4-4-2017

Beyond Serial Founder Effects: The Impact of Admixture and Localized Gene Flow on Patterns of Regional Genetic Diversity

Keith Hunley
University of New Mexico, khunley@unm.edu

Graciela S. Cabana
University of Tennessee, gcabana@utk.edu

BEYOND SERIAL FOUNDER EFFECTS: THE IMPACT OF ADMIXTURE AND LOCALIZED GENE FLOW ON PATTERNS OF REGIONAL GENETIC DIVERSITY

Keith L. Hunley[1]*, Graciela S. Cabana[2]

[1]Department of Anthropology, University of New Mexico, Albuquerque, NM 87131

[2]Department of Anthropology, University of Tennessee, Knoxville, TN 37996

Short title: Founder effects, admixture and gene flow in human evolution

*Corresponding author

E-mail: khunley@unm.edu

Header: Founder effects, admixture and gene flow in human evolution

Keywords: serial founder effect, admixture, gene flow

ABSTRACT

Objectives. Geneticists have argued that the linear decay in within-population genetic diversity with increasing geographic distance from East Africa is best explained by a phylogenetic process of founder effects, growth, and isolation termed serial founder effects (SFE). However, the SFE process has not yet been adequately vetted against other evolutionary processes that may also affect geospatial patterns of diversity. Additionally, studies of SFE have been largely based on a limited 52 population sample from the HGDP-CEPH. Here, we assess the effects of SFE, admixture, and localized gene flow processes on patterns of global and regional diversity using

a published dataset consisting of 645 autosomal microsatellite genotypes from 5,415 individuals in 248 widespread populations.

Materials and Methods. Because SFE is a phylogenetic process, we used a formal tree-fitting approach to explore the role of the process in shaping patterns of global and regional diversity. The approach involved fitting global and regional population trees to extant patterns of gene diversity and then systematically examining the deviations in fit. We also informally tested the SFE process using linear models of gene diversity vs. waypoint geographic distances from Africa. Because gene flow and phylogenetic processes can both shape geospatial patterns of diversity, we tested the role of localized gene flow using partial Mantel correlograms of gene diversity vs. geographic distance controlling for the confounding effects of tree-like genetic structure.

Results. We corroborate previous findings that global patterns of diversity, both within and between populations, are the product of an out-of-Africa SFE process. Within regions, however, diversity within populations is uncorrelated with geographic distance from Africa. Instead, deviations in the fit of regional population trees are largely the product of recent inter-regional admixture. Additionally, in several regions, we found that positive correlations between pairwise gene diversity and geographic distance, frequently attributed to localized gene flow, were instead the product of phylogenetic processes associated with initial peopling or subsequent range expansions.

Conclusions. Detailed analyses of the pattern of diversity within and between populations reveal that the signatures of different evolutionary processes dominate at different geographic

scales. These findings have important implications for recent publications on the biology of race.

INTRODUCTION

In recent years, a consensus has begun to emerge that an out-of-Africa serial founder effect process (SFE) played a central role in shaping neutral genetic diversity in humans. SFE is a special case of the phylogenetic model, which assumes that extant populations formed through a tree-like process of splitting and isolation (Cavalli-Sforza and Edwards, 1967). The SFE process incorporates splitting and isolation, but it also includes founder effects, growth of descendant populations, and a geographic pattern to the populations splits formed by steady movement away from an African homeland (Harpending and Rogers, 2000; Prugnolle et al., 2005; Ramachandran et al., 2005). This continual movement away from a single source location is hypothesized to account for the linear decay in genetic diversity within the 52 populations of the HGDP-CEPH with increasing geographic distance from East Africa (Harpending and Rogers, 2000; Tishkoff and Kidd, 2004; Prugnolle et al., 2005; Ramachandran et al., 2005; DeGiorgio et al., 2009, 2011; Hunley et al., 2009).

