New kinship and F_{ST} estimates reveal higher levels of differentiation in 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

∗ Corresponding authors: alejandro.ochoa@duke.edu and jstorey@princeton.edu

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. We recently derived new kinship and F_{ST} estimators for arbitrary population structures [1, 2]. Our estimates on 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 of 26% for native worldwide human populations and 23% for admixed Hispanic individuals, whereas the existing approach estimates 9.8% and 2.6%, respectively. While previous work correctly measured F_{ST} between subpopulation pairs, our generalized F_{ST} measures genetic distances among all individuals and their most recent common ancestor (MRCA) population, revealing that genetic differentiation is greater than previously appreciated. This analysis demonstrates that estimating kinship and F_{ST} under more realistic assumptions is important for modern population genetic analysis.

Kinship coefficients and F_{ST} are defined as probabilities of identity-by-descent [3–5]. Kinship matrices are crucial for accurate inference under population structure in many important biomedical applications, including genome-wide association studies [6–13] and heritability estimation [14, 15]. However, the most commonly-used standard kinship estimator [9, 10, 13–19] is accurate only in the absence of population structure [2, 20]. Likewise, current F_{ST} estimators assume that individuals are partitioned into statistically-independent subpopulations [4, 5, 21–23], which does not hold for human and other complex population structures. The human genetic population structure is remarkably complex, shaped by geography and population bottlenecks in migrations out of Africa [24–34] and admixture events [35–39]. We use human data to illustrate the improvements provided by our new approach.

Models and methods. Our new kinship and F_{ST} estimators were derived assuming arbitrary population structures, and they yield nearly unbiased estimates [2]. Suppose there are n individuals genotyped at m biallelic autosomal loci, such as SNPs. Our kinship estimator $\hat{\varphi}^{\text{new}}_{jk}$ is given by

$$\begin{aligned} A_{jk} &= \frac{1}{m} \sum_{i=1}^{m} (x_{ij} - 1)(x_{ik} - 1) - 1, \\ \hat{A}_{\text{min}} &= \min_{u \neq v} \frac{1}{|S_u||S_v|} \sum_{j \in S_u} \sum_{k \in S_v} A_{jk}, \\ \hat{\varphi}^{\text{new}}_{jk} &= 1 - \frac{A_{jk}}{\hat{A}_{\text{min}}}, \end{aligned}$$

where the genotypes $x_{ij} \in \{0, 1, 2\}$ count the number of reference alleles at locus i for individual j. For simplicity, here $\hat{A}_{\text{min}}$ uses a partition of individuals into subpopulations S_u for $u \in \{1, ..., K\}$ used solely to estimate the minimum kinship, which is set to zero ($\hat{\varphi}^{\text{new}}_{jk}$ has individual-level resolution; the general framework does not need subpopulations [2]). Under our model $\mathrm{E}[A_{jk}] = (\varphi_{jk} - 1)v$ contains the desired kinship coefficient φ_{jk} and a nuisance parameter v shared by all individuals. Assuming zero kinship across the two least related individuals, $\mathrm{E}[\hat{A}_{\text{min}}] \approx \min_{j,k} \mathrm{E}[A_{jk}] = -v$ yields the nuisance parameter, enabling consistent kinship estimates: $\hat{\varphi}^{\text{new}}_{jk} \xrightarrow[m \to \infty]{\text{a.s.}} \varphi_{jk}$.

We compare to the widely-used standard kinship estimator

$$\hat{\varphi}^{\text{std}}_{jk} = \frac{1}{m} \sum_{i=1}^{m} \frac{(x_{ij} - 2\hat{p}_i)(x_{ik} - 2\hat{p}_i)}{4\hat{p}_i (1 - \hat{p}_i)}, \quad \hat{p}_i = \frac{1}{2n} \sum_{j=1}^{n} x_{ij},$$

which has a complex bias non-linear in φ_{jk} in structured populations [2, 20]. The limit of $\hat{\varphi}^{\text{std}}_{jk}$ as the number of loci $m \to \infty$ is well-approximated by

$$\frac{\varphi_{jk} - \bar{\varphi}_j - \bar{\varphi}_k + \bar{\varphi}}{1 - \bar{\varphi}},$$

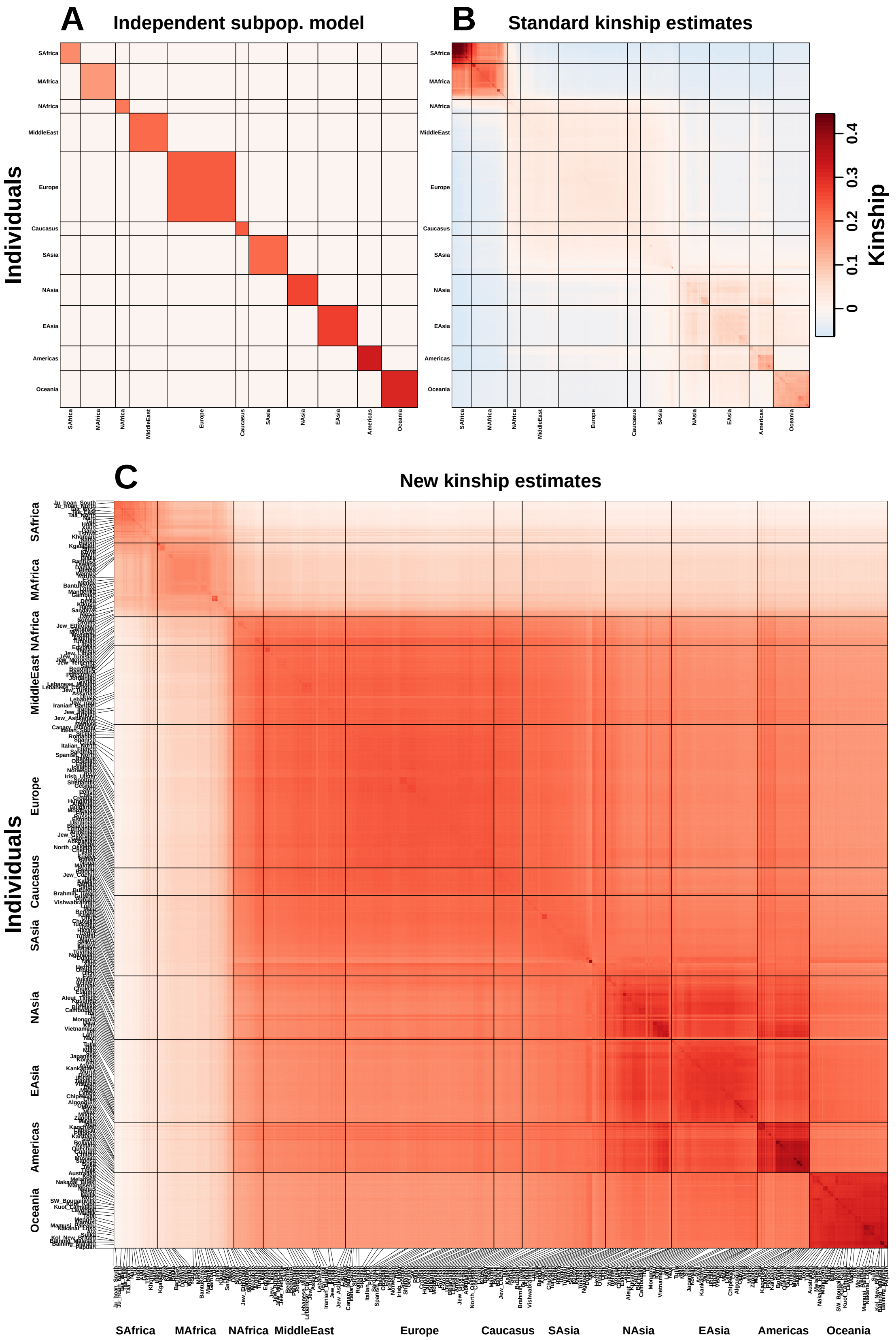
where $\bar{\varphi}_j = \frac{1}{n} \sum_{k'=1}^{n} \varphi_{jk'}$ is the mean kinship of individual j with all others and $\bar{\varphi} = \frac{1}{n^2} \sum_{j'=1}^{n} \sum_{k'=1}^{n} \varphi_{j'k'}$ is the overall mean kinship in the data [2]. This estimator is widely-used in approaches for structured populations, including genetic association studies and heritability estimation [9, 10, 13–19].

