F_{ST} generalized for arbitrary population structures

Alejandro Ochoa and John Storey

Lewis-Sigler Institute for Integrative Genomics, Department of Molecular Biology, and Center for Statistics and Machine Learning, Princeton University

Summary

F_{ST} is a measure of differentiation between two or more populations. We generalize its definition and find large biases in existing estimators in this setting.

- Generalized F_{ST} for arbitrary populations, dropping need for clear population boundaries or homogeneity.
- Clarified connections between F_{ST} and probabilities of Identity-By-Descent (IBD): inbreeding, kinship, coancestry coefficients.
- Characterized the convergence properties of F_{ST} and kinship Method-of-Moment (MM) estimators.
- Calculated the F_{ST} of admixture models.

Introduction

A population is structured if its individuals do not mate randomly. Natural populations are structured due to distance and geography. Population structures confound association studies, since physically unlinked alleles may have correlated frequencies in subpopulations.

The inbreeding coefficient f_j is the probability that the two alleles of an individual j, at a random locus, were inherited from a single ancestor (IBD). The kinship coefficient φ_{jk} is the probability that two random alleles, one from each individual j, k, are IBD [1]. Note that $\varphi_{jj} = \frac{1+f_j}{2}$.

F_{ST} is the mean inbreeding coefficient in a subpopulation [1, 2]. The F_{ST} of many populations is the average F_{ST} of each population from their last common ancestor population.

The Weir-Cockerham (WC) F_{ST} estimator is a consistent (asymptotically unbiased) estimator for islands of different sample sizes, but assumed equal per-island F_{ST} [3]. The newer Hudson F_{ST} estimator is consistent for islands with differing F_{ST} values, but assumes random mating within islands [4]. All F_{ST} estimators assume independently-evolving populations.

Genotype model and moments

Let $x_{ij} \in \{0, 1, 2\}$ be the genotype of individual j on a biallelic SNP i, counting reference alleles. Alleles at i are drawn randomly with probability p_i (ancestral allele frequency). The moments given known relations f_j, φ_{jk} are [1, 2]

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

Generalized F_{ST} in terms of individuals

The individual analog of F_{ST} is $f^T_{L_j}$, the inbreeding between the ancestral population T and the local population L_j of j.

The generalized F_{ST} for a set of individuals is a summary of the individual parameters,

$$F_{\text{ST}} = \sum_j w_j f^T_{L_j},$$

where $\sum_j w_j = 1$ are arbitrary weights. This is backward-compatible with old island F_{ST}, using appropriate weights.