Pickrell and Reich (2014) recently challenged this interpretation of the pattern of diversity, arguing that long-range population movements and admixture have been so pervasive in recent human evolution that any genetic signature of founder effects associated with the initial radiation from Africa would have been erased. Further, they demonstrated using simulations that inter-regional admixture alone could have produced the observed decay in diversity from East Africa. In support of this view, there is substantial empirical evidence that long-range movement and admixture has been common in the past several thousand years, and that this admixture has affected patterns of regional genetic diversity, particularly in the Americas

(Rosenberg et al., 2002; Wang et al., 2007; Hunley and Healy, 2011; Pickrell and Pritchard, 2012). Additionally, though we focus here on recent admixture, long distance dispersals into new or previously colonized areas, contractions and expansions associated with the last glacial maximum, and range expansions associated with the rise of agriculture may also have obscured or erased evidence of SFE (Alves et al., 2016).

Localized gene flow between neighboring populations may also have shaped patterns of diversity at various geographic scales (Relethford, 2004, 2009; Handley et al., 2007). Given sufficient time, localized gene flow will erase genetic signatures of earlier evolutionary processes, an outcome often referred to as isolation by distance. Importantly, localized gene flow is easy to confound with geographically-patterned, tree-like range expansion processes like SFE (Meirmans, 2012), because both processes affect the relationship between geographic distance and the level of divergence between populations (Ramachandran et al., 2005; Hunley et al., 2009). For this reason, tests of localized gene flow must control for the effects of genetic structure caused by geographically-patterned range expansions.

The goal of this study is to reevaluate the SFE model at global and regional levels through joint analyses of patterns of within- and between-population variation in 248 globally-distributed populations. This joint analysis permitted us to disentangle the effects on patterns of global and regional diversity of evolutionary process such as geographically-patterned range expansions, admixture, and localized gene flow. Additionally, the large sample of populations overcomes several potential limitations of the HGDP-CEPH, including small numbers of populations per

region, and the potentially disproportionate representation of relatively isolated populations (Cavalli-Sforza, 2005).

METHODS

Data. Autosomal microsatellite loci are ideal for this study because they lack the ascertainment bias of single nucleotide polymorphisms, which may lead to inaccurate estimates of genetic diversity within and between populations (Rogers and Jorde, 1996), and because they have been typed in large numbers of populations in each region. The 645 loci used in this study were collated by Pemberton et al. (2013) from Cann et al. (2002), Friedlaender et al. (2008), Kopelman et al. (2009), Pemberton et al. (2012), Rosenberg et al. (2002, 2005, 2006), Tishkoff et al. (2009), and Wang et al. (2007, 2008). In compiling the data, Pemberton et al. excluded one individual from each pair of monozygotic twins and first degree relative pairs, both within and between all populations except the Karitiana and Surui. We additionally excluded admixed Hispanic and African American populations as well as the African Dogon (n=3), Eton (n=4), and Ewondo (n=3) populations due to small sample sizes. Our final sample consisted of 5,415 individuals from 248 populations. We divided the African sample into East and Central West subgroupings. The East African sample consists of a cluster of 41 populations spread roughly along the Rift Valley across Kenya, Ethiopia and Tanzania. The Central West African sample consists of a cluster of 49 populations located in Cameroon, Chad, and Nigeria. “Other” African includes populations not located in the East and Central West African clusters, and four pygmy populations in Central West Africa.

Statistical Methods. Following other studies that have used trees to examine SFE and recent admixture (Jakobsson et al., 2008; Li et al., 2008; Pickrell and Pritchard, 2012; Hunley et al., 2015), our broad approach was to construct population trees from the data and then to test the fit of the trees both formally and informally. We chose a tree-fitting approach for several reasons. First, as noted, SFE is a phylogenetic process. If SFE were the only process that shaped the structure of neutral genetic diversity in our species, a population tree constructed from genetic data would perfectly capture the history of population splits and the level of diversity within and between extant human populations. Previous studies have in fact shown that population trees capture the global structure of human diversity well (Hunley et al., 2015). Second, by systematically examining specific causes of lack of fit of trees to the data, we can gain direct insights into the effects of recent admixture and localized gene flow on patterns of diversity. Third, even if other evolutionary processes like admixture and localized gene flow have dominated recent evolution, residual effects of an SFE process will persist. The tree-fitting approach we employ here will both capture these residual effects and provide a mechanism for controlling for those effects on tests of admixture and gene flow. Miermans (2012) recently demonstrated the need for such control when he showed than both isolation by distance and geographically-patterned hierarchical processes (like SFE) can produce spatial autocorrelation in allele frequencies. For these reasons, the tree-fitting approach can complement other approaches that have been employed to study the structure of human diversity and the evolutionary and demographic processes that have shaped that diversity (Rosenberg et al., 2002; Prugnolle et al., 2005; Ramachandran et al., 2005; Novembre et al., 2008; Creanza et al., 2015; Alves et al., 2016).