The original F_{ST} measures inbreeding in a subpopulation relative to an ancestral population [4], excluding local inbreeding if present [5]. Existing approaches estimate the mean F_{ST} between two or more independent subpopulations relative to their MRCA population [21, 23, 40], but have a downward bias otherwise [2]. In our new approach, inbreeding coefficients are estimated from kinship (measured from the MRCA population) using $\hat{f}^{\text{new}}_j = 2\hat{\varphi}^{\text{new}}_{jj} - 1$ and the generalized F_{ST} is estimated using $\hat{F}^{\text{new}}_{\text{ST}} = \sum_{j=1}^{n} w_j \hat{f}^{\text{new}}_j$ (valid for locally-outbred individuals [1]), where w_j are weights to account for geographically-imbalanced sample sizes. Our generalized F_{ST} does not require subpopulations, is the first to be applicable to arbitrary population structures [1], and our estimator is accurate in this setting [2]. We compare to the existing F_{ST} estimators Weir-Cockerham [21], HudsonK (for two subpopulations [23] generalized in [2]), and BayeScan [40]. These estimators assume independent subpopulations and homogeneous inbreeding within subpopulations, which causes downward biases in more complex population structures and implicitly admit negative kinship coefficients [2]. The classical F_{ST} interpretation—the proportion of variance explained by differences between subpopulation pairs—is not appropriate when subpopulations are not independent, and is not clearly defined in the absence of obvious subpopulations (such as for admixed individuals). Instead, our generalized F_{ST} measures the genetic drift of individuals from the MRCA population, which ensures valid underlying kinship coefficients [1].

Results. We first analyze the Human Origins datasets of native populations [41–43], which consists of **2922 individuals** from **243 sub-subpopulations** grouped into **11 subpopulations** (Supplementary Information). Sub-subpopulation abbreviations are defined in [41–43], while the subpopulation labels are defined in Fig. S1 (Supplementary Information). If these subpopulations were independent and internally unstructured, as assumed by existing F_{ST} estimators, the kinship matrix would have zero values between subpopulations and equal kinship within subpopulations (Fig. 1A). Instead, our approach reveals substantial kinship between subpopulations and heterogeneity within subpopulations (Fig. 1C).

The kinship matrix of Fig. 1C can be interpreted under the African origins serial founder model, as follows. Recall that a population size reduction (bottleneck) increases kinship and F_{ST} relative to the ancestral population [3–5]. The first population split occured roughly between individuals from Sub-Saharan African (KhoeSan-speaking hunter-gatherers (`SAfrica`) and Bantu-speakers and other agro-pastoralists (`MAfrica`)) and individuals outside of Sub-Saharan Africa. This split resulted in bottlenecks that increased kinship in each side relative to the ancestral value (which equals the

Figure 1 *(following page)*: **Population-wide kinship estimates in Human Origins.** As a visual aid, individuals are arranged into a hierarchy with subpopulations (rough continental clusters, i.e., `SAfrica`) and sub-subpopulations (locations potentially separated by ethnicity or religion, i.e., `Lebanese_Christian`). However, we estimate individual-level kinship without using this hierarchy. Color corresponds to kinship (φ_{jk}) for every pair of individuals j (x-axis) and k (y-axis) and inbreeding coef. (f_j) along the diagonal. **A.** Kinship matrix assumed by the independent subpopulations model prevalent in F_{ST} estimation: fixed φ_{jk} within subpopulations, $\varphi_{jk} = 0$ between subpopulations. **B.** Biased standard kinship estimates. The overall downward bias causes many negative estimates (blue) and strong distortions across the matrix (incorrectly assigns highest kinship within `SAfrica`-`MAfrica`), as predicted by our theory. Also note comparable kinship estimates between each of `MAfrica`, `Europe`, and `EAsia`, which contradicts the African origins model. **C.** New kinship estimates. Our new estimates reveal substantial kinship between subpopulations and heterogeneity within subpopulations. For an improved dynamic range, all displayed f_j, φ_{jk} values in panels B and C were capped at the 99 percentile of the estimated f_j values of panel C (full f_j distribution in Fig. 3). Additionally, panel B was capped below to the 1 percentile of its distribution.


A
Independent subpop. model
B
Standard kinship estimates
Individuals
SAfrica
MAfrica
NAfrica
MiddleEast
Europe
Caucasus
SAsia
NAsia
EAsia
Americas
Oceania
Kinship
0
0.1
0.2
0.3
0.4
C
New kinship estimates
Individuals
SAfrica
MAfrica
NAfrica
MiddleEast
Europe
Caucasus
SAsia
NAsia
EAsia
Americas
Oceania

kinship between the two subpopulations). The next split was roughly between West Eurasians and the rest, again increasing kinship within each side. Among West Eurasians, kinship is higher within `Europe`, reflecting another bottleneck. `Americas` (Native Americans) and `Oceania` have the highest kinship values within, reflecting further bottlenecks in their trek out of Africa. Note that the European admixture in `Americas` (calculated in Supplementary Information) is evident in individuals with lower kinship relative to other `Americas` individuals and greater kinship with `Europe` (Fig. 1C). Overall, our observations are coherent with previous work [27, 30, 31, 35], but our approach is the first to use a nearly unbiased estimator of kinship coefficients under assumptions aligned with the data. Our approach accurately estimates kinship at individual-level resolution and successfully uncovers a complex population structure where individuals may be related to each other in arbitrary ways.

The MRCA population of living humans is estimated to have existed in Africa 100-200K years ago [26, 27, 34], which first split into the ancestral KhoeSan population (who speak so-called "click" languages of the Kx'a, Tuu, and Khoe families, grouped into `SAfrica`) and the rest [26, 27, 31, 32, 34]. This MRCA population excludes ancestry from the Neanderthal and Denisovan introgressions [36, 37], but their limited contribution makes it a reasonable approximation. In our estimates, the minimum per-sub-subpopulations mean kinship is between `Ju_Hoan_North` (`SAfrica`) and `Kol_New_Britain` (`Oceania`). Moreover, the 2114 pairs with the smallest kinship values all consisted of pairs where one sub-subpopulation was from `SAfrica` (most commonly `Ju_Hoan_North` and `Ju_Hoan_South`) and the other was from outside of `SAfrica` and `MAfrica`. Therefore, we infer the first population split to have been between the ancestral KhoeSan population (`SAfrica`) and the rest, agreeing with previous work using independent mtDNA [26], Y chromosome [34], and microsatellite data [27, 32] (not used by our approach), as well as SNPs [31]. High kinship between `SAfrica` and `MAfrica` (Fig. 1C) suggests recent admixture [32] or an isolation-by-distance structure [44].

