# New kinship and $F_{ST}$ estimates applied to the global human population

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

[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, Princeton University, Princeton, NJ 08544, USA

ASHG VIRTUAL MEETING 2020 OCTOBER 27-30

## Abstract

**Background:** Kinship coefficients and $F_{ST}$, which measure genetic relatedness and the overall population structure, respectively, have important biomedical applications. However, existing estimators are only accurate under restrictive conditions that most natural population structures do not satisfy. **Methods:** We derive new population kinship and $F_{ST}$ estimators for arbitrary population structures that are validated by theory and simulations, and are the only unbiased approaches currently available. Our approach is built from a new generalized $F_{ST}$ definition that, for the first time, is applicable to individuals without discrete subpopulations, yet agrees with the classical definition when there are subpopulations. We also characterize with theory and simulations existing approaches for kinship and $F_{ST}$ estimation, and find them to be severely downwardly biased in common settings such as admixture models and the presence of 3 or more subpopulations on a tree. All existing approaches misspecify the Most Recent Common Ancestor (MRCA) population, which admits negative kinship estimates and results in the observed bias. **Results:** Our estimates on various human datasets reveal a complex population structure driven by founder effects due to dispersal from Africa and admixture. Notably, our new approach estimates larger $F_{ST}$ values for worldwide human populations (31% on HGDP, 26% on Human Origins, and 22% on 1000 Genomes) than existing approaches (12%, 11%, and 8.5%, respectively, using the Weir-Cockerham estimator). We find that the two whole-genome sequencing datasets (HGDP and 1000 Genomes) agree on kinship for comparable subpopulations, whereas the microarray genotyping dataset Human Origins underestimates kinship outside of (and overestimates kinship within) Sub-Saharan Africa. Lastly, we showcase the generality of our approach by estimating kinship and $F_{ST}$ of 20% in admixed Hispanic individuals, who do not have discrete genetic subpopulations (a forced application of the Weir-Cockerham estimator yields $F_{ST} = 2.5\%$). While previous work correctly measured $F_{ST}$ between subpopulation pairs, our generalized $F_{ST}$ measures genetic distances among all individuals and their MRCA population, revealing that genetic differentiation is greater than previously appreciated. **Conclusion:** This analysis demonstrates that estimating kinship and $F_{ST}$ under more realistic assumptions is important for modern population genetic analysis. Potential improvements in applications that use kinship matrices include genetic association studies and heritability estimation.
Research supported in part by NIH grant R01 HG006448.

## Kinship model

$x_{ij} \in \{0, 1, 2\}$: genotype of ind. $j$, biallelic SNP $i$, counting ref alelles. $p_i$: ancestral allele frequency. $\varphi_{jk}$: 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]: $F_{ST} = \sum_{j=1}^{n} w_j f_j$.

## Standard kinship estimator

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

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

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

## New kinship estimator: `popkin`

Derived for arbitrary population structures, practically unbiased [2].

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

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

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

## Results and Conclusions

Our theory, simulations [1, 2], and real data analysis [4] show that existing kinship and $F_{ST}$ estimators are biased in arbitrary population structures, and our new approach overcomes these limitations.

We analyzed 3 global human datasets: Human Origins (Fig. 1), HGDP (Fig. 2), and 1000 Genomes (Fig. 3). Standard estimates are often negative and inconsistent with African origins. We estimate a complex population structure and differentiation patterns consistent with African origins. Our approach estimates higher $F_{ST}$ than existing approaches ($\approx 10\%$ or less) which assume independent subpopulations (Fig. 5). Human Origins underestimated kinship and $F_{ST} \approx 26\%$ due to inherent biases in its genotyping array platform (the other datasets are WGS). HGDP has largest $F_{ST} \approx 31\%$ since it has greater diversity than 1000 Genomes ($F_{ST} \approx 22\%$), but they agree in comparable subsets.

The Hispanic subset of 1000 Genomes demonstrates population kinship consistent with admixture, and $F_{ST} \approx 20\%$ estimation in the absence of discrete subpopulations (Fig. 4).