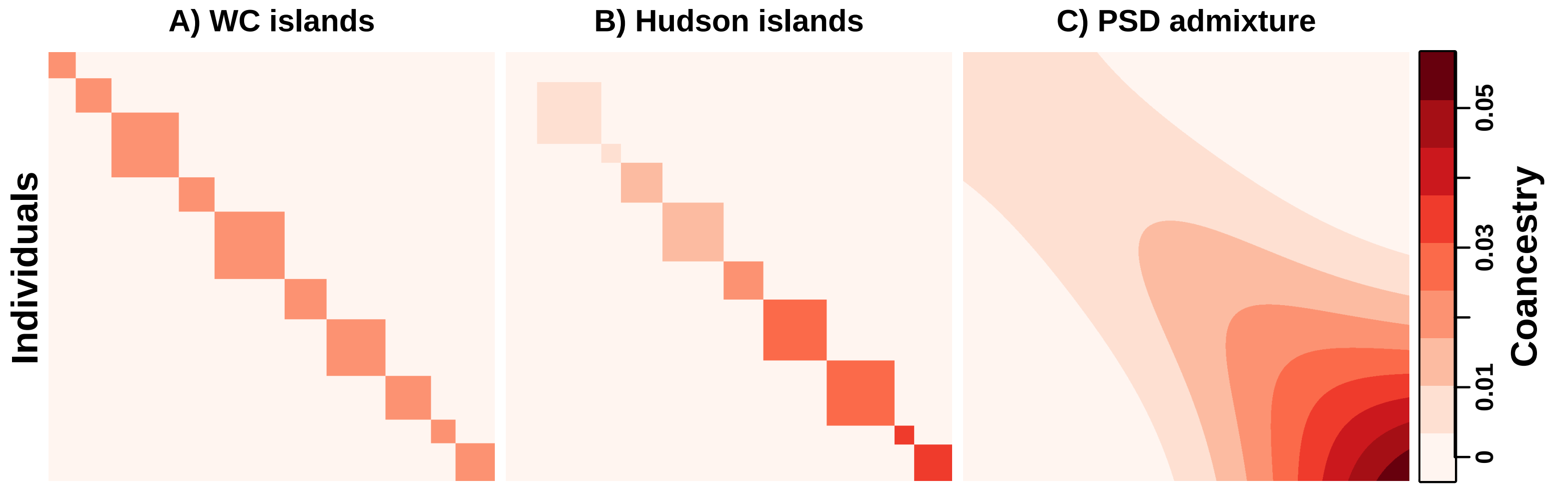

Figure 1: **Coancestry matrices of our simulations.** All simulations have $n = 1000$ individuals and $K = 10$ populations, and comparable F_{ST} (0.02 for the islands, 0.0187 for admixture). A) Weir-Cockerham islands have equal F_{ST} per island. B) Hudson islands has different F_{ST} per island. C) Extensive admixture and different F_{ST} per intermediate population.

Coancestry and individual allele frequencies

Let π_{ij} denote the individual allele frequency (IAF) of individual j at SNP i. The IAF moments are

$$\begin{aligned} \mathrm{E}[\pi_{ij}] &= p_i, \\ \mathrm{Cov}(\pi_{ij}, \pi_{ik}) &= p_i(1-p_i)\theta_{jk}, \\ x_{ij}|\pi_{ij} &\sim \mathrm{Binomial}(2, \pi_{ij}), \end{aligned}$$

where $\theta_{jk} \in [0, 1]$ are individual coancestry coefficients. This models locally outbred and locally unrelated individuals, and generalizes the model of [5]. Under these assumptions, kinship and coancestry are the same:

$$\theta_{jk} = \begin{cases} \varphi_{jk} & \text{if} \quad j \neq k, \\ 2\varphi_{jj} - 1 = f_j & \text{if} \quad j = k. \end{cases}$$

We also have an F_{ST} analogous to a previous definition [5]:

$$F_{\text{ST}} = \sum_j w_j \theta_{jj}.$$

Inconsistency in MM estimators

The "naive" MM coancestry estimator is, and converges to,

$$\hat{\theta}_{jk} = \frac{\sum_i (\pi_{ij} - \hat{p}_i)(\pi_{ik} - \hat{p}_i)}{\sum_i \hat{p}_i(1-\hat{p}_i)} \xrightarrow{\text{a.s.}} \frac{\theta_{jk} - \bar{\theta}_j - \bar{\theta}_k + \bar{\theta}}{1 - \bar{\theta}},$$

as the number of SNPs $m \to \infty$, where $\hat{p}_i = \sum_j w_j \pi_{ij}$, $\bar{\theta}_j = \sum w_j \theta_{jk}$ and $\bar{\theta} = \sum_j \sum_k w_j w_k \theta_{jk}$. So these estimates suffer from column- and row-specific distortions. The genotype version ($\pi_{ij} \to x_{ij}/2$) is a popular kinship estimator with similar biases. The "naive" F_{ST} estimator from $\hat{\theta}_{jj}$ is, and converges to,

$$\hat{F}_{\text{ST}} = \sum_j w_j \hat{\theta}_{jj} = \frac{\sum_i \sum_j w_j (\pi_{ij} - \hat{p}_i)^2}{\sum_i \hat{p}_i(1-\hat{p}_i)} \xrightarrow{\text{a.s.}} \frac{F_{\text{ST}} - \bar{\theta}}{1 - \bar{\theta}},$$

analogous to a previous result for populations [5]. Since $0 \leq \bar{\theta} \leq F_{\text{ST}}$, $\hat{F}_{\text{ST}}$ may be arbitrarily close to zero, even for large true F_{ST}. In practice $\bar{\theta}$ is unknown.