The diagonal of the kinship matrix of Fig. 1C contains inbreeding coefficients f_j, which are individual-specific F_{ST} values (for locally-outbred individuals, which most humans are). Every individual is differentiated (first percentile $f_j = 0.149$, where the zero value would correspond to the f_j of the child of the two most unrelated individuals in the data), and differentiation increases with distance from southern Africa (shown geographically in Fig. 2 and using distributions in Fig. 3), as expected under the African origins model and agreeing with previous work [26, 27, 29–34]. Remarkably, our estimated F_{ST} of 0.260 is substantially larger than estimates around 0.098 from existing approaches (Fig. 3) and previous measurements based on F_{ST} [30, 45] or related variance component models [31, 46, 47] — except for some AMOVA ϕ_{ST} estimates [48] (pairwise F_{ST} estimates [23, 49–52] are not generally comparable to our estimate). Existing approaches underestimate F_{ST} because they assume zero kinship between subpopulations, clearly incorrect as seen in Fig. 1C, whereas our new approach models arbitrary kinship between individuals and leverages kinship to estimate F_{ST}.

The popular standard kinship estimator [9, 10, 13–19] has a nonlinear bias in structured pop-

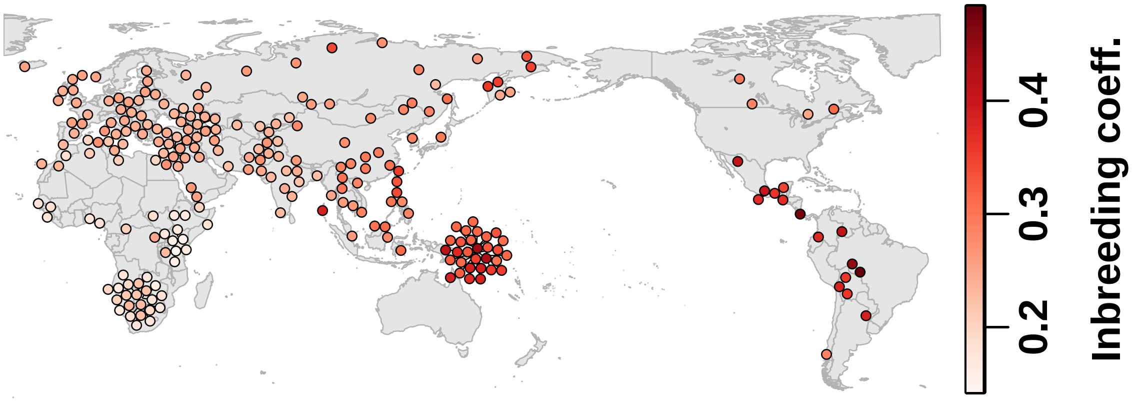


Figure 2: **Geographical distribution of population-level inbreeding.** Colors in circles denote the mean individual inbreeding f_j within each Human Origins sub-subpopulation. These mean f_j values increase smoothly with distance from southern Africa, as expected under the African origins serial founder model.

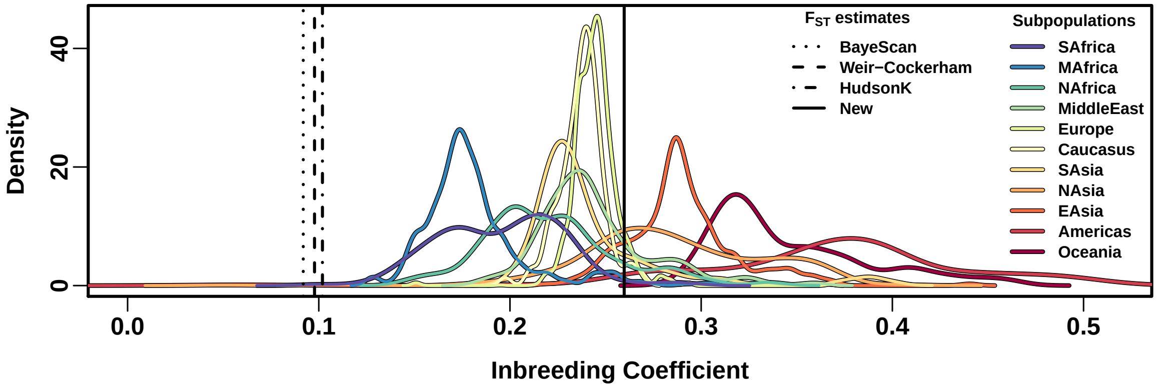


Figure 3: **Inbreeding and F_{ST} estimates in Human Origins.** Our new approach yields individual f_j inbreeding values and a generalized F_{ST} estimate, which is their weighted mean. Per-subpopulation f_j distributions show increasing values with distance from Africa (where densities are estimated with equal sub-subpopulation weights). Weir-Cockerham, HudsonK, and BayeScan assumed the $K = 11$ subpopulations are independent (Fig. 1A), which causes downward bias.

ulations [2, 20]. Standard kinship matrix estimates have abundant negative values and strong distortions (Fig. 1B *versus* our estimates in Fig. 1C, direct comparison in Fig. S2A). These estimates disagree with the African origins model, assigning greater kinship within `SAfrica-MAfrica` than to any other subpopulation, and comparable kinship between `Europe`, `EAsia` and `MAfrica` (incorrect since `MAfrica` split first [23]). The biases in the standard kinship and existing F_{ST} estimators are both fundamentally due to assuming that most kinship values are zero, and explicitly or implicitly admit negative kinship values [2, 22]. Our new kinship estimator is developed for arbitrary population structures and yields more interpretable estimates in human data.

Hispanic Latin Americans have a complex population structure, being recently admixed from European (`EUR`), Native American (`AMR`), and Sub-Saharan African (`AFR`) populations [53–57]. Here we show that Hispanics in the 1000 Genomes Project (TGP) do not have discrete subpopulations, so the classical F_{ST} definition does not apply. Using our approach, we estimate the kinship matrix of the 347 TGP Hispanic individuals (`PUR`: Puerto Rican; `CLM`: Colombian; `PEL`: Peruvian; `MXL`: Mexican-American; standard kinship in Fig. 4A, our new approach in Fig. 4B). We also estimate individual-specific admixture proportions of `EUR`, `AMR`, and `AFR` ancestry (Fig. 4C), detailed in Supplementary Information. We confirm previous observations that relatedness in Hispanics varies along a continuum driven by admixture [53–57]. In particular, since differentiation increases from `AFR` to `EUR` to `AMR` (Fig. 3), the greatest kinship is between individuals with higher `AMR` ancestry, and the lowest kinship is between individuals with higher `AFR` ancestry (Fig. 4B and C). Standard kinship estimates are also biased and distorted in Hispanics (Fig. 4A and Fig. S2B) and lack the interpretability of our estimates. Lastly, our approach estimates F_{ST} to be 0.233, which is comparable to f_j estimates for `MAfrica` and `Europe` in Human Origins (Fig. 3). In contrast, Weir-Cockerham estimates an unrealistically small F_{ST} of 0.0260, which is downwardly biased because it requires subpopulation labels (we used sampling locations: `PUR`, `CLM`, `PEL`, `MXL`; see the thin colored bar outside each kinship matrix in Fig. 4A and B), which erases the considerable substructure within subpopulations and models the large kinship values between individuals of different subpopulations as zero. We emphasize that existing F_{ST} approaches were designed for and require non-overlapping, independently evolving subpopulations, so they do not apply to individuals with variable admixture proportions such as Hispanics (as shown here) and African Americans [32, 58]. Our results are not a critique of those important advances, but a demonstration that modern data requires new estimators. Our approach overcomes these challenges by estimating kinship without assuming subpopulations (Supplementary Information) and estimating F_{ST} from these individual-specific parameters.