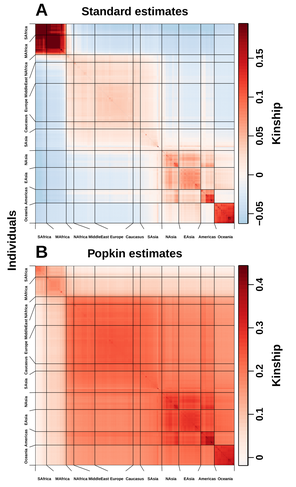


Figure 1:**Kinship in Human Origins.**

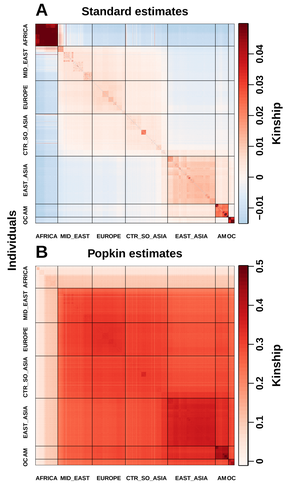


Figure 2:**Kinship in HGDP.**

Color = kinship for individual pairs (individuals along x and y axes), inbreeding coefficients along diagonal.

## References

[1] Alejandro Ochoa and John D. Storey. "$F_{ST}$ and kinship for arbitrary population structures I: Generalized definitions". *bioRxiv* (10.1101/083915) (2016). https://doi.org/10.1101/083915.

[2] Alejandro Ochoa and John D. Storey. "Estimating $F_{ST}$ and kinship for arbitrary population structures". *PLoS Genetics* (2019). In revision.

[3] Bruce S. Weir and Jérôme Goudet. "A Unified Characterization of Population Structure and Relatedness". *Genetics* 206(4) (2017), pp. 2085–2103.

[4] Alejandro Ochoa and John D. Storey. "New kinship and $F_{ST}$ estimates reveal higher levels of differentiation in the global human population". *bioRxiv* (10.1101/653279) (2019). https://doi.org/10.1101/653279.

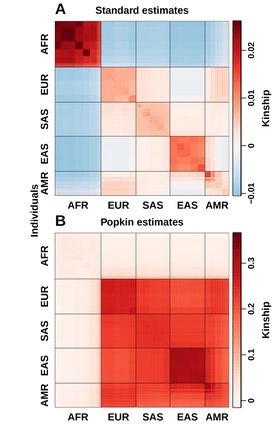


Figure 3:**Kinship in 1000 Genomes.**

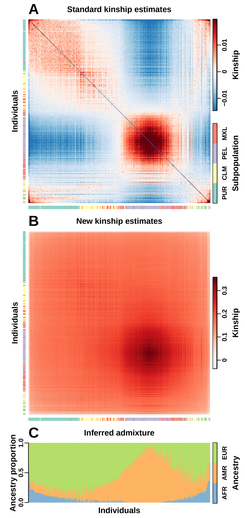


Figure 4:**Kinship of Hispanics in 1000 Genomes.**

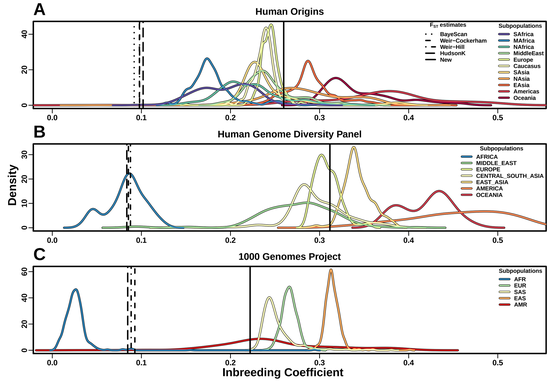


Figure 5:**Population inbreeding and $F_{ST}$ estimates in human datasets.** Weir-Cockerham, Weir-Hill, HudsonK, and BayeScan assumed the $K$ subpopulations are independent, which causes downward bias. **A.** Human Origins. $K = 11$. **B.** HGDP. $K = 7$. **C.** 1000 Genomes. $K = 5$.

Nothing to Disclose

American Society of Human Genetics