Relatedness and Differentiation in Arbitrary Population Structures

Alejandro Ochoa[1,2] and John D. Storey[3,4]

[1]Duke Center for Statistical Genetics and Genomics, and [2]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA
[3]Lewis-Sigler Institute for Integrative Genomics, and [4]Center for Statistics and Machine Learning, Princeton University, Princeton, NJ 08544, USA

Abstract

Important biomedical applications, including genome-wide association studies and heritability estimation for complex traits, require accurate modeling of the covariance structure of genetic variants. This dependence structure between individuals is parametrized by kinship coefficients, which are the probability that alleles at random loci are "identical by descent" (IBD). The fixation index F_{ST} is also an IBD probability that measures the overall population structure. My work is focused on extending current models and estimation approaches for kinship and F_{ST} to arbitrary population structures, where individuals are not assumed to belong to subpopulations that are disjoint, homogeneous, and statistically independent, such as African-Americans and Hispanics. Here I show how my approach improves upon previous approaches in real human datasets containing world-wide samples and in simulations where the true parameters are known. My work led to a novel kinship and F_{ST} estimation framework with greatly improved accuracy, implemented in the R package `popkin` available on CRAN. These results have direct implications in the estimation of heritability, association studies, and other analyses where population structure is a confounder. Research supported in part by NIH grant R01 HG006448.

Model

$x_{ij} \in \{0, 1, 2\}$: genotype of ind. j, biallelic SNP i, counting ref alellles. p_i: ancestral allele frequency. φ_{jk}: kinship coefficient. $f_j = 2\varphi_{jj} - 1$: inbreeding coefficient. Genotype moments:

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

Generalized F_{ST} for locally-outbred individuals [1]:

$$F_{\text{ST}} = \sum_{j=1}^{n} w_j f_j.$$

Standard kinship estimator

Biased when there is population structure [2, 3].

$$\hat{p}_i = \frac{1}{2n}\sum_{j=1}^{n} x_{ij}, \qquad \mathrm{E}\left[\hat{p}_i\left(1-\hat{p}_i\right)\right] = p_i\left(1-p_i\right)\left(1-\bar{\varphi}\right),$$

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

New kinship estimator: `popkin`

Derived for arbitrary population structures [2].

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

$$A_{\text{min}} = \min_{u \neq v} \frac{1}{|S_u||S_v|}\sum_{j\in S_u}\sum_{k \in S_v} A_{jk}, \qquad \hat{\varphi}_{jk} = 1 - \frac{A_{jk}}{A_{\text{min}}} \xrightarrow[m\to\infty]{\text{a.s.}} \varphi_{jk}.$$