Conclusion. Our analysis of the Human Origins and 1000 Genomes Project datasets reveals a complex population structure with predominantly non-zero kinship values that vary along a continuum. Our new approach does not artificially partition individuals into subpopulations, which enables us to capture population structure with individual-level resolution and for the first time

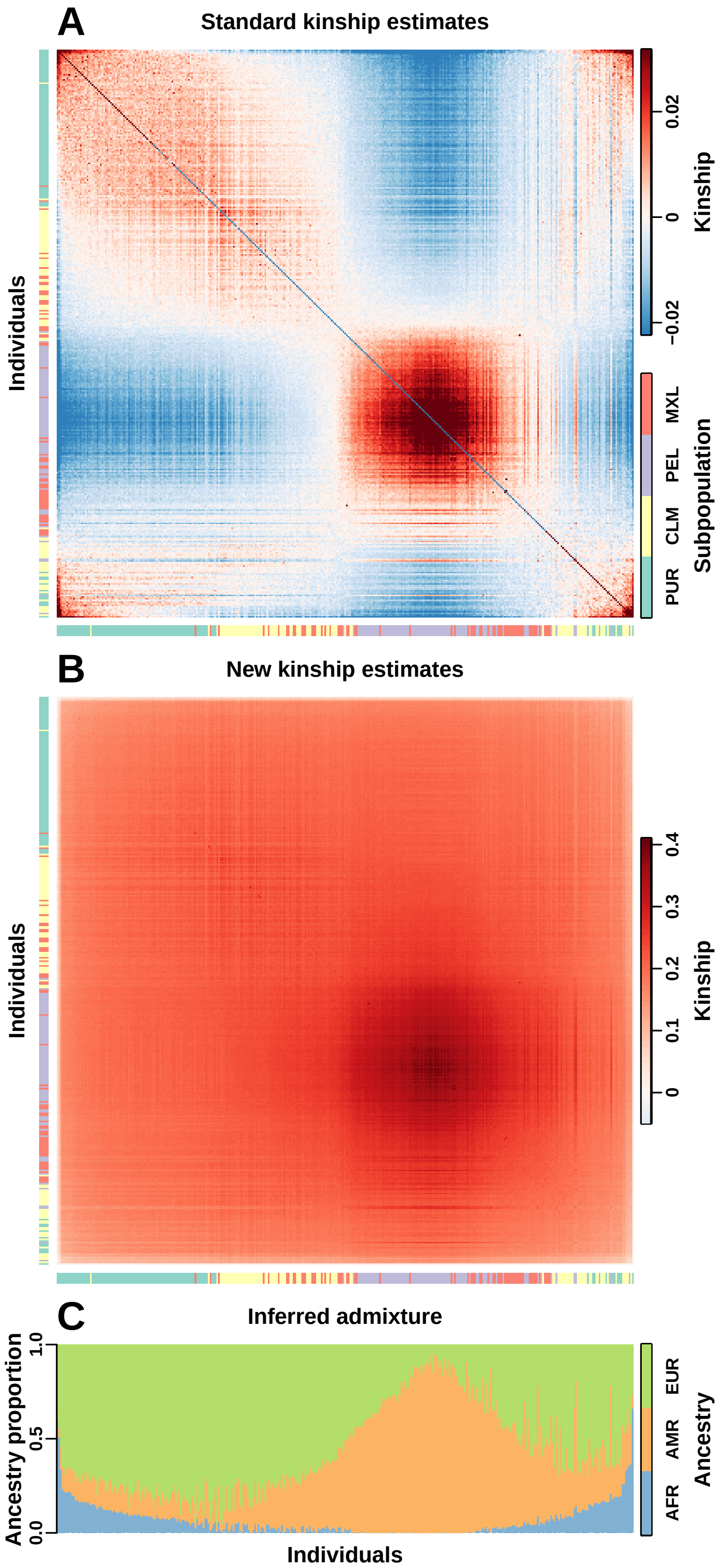


Figure 4: **Kinship of Hispanics in 1000 Genomes.** The colors in the kinship heatmaps correspond to φ_{jk} kinship values for every pair of individuals j (x-axis) and k (y-axis), and f_j along the diagonal. The color bars outside kinship matrices mark the subpopulation (sampling location) of every individual, which is poorly correlated with kinship (panel B) or admixture proportions (panel C). **A**: Biased standard kinship estimates. Note the overall downward bias causes negative estimates (blue), discontinuities between f_j (diagonal) and φ_{jk} (off-diagonal), and distortions (e.g., relatively higher kinship between individuals with higher AFR ancestry). Displayed φ_{jk}, f_j values are capped to the 1 and 99 percentiles of their distribution. **B**: New kinship estimates reveal a smooth population structure without discrete subpopulations, and much greater kinship values (note the different color scales for panels A and B). Individuals were ordered using seriation, which places low φ_{jk} away from the diagonal [59, 60]. **C**: Admixture proportions of every individual for Sub-Saharan African (AFR), European (EUR), and Native American (AMR) ancestry (calculated in Supplementary Information).

yields accurate population-level kinship and F_{ST} estimates for world-wide human SNP data. New approaches that make minimal assumptions about relatedness and structure are necessary for many biomedical applications—including genetic association studies in multiethnic panels and admixed individuals—and our new framework provides the foundation that enables this goal.

Data and Software. This approach is implemented in the R package `popkin` available online (`https://cran.r-project.org/package=popkin` and `https://github.com/StoreyLab/popkin`). Public data and code reproducing these analyses are available at
`https://github.com/StoreyLab/human-differentiation-manuscript`.

Acknowledgments

This research was supported in part by NIH grant R01 HG006448.

References

[1] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures I: Generalized definitions". *bioRxiv* (10.1101/083915) (2019). `https://doi.org/10.1101/083915`.

[2] Alejandro Ochoa and John D. Storey. "F_{ST} and kinship for arbitrary population structures II: Method of moments estimators". *bioRxiv* (10.1101/083923) (2019). `https://doi.org/10.1101/083923`.

[3] Sewall Wright. "Systems of Mating. IV. the Effects of Selection". *Genetics* 6(2) (1921), pp. 162–166.

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

[5] S. Wright. "The genetical structure of populations". *Ann Eugen* 15(4) (1951), pp. 323–354.