Simulation results

We constructed an admixture simulation that induces extreme biases in existing F_{ST} estimators (fig. 1). While the WC and Hudson F_{ST} estimators are unbiased under their respective models, they are indeed severely biased in our admixture model (fig. 2). Our simulation also illustrates the downward bias and gross distortions of estimated coancestries (and kinships) using the MM approach (fig. 3).

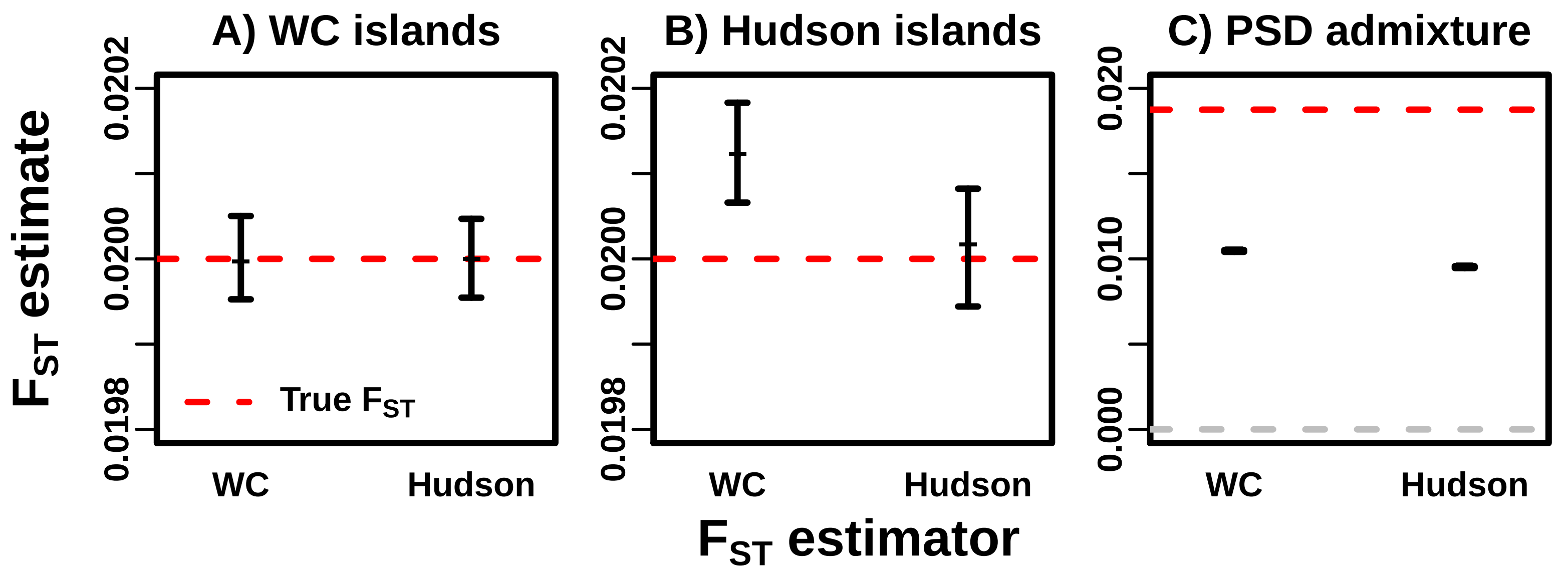

Figure 2: **WC and Hudson F_{ST} estimates are severely biased in admixture simulation.** WC and Hudson estimators evaluated on simulated genotypes. A) The island model assumed by WC. B) The island model assumed by Hudson. C) The admixture constructed to give a very biased F_{ST} estimator (note different y-axis scale). Bars are prediction intervals.