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

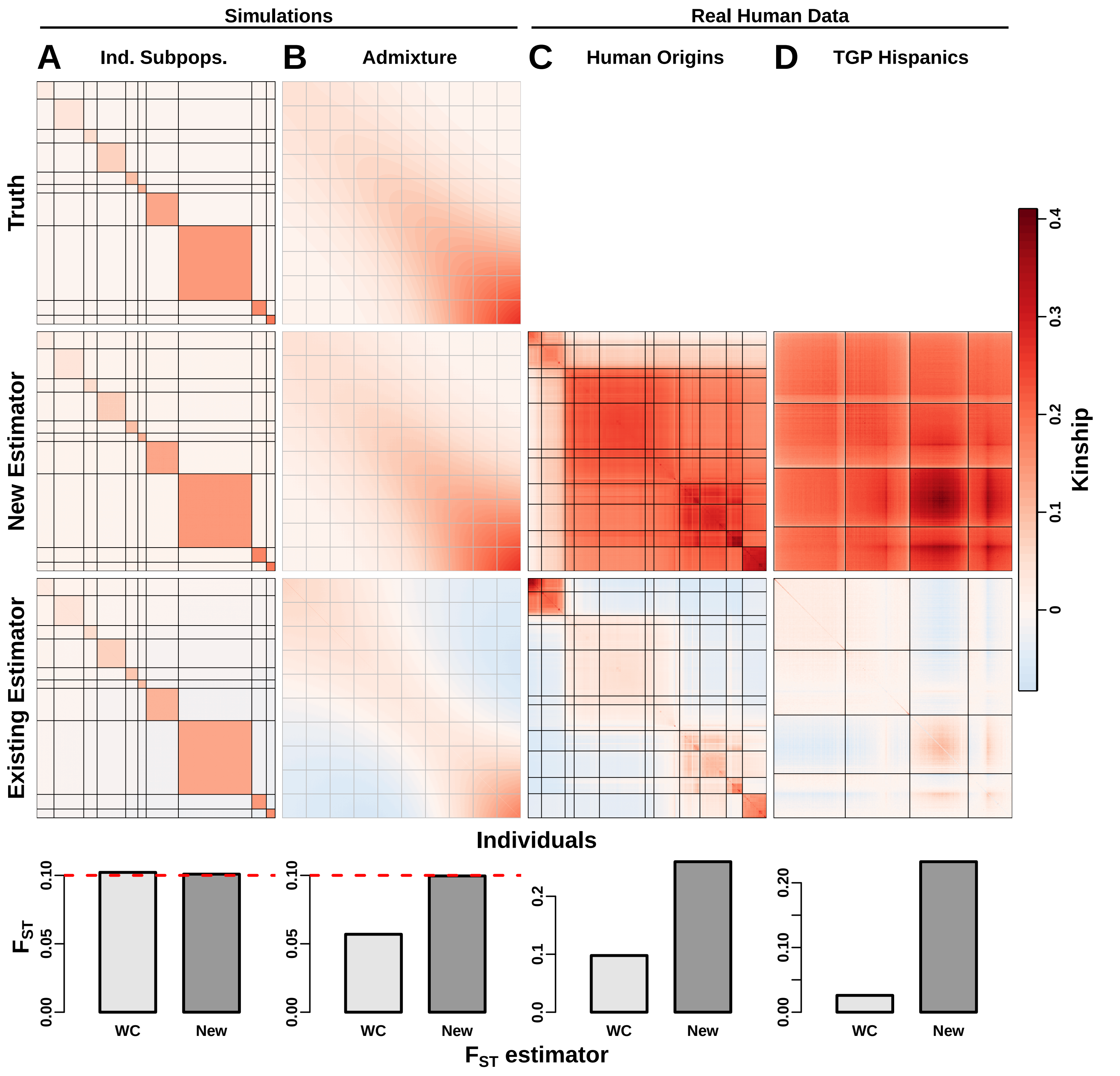

Figure 1: **New and existing kinship and F_{ST} estimates in simulations and real human data.** Rows: (1) true kinship matrix, (2) new kinship estimates, (3) standard kinship estimates, and (4) comparison of Weir-Cockerham to our new F_{ST} estimates and true F_{ST} (red dashed lines). Kinship matrix colors are φ_{jk} values between individual pairs j, k, and f_j along diagonal. Columns are datasets: **A.** Independent subpopulations simulation (assumed by existing F_{ST} estimators). **B.** Spatial admixture simulation (demonstrates biases in existing methods and superior performance of our new approach). **C.** Human Origins dataset (global native populations). **D.** Hispanic subset of 1000 Genomes Project.

Results and Conclusions

Our theory and simulations [1, 2] and real data analysis show that existing kinship and F_{ST} estimation approaches are biased in arbitrary population structures, and our new approach overcomes these limitations (Fig. 1). We estimate a complex population structure in native Human samples (Fig. 2) and a pattern of differentiation consistent with the Out-of-Africa model (Fig. 3). Our approach estimates a higher $F_{\text{ST}} \approx 26\%$ than existing approaches ($\approx 10\%$ or less) which assume independent subpopulations (Fig. 4).

References

[1] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". Submitted, preprint at `http://biorxiv.org/content/early/2016/10/27/083915`. 2016.

[2] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures II: Method of moments estimators". Submitted, preprint at `http://biorxiv.org/content/early/2016/10/27/083923`. 2016.

[3] Bruce S. Weir and Jérôme Goudet. "A Unified Characterization of Population Structure and Relatedness". *Genetics* (2017), genetics.116.198424.