[6] C. Xie, D. D. Gessler, and S. Xu. "Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method". *Genetics* 149(2) (1998), pp. 1139–1146.

[7] Jianming Yu et al. "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness". *Nat. Genet.* 38(2) (2006), pp. 203–208.

[8] Yurii S. Aulchenko, Dirk-Jan de Koning, and Chris Haley. "Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis". *Genetics* 177(1) (2007), pp. 577–585.

[9] Alkes L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". *Nat. Genet.* 38(8) (2006), pp. 904–909.

[10] William Astle and David J. Balding. “Population Structure and Cryptic Relatedness in Genetic Association Studies”. *Statist. Sci.* 24(4) (2009). Mathematical Reviews number (MathSciNet): MR2779337, pp. 451–471.

[11] Hyun Min Kang et al. “Efficient control of population structure in model organism association mapping”. *Genetics* 178(3) (2008), pp. 1709–1723.

[12] Hyun Min Kang et al. “Variance component model to account for sample structure in genome-wide association studies”. *Nat. Genet.* 42(4) (2010), pp. 348–354.

[13] Xiang Zhou and Matthew Stephens. “Genome-wide efficient mixed-model analysis for association studies”. *Nat. Genet.* 44(7) (2012), pp. 821–824.

[14] Jian Yang et al. “Common SNPs explain a large proportion of the heritability for human height”. *Nat. Genet.* 42(7) (2010), pp. 565–569.

[15] Jian Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. *Am. J. Hum. Genet.* 88(1) (2011), pp. 76–82.

[16] Cyril S. Rakovski and Daniel O. Stram. “A kinship-based modification of the armitage trend test to address hidden population structure and small differential genotyping errors”. *PLoS ONE* 4(6) (2009), e5825.

[17] Timothy Thornton and Mary Sara McPeek. “ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure”. *Am. J. Hum. Genet.* 86(2) (2010), pp. 172–184.

[18] Doug Speed and David J. Balding. “Relatedness in the post-genomic era: is it still useful?” *Nat. Rev. Genet.* 16(1) (2015), pp. 33–44.

[19] Bowen Wang, Serge Sverdlov, and Elizabeth Thompson. “Efficient Estimation of Realized Kinship from SNP Genotypes”. *Genetics* (2017), genetics.116.197004.

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

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

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

[23] Gaurav Bhatia et al. “Estimating and interpreting FST: the impact of rare variants”. *Genome Res.* 23(9) (2013), pp. 1514–1521.

[24] A. M. Bowcock et al. “Drift, admixture, and selection in human evolution: a study with DNA polymorphisms”. *PNAS* 88(3) (1991), pp. 839–843.

[25] A. M. Bowcock et al. “High resolution of human evolutionary trees with polymorphic microsatellites”. *Nature* 368(6470) (1994), pp. 455–457.

[26] Yu-Sheng Chen et al. “mtDNA Variation in the South African Kung and Khwe—and Their Genetic Relationships to Other African Populations”. *The American Journal of Human Genetics* 66(4) (2000), pp. 1362–1383.

[27] Lev A. Zhivotovsky, Noah A. Rosenberg, and Marcus W. Feldman. “Features of Evolution and Expansion of Modern Humans, Inferred from Genomewide Microsatellite Markers”. *The American Journal of Human Genetics* 72(5) (2003), pp. 1171–1186.

[28] Gabor T Marth et al. “The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations.” *Genetics* 166(1) (2004), pp. 351–372.

[29] Sohini Ramachandran et al. “Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa”. *Proc Natl Acad Sci U S A* 102(44) (2005), pp. 15942–15947.

[30] Matthieu Foll and Oscar Gaggiotti. “Identifying the Environmental Factors That Determine the Genetic Structure of Populations”. *Genetics* 174(2) (2006), pp. 875–891.

[31] Jun Z. Li et al. “Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation”. *Science* 319(5866) (2008), pp. 1100–1104.

[32] Sarah A. Tishkoff et al. “The Genetic Structure and History of Africans and African Americans”. *Science* 324(5930) (2009), pp. 1035–1044.

[33] Graham Coop et al. “The Role of Geography in Human Adaptation”. *PLoS Genet* 5(6) (2009), e1000500.

[34] G. David Poznik et al. “Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females”. *Science* 341(6145) (2013), pp. 562–565.

[35] Noah A. Rosenberg et al. “Genetic Structure of Human Populations”. *Science* 298(5602) (2002), pp. 2381–2385.

[36] Richard E. Green et al. “A draft sequence of the Neandertal genome”. *Science* 328(5979) (2010), pp. 710–722.

[37] David Reich et al. “Genetic history of an archaic hominin group from Denisova Cave in Siberia”. *Nature* 468(7327) (2010), pp. 1053–1060.

[38] Joseph K. Pickrell and Jonathan K. Pritchard. “Inference of population splits and mixtures from genome-wide allele frequency data”. *PLoS Genet.* 8(11) (2012), e1002967.

[39] Nick Patterson et al. “Ancient admixture in human history”. *Genetics* 192(3) (2012), pp. 1065–1093.

[40] Matthieu Foll and Oscar Gaggiotti. “A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective”. *Genetics* 180(2) (2008), pp. 977–993.

[41] Iosif Lazaridis et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans". *Nature* 513(7518) (2014), pp. 409–413.

[42] Iosif Lazaridis et al. "Genomic insights into the origin of farming in the ancient Near East". *Nature* 536(7617) (2016), pp. 419–424.

[43] Pontus Skoglund et al. "Genomic insights into the peopling of the Southwest Pacific". *Nature* 538(7626) (2016), pp. 510–513.

[44] Sewall Wright. "Isolation by Distance". *Genetics* 28(2) (1943), pp. 114–138.

[45] Nuno M. Silva et al. "Human Neutral Genetic Variation and Forensic STR Data". *PLOS ONE* 7(11) (2012), e49666.

[46] R. C. Lewontin. "The Apportionment of Human Diversity". *Evolutionary Biology*. Ed. by Theodosius Dobzhansky, Max K. Hecht, and William C. Steere. Springer US, 1995, pp. 381–398.

[47] Guido Barbujani et al. "An apportionment of human DNA diversity". *PNAS* 94(9) (1997), pp. 4516–4519.

[48] L. Excoffier, P. E. Smouse, and J. M. Quattro. "Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data." *Genetics* 131(2) (1992), pp. 479–491.

[49] Mari Nelis et al. "Genetic Structure of Europeans: A View from the North–East". *PLOS ONE* 4(5) (2009), e5472.

[50] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing". *Nature* 467(7319) (2010), pp. 1061–1073.

[51] The 1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491(7422) (2012), pp. 56–65.

[52] Christopher D. Steele, Denise Syndercombe Court, and David J. Balding. "Worldwide FST Estimates Relative to Five Continental-Scale Populations". *Annals of Human Genetics* 78(6) (2014), pp. 468–477.

[53] Katarzyna Bryc et al. "Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations". *Proc. Natl. Acad. Sci. U.S.A.* 107 Suppl 2 (2010), pp. 8954–8961.

[54] Timothy Thornton et al. "Estimating kinship in admixed populations". *Am. J. Hum. Genet.* 91(1) (2012), pp. 122–138.

[55] Andrés Moreno-Estrada et al. "Reconstructing the Population Genetic History of the Caribbean". *PLOS Genetics* 9(11) (2013), e1003925.

