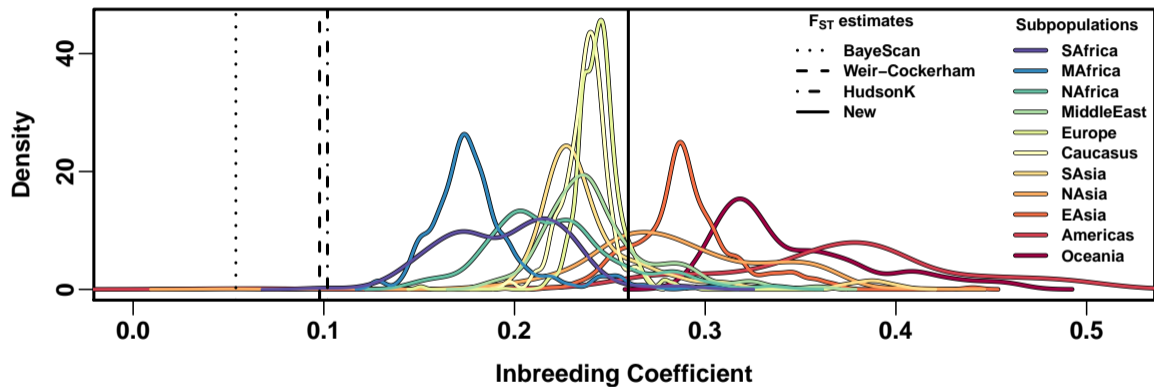
# Standard kinship estimates

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



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

# Differentiation ( $F_{ST}$ ) previously underestimated

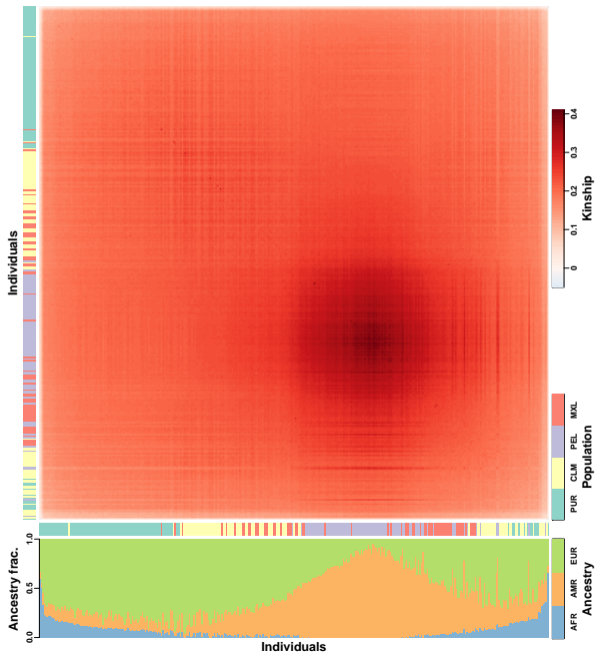




# Kinship driven by admixture in Hispanics

## Our new kinship estimates

Genotypes from the 1000 Genomes Project (2013)



# Better kinship estimates will improve GWAS!

¡Muchas gracias!

Princeton University

**John D. Storey**

Wei Hao

University of Warsaw

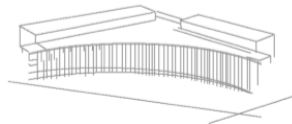
Neo Christopher Chung

Funding

National Institutes of Health

Otsuka Pharmaceutical

Duke StatGen Center  
Biostatistics & Bioinformatics



Lewis-Sigler Institute for Integrative Genomics