To implement the tree-fitting approach, we first estimated gene diversity (Nei, 1987) within and between populations, visualized the geographic pattern using scatter plots, and fit linear models to the data at global and regional levels. Great circle geographic distances were estimated through waypoints on land as described in Ramachandran et al. (2005).

We computed five measures of genetic distance or difference (Reynolds et al., 1983; Weir and Cockerham, 1984; Nei, 1987; Goldstein et al., 1995; Slatkin, 1995) under the infinite alleles and stepwise mutation models (Kimura and Crow, 1964; Ohta and Kimura, 1973). The five measures were strongly correlated with one another. We next constructed a global population tree (GPT) using the Fitch-Margoliash method (Fitch and Margoliash, 1967). We rooted the GPT using chimpanzee data from Pemberton et al. (2013).

We also constructed four types of regional population trees using the Neighbor Joining (NJ) method (Saitou and Nei, 1987) and formally tested their fit to regional patterns of gene diversity. Variation in the topology and fit of these trees is informative about the combined effects of SFE and admixture on patterns of regional diversity. The first type, termed the "island" model, assumes independent evolution among all populations within a region after divergence from a single ancestral population. The island models are intended to represent the simplest possible model of tree-like evolution in each region. The second type consisted of regional "sub-trees" constructed by simply pruning out regional portions of the GPT. Because the topologies and branch lengths were extracted intact from the GPT, the relationships among the populations in a given region were directly affected by their relationships to populations in other regions. The third type consisted of new "region-only" population trees constructed using

only the genetic distances between populations in each region. Since these region-only trees were made solely from the distances in each region, their topologies and branch lengths were not directly affected by their genetic distance to populations outside of the region. The fourth type of tree was constructed independently from genetic distances in each region after removing populations with greater than 10% ancestry from another region, a necessarily arbitrary value chosen to simply assess the general impact of inter-regional admixture on the relative fit of regional population trees. We refer to these as “un-admixed” trees.

We formally tested the fit of each type of tree to the actual pattern of gene diversity using a maximum likelihood-based, generalized hierarchical modeling method (GHM).The GHM method estimates a set of ‘‘expected’’ gene diversities contingent on a given tree’s topology and the actual gene diversities (Cavalli-Sforza and Piazza, 1975; Long and Kittles, 2003). The fit of each tree was assessed using a likelihood ratio statistic, Λ, which, under the limit of large sample size, is distributed as a $\chi 2$ random variable with degrees of freedom equal to the number of populations plus the number of pairs of populations minus the number of nodes in the tree (Urbanek et al., 1996; Long and Kittles, 2003). The GHM method was also used to identify the root of the region-only trees under the assumption that the best fitting tree contained the correct root (Hunley et al., 2015). In these analyses, we first constructed an unrooted region-only NJ tree for each region, placed the root at every branch in the tree, and tested the fit of each tree using the GHM method. For these analyses, we combined the African sub-regions into a single African region, and we combined the Middle Eastern and European samples into a single region.

We examined population-level deviations in the fit of the four types of trees in the Americas as follows: First, after fitting the trees using the GHM method, we calculated Nei's minimum genetic distances from the actual gene diversities and from the GHM-estimated gene diversities. We then fit a linear model to these actual vs. expected genetic distances, and visualized the residuals using violin plots (Hintze and Nelson, 1998).