[56] Andrés Moreno-Estrada et al. "The genetics of Mexico recapitulates Native American substructure and affects biomedical traits". *Science* 344(6189) (2014), pp. 1280–1285.

[57] Julian R. Homburger et al. "Genomic Insights into the Ancestry and Demographic History of South America". *PLoS Genet.* 11(12) (2015), e1005602.

[58] Soheil Baharian et al. "The Great Migration and African-American Genomic Diversity". *PLoS Genet.* 12(5) (2016), e1006059.

[59] W. S. Robinson. "A Method for Chronologically Ordering Archaeological Deposits". *American Antiquity* 16(4) (1951), pp. 293–301.

[60] Lawrence Hubert, Phipps Arabie, and Jacqueline Meulman. *Combinatorial Data Analysis: Optimization by Dynamic Programming.* SIAM, 2001. 172 pp.

[61] Christopher C. Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". *GigaScience* 4(1) (2015), p. 7.

[62] David H. Alexander, John Novembre, and Kenneth Lange. "Fast model-based estimation of ancestry in unrelated individuals". *Genome Res.* 19(9) (2009), pp. 1655–1664.

Supplementary Information: New kinship and F_{ST} estimates reveal higher levels of differentiation in 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

∗ Corresponding authors: alejandro.ochoa@duke.edu and jstorey@princeton.edu

Contents

S1 Human Origins

S1.1 Data Processing

The Human Origins data described in the main text is merged from several sources [41–43] and processed as follows. Both Original and Pacific in Table S1 refer to the full datasets (public and non-public portions) described in [41–43] after removing non-autosomal loci and excluding ancient individuals. The Final dataset described in the main text is the union of individuals in Original and Pacific and the intersection of loci, after additional filters described below. This dataset was processed using the plink2 software [61].

Table S1: **Overview of Human Origins datasets and filters**

Dataset	Loci (m)	Individuals (n)	Sub-subpopulations	Subpopulations
Original [42]	616,938	2583	214	11
Pacific [43]	593,124	356	38	2
Final	588,091	2922	243	11

We obtained the full (public and non-public) Human Origins data presented in [41–43] by contacting the authors and agreeing to their usage restrictions. The public subset of these data is available at `https://reich.hms.harvard.edu/datasets`. The Original dataset described in [41, 42] was obtained as a single, merged dataset. The Pacific dataset described in [43] was obtained as a separate dataset. Geographical coordinates for these individuals were obtained from supplementary data [41–43].

Both Original and Pacific were genotyped on the same microarray platform, but a small subset of loci present in Original were removed from Pacific due to more stringent quality checks (David Reich, personal communication). For that reason, we merged Original and Pacific by considering the *union* of individuals but the *intersection* of loci (Table S1).

These datasets include labels that group individuals, called here *sub-subpopulations*. We removed individuals from the singleton sub-subpopulations "`Ignore_Adygei (relative_of_HGDP01382)`", `Saami_WGA`, `Wayuu`, `Ticuna`, and `Chane`, as well as `AA` (African Americans) since they were the only non-native sub-subpopulation. Then we removed non-polymorphic loci.

We then edited a few sub-subpopulation labels, as follows. We merged all individuals from the four subgroups `GujaratiA-D` into `Gujarati`. Lastly, the label `Southwest_Bougainville` was shortened to `SW_Bougainville` in figures. The resulting 243 native sub-subpopulations were manually grouped for visual aid into 11 subpopulations by taking into account geography, population history, and our kinship estimates (Fig. S1).

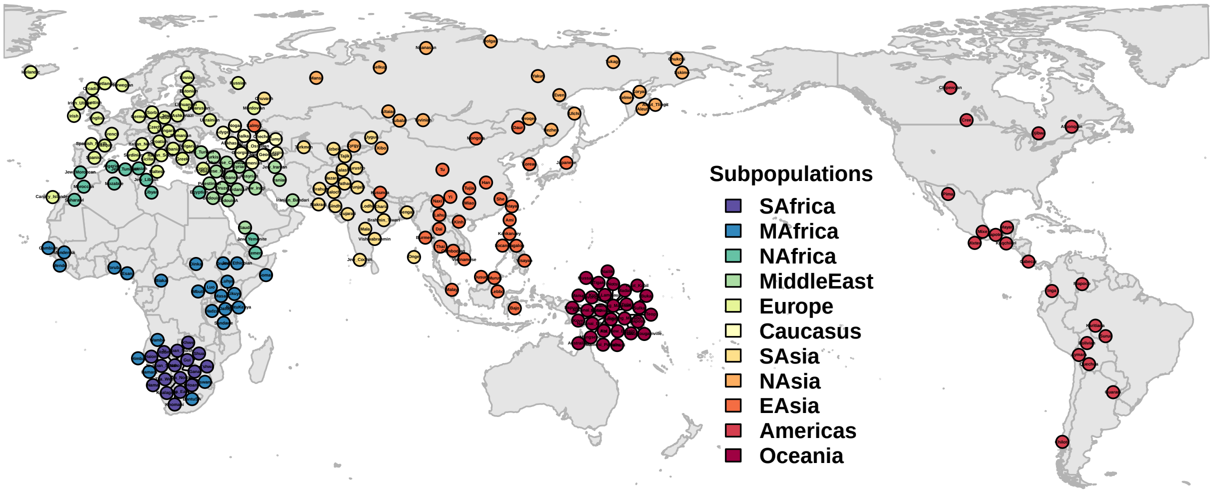


Figure S1: **Grouping of Human Origins sub-subpopulations into $K = 11$ subpopulations.** Each circle represents a sub-subpopulation, which was placed near its sampling location but nudged if necessary so circles do not overlap. The color of the circles corresponds to the subpopulation cluster we assigned. This partition into subpopulation is based on geography, history, language families, and our kinship estimates.

S1.2 Weights for individuals

The Final dataset has a wide distribution of sub-subpopulation sample sizes, with a median sub-subpopulation size of 10 individuals, a minimum of 2 (`Canary_Islander`), and a maximum of 70 (`Yoruba`).

To calculate our generalized F_{ST} estimate, weights were constructed so that every subpopulation is weighed equally, and every sub-subpopulation weighs equally within each subpopulation, as follows. For every individual j, let n_j be the number of individuals in the sub-subpopulation of j, and m_j be the number of sub-subpopulations in the subpopulation of j. The weights used are then $w_j = \frac{1}{K n_j m_j}$, where K is the number of subpopulations.

S1.3 Comparison of new and existing kinship estimates

A direct comparison of each our new kinship and inbreeding estimates on the real data to those from the standard kinship estimator are presented in Fig. S2. We found previously that bias in the standard kinship estimator varies for every pair of individuals depending on their mean kinship to everybody else in the dataset [2]. This effect is evident in our comparisons, resulting in complex patterns for the standard kinship estimates that are not linear and are not functions of the true kinship values (estimated without bias by our new estimator).

Biases in standard inbreeding coefficients (estimated as $\hat{f}_j^{\text{std}} = 2\hat{\varphi}_{jk}^{\text{std}} - 1$; second row of Fig. S2) are considerably more extreme compared to the biases of the kinship values between different indivdiuals (first row of Fig. S2). In particular, standard inbreeding estimates often exceed 1, and in the case of Hispanics from 1000 Genomes they are negatively correlated with their true values.