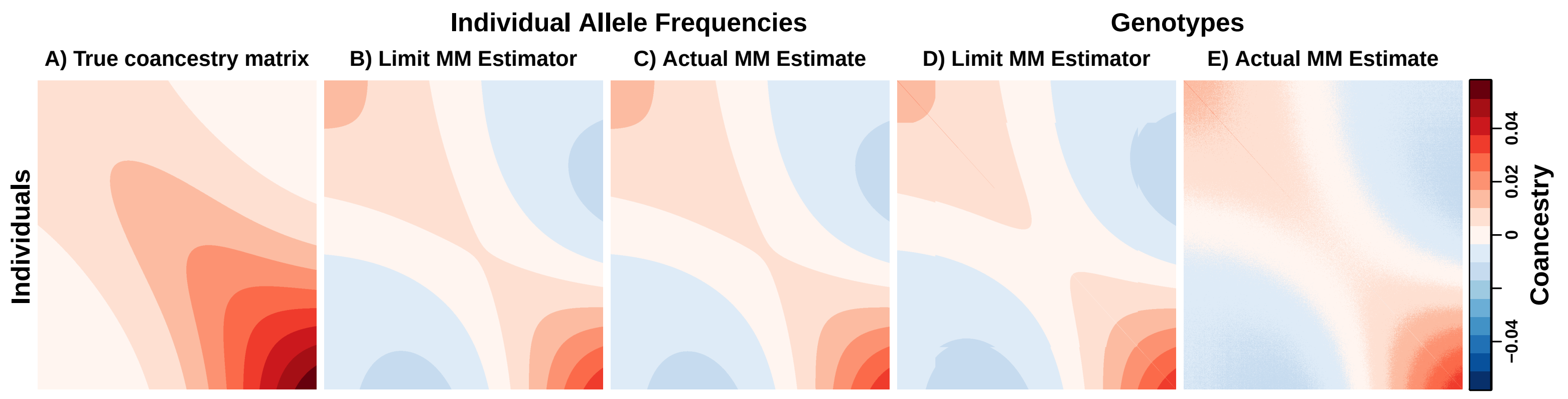

Figure 3: **Distorted MM estimates of coancestry.** Estimates of the coancestry matrix Θ for the 1000 admixed individuals of our simulation. Estimated coancestries from simulated IAFs and genotypes agree with the calculated limits for infinite SNPs. A) True Θ. B) Limit of $\hat{\Theta}$ from IAFs. C) $\hat{\Theta}$ from IAFs. D) Limit of $\hat{\Theta}$ from genotypes. E) $\hat{\Theta}$ from genotypes.

Conclusion

We generalized F_{ST} for arbitrary population structures. In this setting, popular MM-based estimators of F_{ST} and kinship/coancestry may be severely biased by arbitrary and unknown amounts. Since real populations are never independent islands, current F_{ST} estimates are actually loose lower bounds of the true F_{ST}. New methods are needed to estimate these quantities without bias.

Contact Information

ochoa@princeton.edu
http://viiia.org

References

[1] Gustave Malécot. Mathématiques de l'hérédité. *Masson et Cie*, 1948.

[2] Sewall Wright. The Genetical Structure of Populations. *Annals of Eugenics*, 15(1):323–354, 1951.

[3] B. S. Weir and C. Clark Cockerham. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358–1370, November 1984.

[4] Gaurav Bhatia, Nick Patterson, Sriram Sankararaman, and Alkes L. Price. Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23(9):1514–1521, September 2013.

[5] B. S. Weir and W. G. Hill. Estimating F-Statistics. *Annual Review of Genetics*, 36(1):721–750, 2002.

PRINCETON UNIVERSITY

CENTER FOR STATISTICS AND MACHINE LEARNING