We constructed Mantel correlograms and partial Mantel correlograms (Oden and Sokal, 1986; Legendre and Legendre, 2012), using the mpmcorrelogram package in R (Matesanz et al., 2011). The correlograms measured the correlation between gene diversity matrices and geographic distance matrices after converting the latter to *n* new binary matrices, where *n* is the number of distance classes. The binary matrices for a given distance class contained 1s if the geographic distances were in the class and 0s if they were not. For the partial correlations, we controlled for genetic structure using GHM-estimated gene diversities for the region-only trees. Distance classes of approximately equal size were created from quantiles of geographic distances. The results reported below were insensitive to the number of distance classes. *P*-values for the statistical tests were set at 0.05 divided by the number of distance classes. To minimize the impact of inter-regional admixture on the analyses, as above, we removed populations with greater than 10% ancestry from a different region, though this removal did not impact our results or their interpretation.

We estimated admixture proportions in individuals and populations using the model-based clustering algorithm implemented in *Structure* (Pritchard et al., 2000). We conducted multiple runs of the program for values of *K* from 3 to 35. The likelihoods were relatively uniform above

values of K = 6. Above this value, additional substructure formed within regions, but estimates of inter-regional admixture proportions remained stable. We chose to display the results for K = 7, which was the lowest value of K that separated the geographic regions into separate clusters.

RESULTS

Gene diversity within populations. Beginning with the within-population pattern of diversity, Figure 1 shows gene diversity vs. geographic distance through waypoints on land from Addis Ababa, Ethiopia. The high R^2 of 0.778 for the 248-population sample is similar to that reported by Ramachandran et al. (2005) for 51 populations in the HGDP-CEPH sample (R^2=0.763) and by Pemberton et al (2013) for 239 populations (R^2 = 0.841). The large circles in the figure mark the mean gene diversity vs. mean geographic distance for each region.

Within regions, however, gene diversity is uncorrelated with waypoint distances from East Africa except in the Americas (R^2=0.257, p=0.005). This correlation in the Americas, however, may be an artifact of a north-south gradient of European admixture. In partial correlation analyses controlling for European ancestry, the squared correlation drops to 0.052 (p = 0.23). In contrast, the correlation between European ancestry and geographic distance from Beringia retains statistical significance when gene diversity is controlled (R^2=0.29, p=0.009).

Gene diversity between populations. Turning to the between-population pattern, we used various methods to construct a global population tree (GPT). All trees shared the following features, evident in Figure 2. First, the trees were rooted at the node connecting the Sub-Saharan African San and !Xun/Khoe to the remaining 246 populations. Second, moving away from the root, multiple nodes along the base of the tree separated Sub-Saharan African

populations on one side of the node from both Sub-Saharan African and non-Sub-Saharan African populations on the other side. Third, outside of Africa, populations clustered by region, and the regions were nested inside one another. Fourth, within regions, the population order in the GPT often paralleled the geographic location of the populations, though, as we show below, there are important exceptions to this pattern. Overall, these features of the GPT are consistent with an African origin, migration out of Africa, and successive founder events associated with the initial entry into each region.

Again, if SFE was the only process that affected diversity, the GPT would capture the history of population splits as well as the level of gene diversity within and between populations. Figure 3 permits us to assess how well the GPT actually fits the gene diversity pattern. It shows gene diversity vs. waypoint geographic distance between all pairs of populations. The circles show the gene diversities between populations within each region, and the dashes show the gene diversities between populations in different regions. The most salient feature of the plot is the layered pattern of variation. In general, populations connected by a node in the GPT have uniform gene diversity to one another, whether they are located nearby or thousands of kilometers apart. The San and !Xun/Khoe for example, have high and approximately uniform gene diversity to the other 246 populations. This aspect of the plot indicates that the GPT captures the gene diversity pattern well.