S1.4 Admixture analysis

We performed an admixture analysis to complement out analysis of kinship and F_{ST}. We used the Admixture software [62] to infer $K = 7$ ancestry clusters from the Final dataset (see Table S1). The admixture proportions can be visualized as stacked barplots on Fig. S3, where individuals are ordered just as in the kinship matrix of the main text. These seven infered ancestry clusters correspond approximately with the 11 subpopulations we constructed based on geography and other criteria (Section S1.1) as follows (Fig. S3):

- Cluster 1 corresponds to `SAfrica`.
- Cluster 2 corresponds to `MAfrica`.
- Cluster 3 corresponds to `Europe`, `NAfrica`, `MiddleEast` and `Caucasus`.
- Cluster 4 corresponds to `SAsia`.
- Cluster 5 corresponds to `EAsia` and `NAsia`.
- Cluster 6 corresponds to `Americas`.
- Cluster 7 corresponds to `Oceania`.

We typically see that each ancestry cluster is concentrated in a certain geographical region, and this ancestry is also present to a lesser extent in neighboring regions and diminishes with geographical distance from its point of greatest concentration. This again argues for a complex population structure where relatedness at the population level falls on a continuum rather than taking on discrete values.

The most notable geographic discontinuities in ancestry were observed for cluster 3, which is roughly West Eurasian ancestry. Unusually high proportions of this ancestry are observed in most individuals of two sub-subpopulations of the Kamchatka Peninsula in Russia (`Aleut` and `Aleut_Tlingit` in subpopulation `NAsia`), as well as the four `Americas` sub-subpopulations from Canada (`Chipewyan`, `Cree`, `Algonquin`, and `Ojibwa`) and `Chilote` from Chile (Fig. S3). We also observed an unusual reduction of cluster 5 ancestry in `Aleut` and `Aleut_Tlingit` relative to its closest `NAsia` sub-subpopulations, and unusually low levels of cluster 6 ancestry in `Chipewyan`, `Cree`, `Algonquin`, `Ojibwa`, and `Chilote` relative to the rest of `Americas` sub-subpopulations. These data point to likely recent European admixture in those individuals.

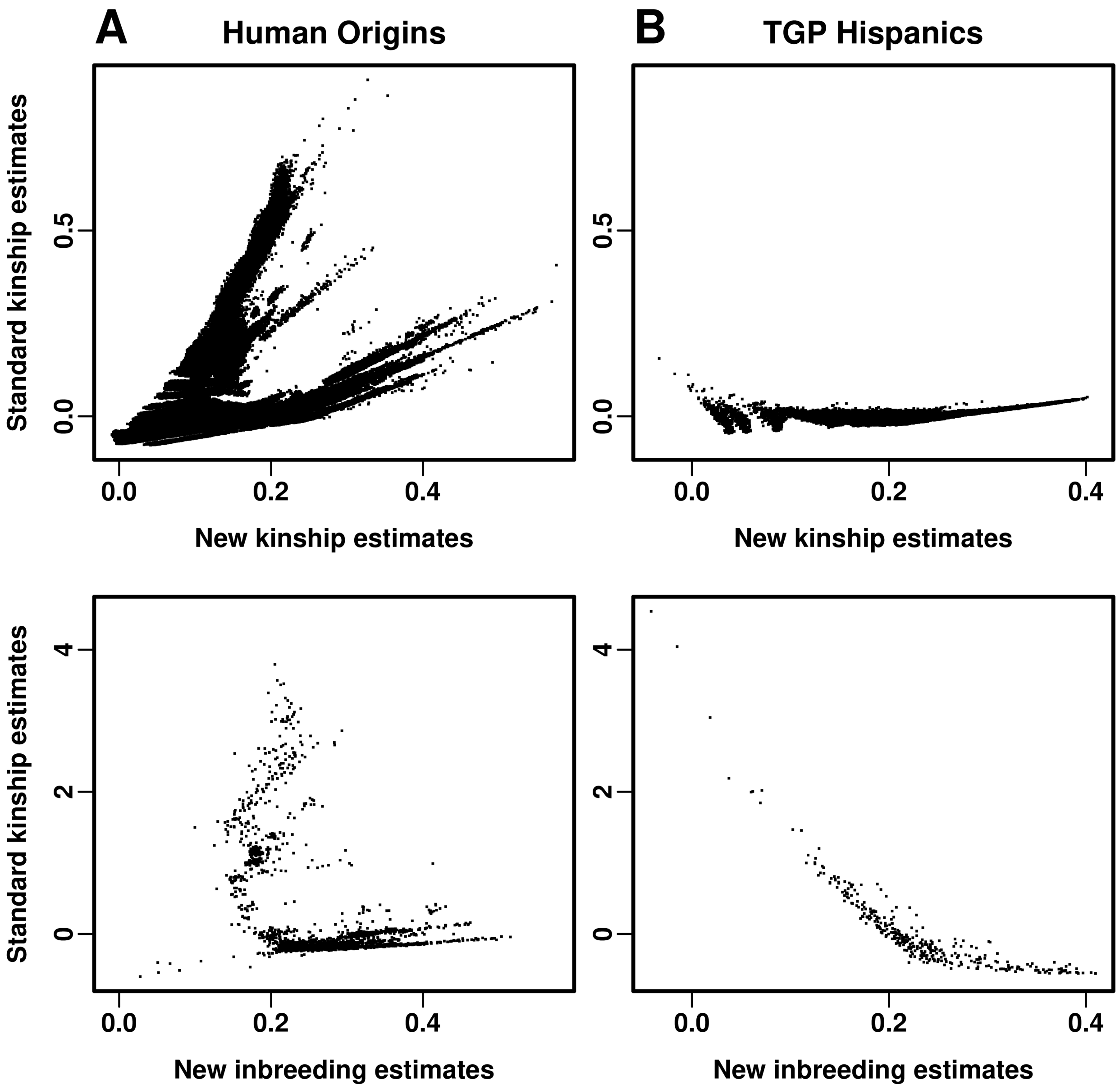


Figure S2: **Comparison of new and existing kinship estimates.** The x-axes are estimates from the new kinship estimator, while the y-axes are estimates from the standard kinship estimator. Inbreeding coefficients (second row) are compared separately from kinship coefficients (between different individuals; first row) since the scales are very different for the standard kinship estimator (but not for the new kinship estimator). Columns: **A.** Comparison of estimates in the Human Origins dataset. **B.** Comparison of estimates in the 1000 Genomes Hispanics dataset.