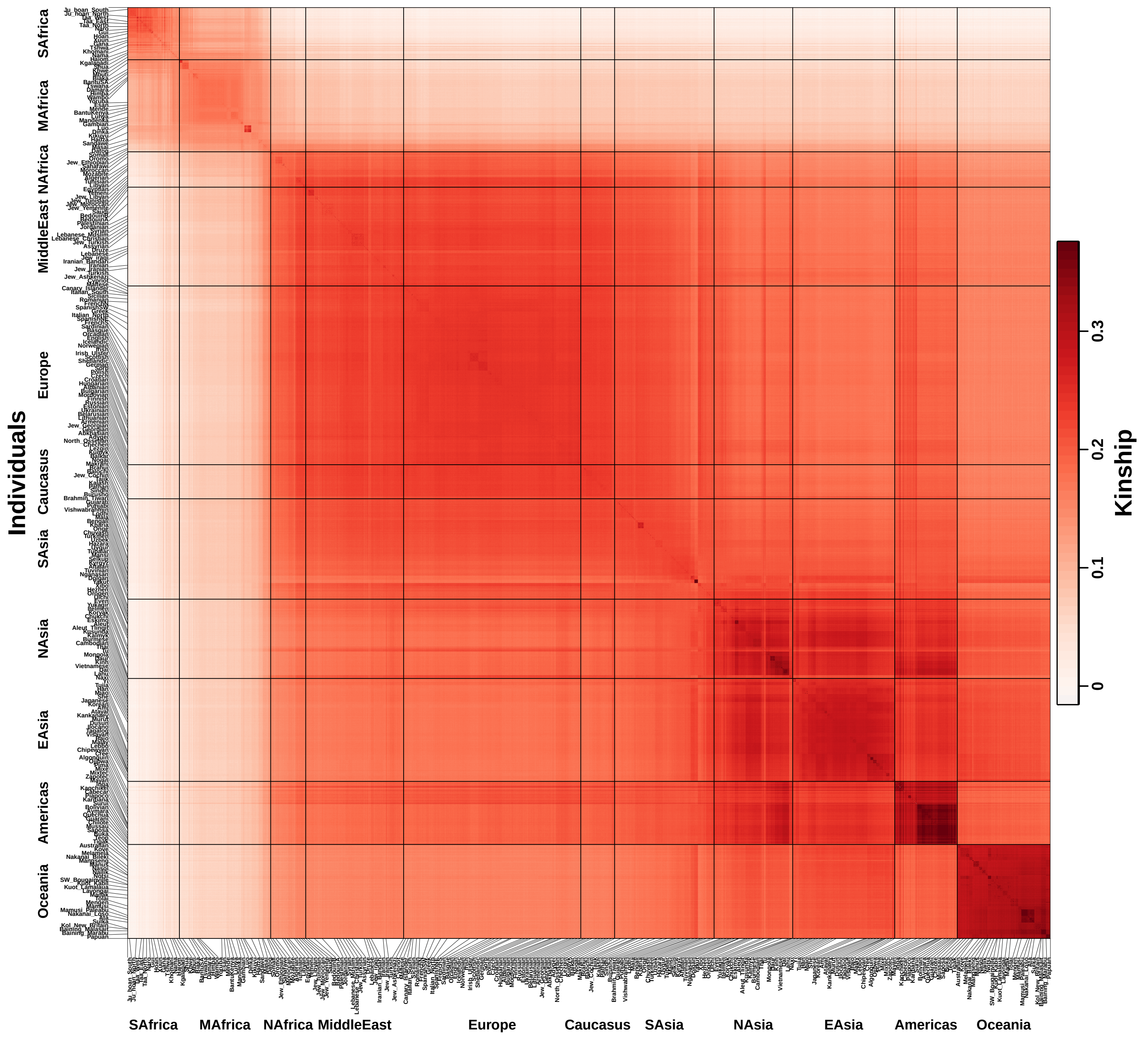

Figure 2: **Population-wide kinship estimates in Human Origins.** Our new estimates reveal substantial kinship between subpopulations and heterogeneity within subpopulations.