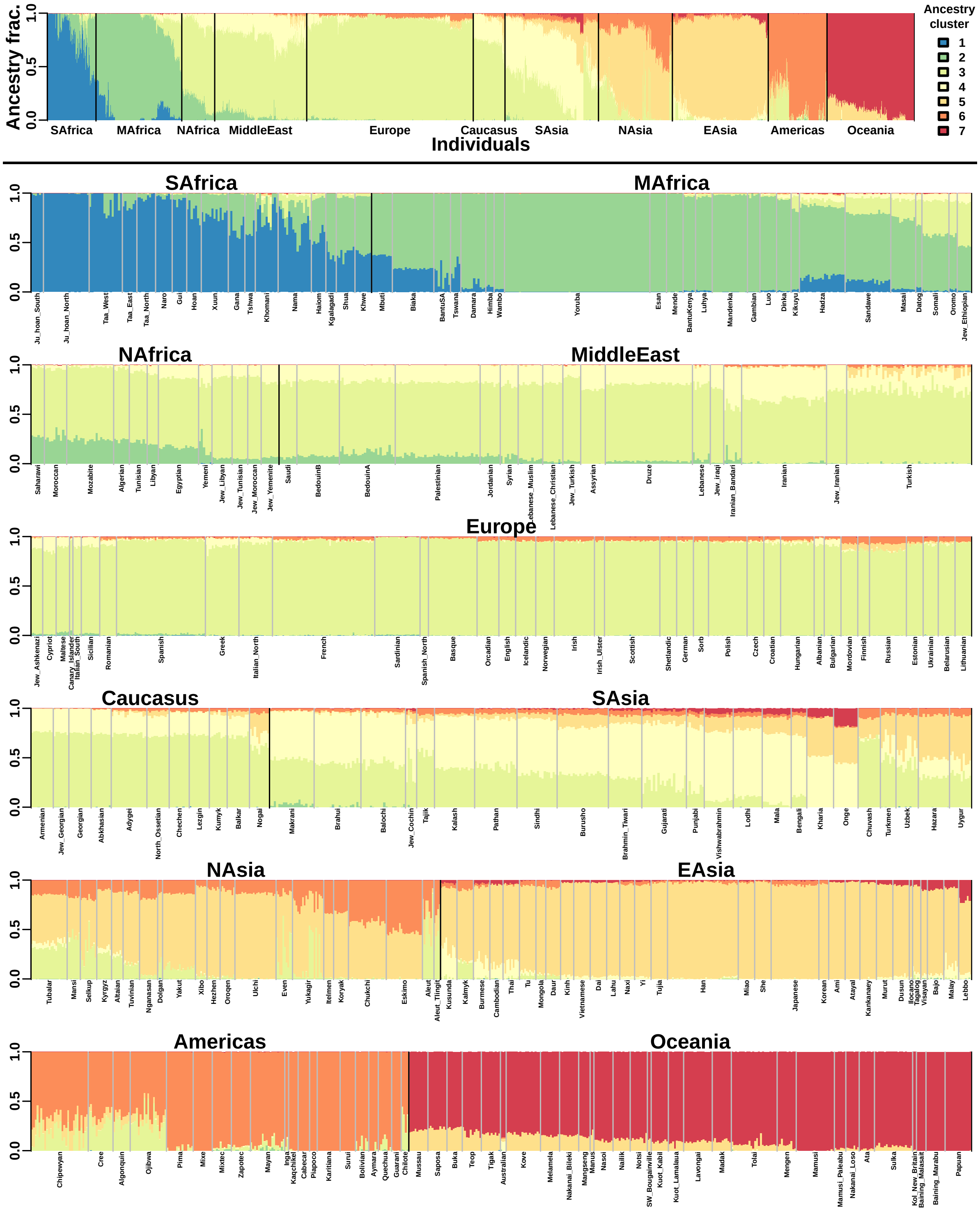


Figure S3: **Admixture analysis of Human Origins with $K = 7$.** The top row shows the full set of admixture proportions for the $K = 7$ infered ancestry clusters and all 11 subpopulations in Human Origins. All other rows show the same data with greater detail, including labels for every sub-subpopulation. The seven ancestry clusters were ordered manually to correspond roughly with distance from southern Africa.

S2 1000 Genomes Project

S2.1 Data processing

The 1000 Genomes Project (TGP) "Phase 3" integrated call data [50, 51] is available at `http://www.internationalgenome.org/data` (dated 2013-05-02). We started from the plink2-formatted version available at `https://www.cog-genomics.org/plink/2.0/resources#1kg_phase3`. This dataset was processed using the plink2 software [61]. Our analysis was restricted to autosomal biallelic SNP loci ascertained in `YRI`, after removing loci with repeated identifiers (20,417,484 loci). Of these, 14,145,583 loci are polymorphic in Hispanics (`PUR`, `CLM`, `PEL`, `MXL`; Table S2).

Table S2: **Overview of 1000 Genomes (TGP) Hispanics dataset**

Dataset	Loci (m)	Individuals (n)	Subpopulations
Full TGP (ascertained in `YRI`, other locus filters)	20,417,484	2504	26
Hispanics	14,145,583	347	4
Hispanics + Admixture Panels	6,216,713	665	7

S2.2 Admixture analysis

The Admixture analysis of the Hispanic individuals was performed with the addition of individuals from the `YRI`, `IBS`, and `CHB` subpopulations to help anchor the $K = 3$ admixture clusters. Only loci with minor allele frequency $\geq 10\%$ across the 7 subpopulations (6,216,713 loci, see Table S2) were input to the Admixture software [62]. The cluster associated with `YRI` was assigned to Sub-Saharan African (AFR) ancestry, `IBS` to European (EUR) ancestry, and `CHB` to Native American (AMR) ancestry by proxy (Fig. S4). There are no Native American subpopulations in 1000 Genomes, but the high AMR ancestry predicted for many `PEL` and `MXL` individuals suggests that AMR ancestry is not being underestimated by this procedure.

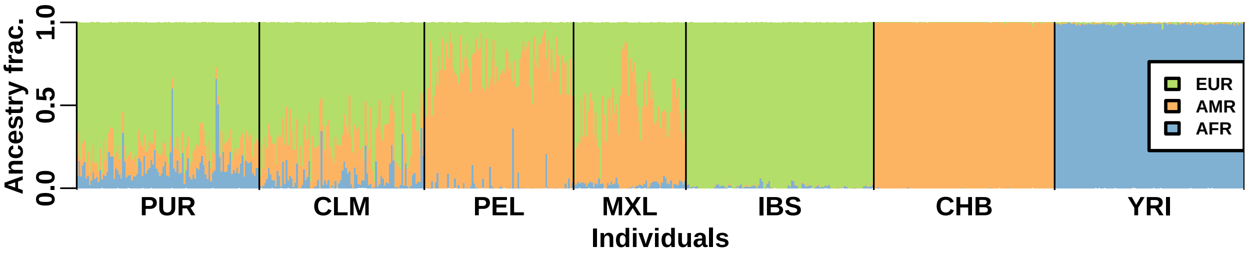


Figure S4: **Ancestry inference in Hispanic individuals.** Admixture proportions of every individual using `YRI`, `IBS`, and `CHB` as reference panels for Sub-Saharan African (AFR), European (EUR), and Native American (AMR) ancestry, respectively.

S2.3 Estimation of minimum kinship

The kinship matrix of the Hispanic individuals was estimated as follows. First, the A_{jk} values were estimated, and the function `seriate` from the R package `seriation` was used with default values to reorder the columns and rows of this matrix so that the lowest kinship values are pushed away from the diagonal [59, 60]. We inspected the individuals at the extremes of the resulting ordering, and found that four individuals with among the highest African admixture proportions also shared the smallest kinship estimates in the data. The two most extreme clusters, (`HG01108`, `HG01242`) and (`HG01551`, `HG01241`), were used to estimate A_{min}, which yields the final kinship estimates $\hat{\varphi}_{jk}^{\text{new}}$.