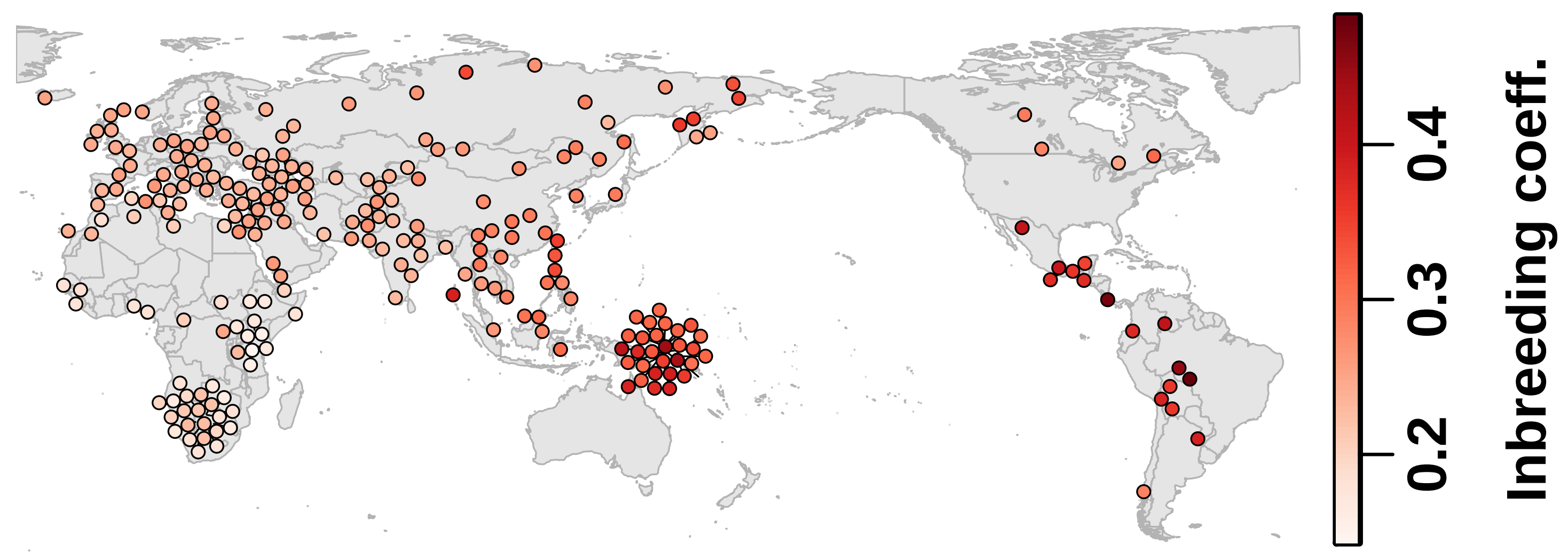

Figure 3: **Geography of population-level inbreeding.** Mean f_j increases with distance from Africa, as expected under the Out-of-Africa serial founder model.

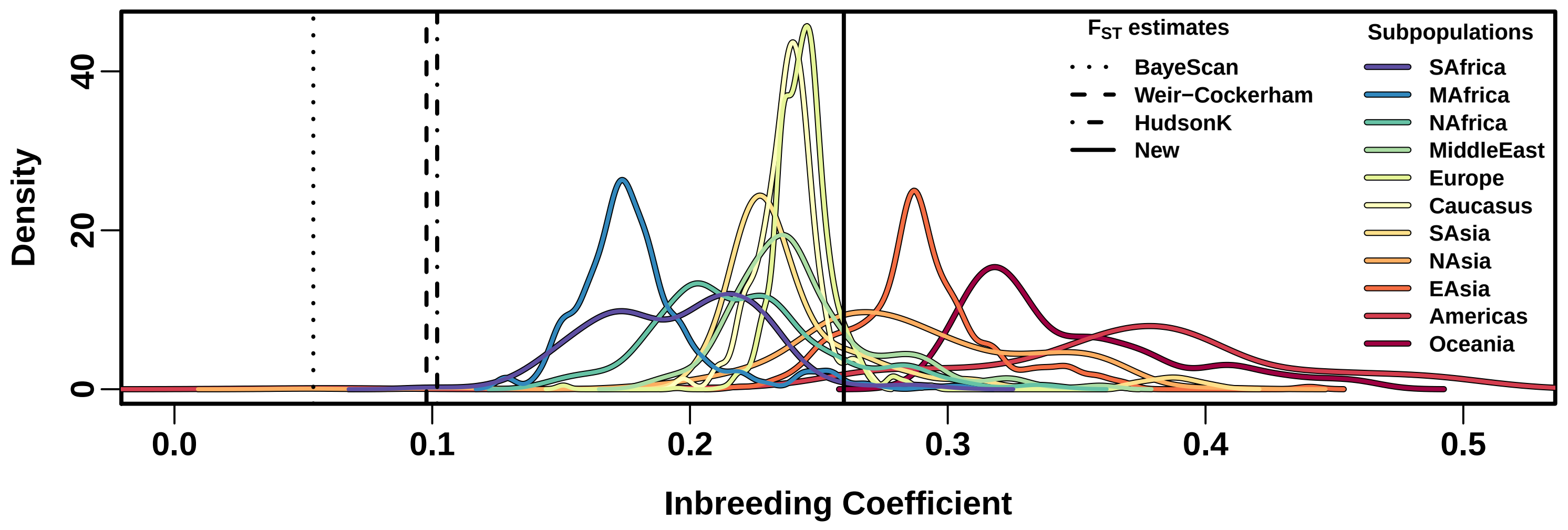

Figure 4: **Inbreeding and F_{ST} estimates in Human Origins.** Weir-Cockerham, HudsonK, and BayeScan assumed the $K = 11$ subpopulations are independent, which causes downward bias.