Another noteworthy feature of the between-population pattern is that the slope of the plot is positive ($r = 0.200$, $p < 0.001$). Such positive correlations are typically assumed to be the outcome of long periods of continuous, localized gene flow, often referred to as isolation by

distance, not a geographically-patterned tree-like process like SFE. However, the layered pattern in the figure is clearly inconsistent with isolation by distance.

We used a model-based clustering method (Pritchard et al., 2000) to examine genetic structure globally, and within each region. In Figure 4, we show the results of an unsupervised cluster analysis at $K = 7$. As expected from previous analyses of the HGDP-CEPH sample (Rosenberg et al., 2002), as well as the structure of the GPT in Figure 2, populations tend to cluster by region. Populations that lie near the borders of regions have ancestry from regions on either side of the border, and the level of this inter-regional ancestry tends to dissipate with increasing geographic distance from the border. In several cases, admixture between proximate regions occurs far away from the regional borders, e.g., the three populations in Remote Oceania have high levels of East Asian ancestry, possibly related to the migration of Austronesian speaking peoples from Southeast Asia beginning about 3,000 years ago (Friedlaender et al., 2008).

The plot shows that populations in most regions also have ancestry from non-proximate regions. The level of this form of inter-regional ancestry varies widely across populations. The most notable example is in the Americas, where the average European ancestry in the 29 Native American populations is 7.8%. The level of European ancestry is highest in the three northernmost populations located in Canada, and lowest in isolated populations in Brazil and Paraguay. European ancestry also varies substantially across individuals within admixed populations; in the three Canadian populations, for example, it ranges from 0.8%-62.9%.

To more closely examine the effects of admixture on patterns of regional diversity, we constructed and tested the fit of four types of region-specific population trees. As an example,

the four tree types for the Americas are shown in Figure 5. The population names are colored by geographic sub-region. The pie charts at the tips of the trees show the proportion of Native American ancestry (white) and non-Native American ancestry (colors from Fig. 4) for each population. The sub-tree in Figure 5B was pruned from the GPT with topology and branch lengths intact. These features of the tree were therefore affected by distances between Native American and non-Native American populations. In the tree, populations with high Middle Eastern and European genetic ancestry tend to cluster near the base of the tree. The region-only tree in Fig. 5C was constructed just from genetic distances in the Native American populations, and was therefore unconstrained by distances to populations outside of the Americas. The un-admixed tree in Fig. 5D was constructed just from genetic distances in the Native American populations after removing nine heavily admixed populations (see Methods). In contrast to the sub-tree, the order of populations in the region-only and un-admixed trees aligns more closely with geographic location, starting in North America and ending in eastern South America. Note, for example, the position of the heavily admixed South American Wayuu, Arhuaco, and Kaingang populations in the sub-tree vs. the region-only tree.

We used a generalized hierarchical modeling (GHM) method to formally test the fit of each type of tree to the actual gene diversity pattern. The comparative fit of the trees is shown in Figure 6. In all but one case, 1) sub-trees, region-only trees, and un-admixed trees fit better than the island trees, 2) region-only trees fit better than the sub-trees, and 3) un-admixed trees fit better than the region-only trees. The lone exception to this pattern was in Africa for the region-only vs. un-admixed tree. In the full African sample, only 12 of the 119 populations had greater than 10% ancestry from another region, and nine of these populations were concentrated in the

distal portion of the tree. Residents of the nine populations speak Afro-asiatic languages, and they are located near northern and eastern borders of Africa. This result suggests that admixture that occurs near the borders of regions has minimal effect on the structure of regional trees and their fit to the gene diversity pattern.

To explore the contribution of individual populations to the lack of fit of the Americas trees, we used violin plots of residuals of linear models of actual vs. GHM-estimated genetic distances. The top four violin plots in Figure 7 show the range of residuals for the four trees; the range decreases steadily from one tree to the next, as expected based on the GHM results. The remaining violin plots in the figure compare the residuals for the sub-tree vs. region-only tree for the nine most heavily admixed populations. The plots confirm the large contribution of the admixed Arhuaco, Wayuu, and Kaingang populations to the lack of fit of the sub-tree.

It is noteworthy that the three northernmost populations in the Americas (Cree, Ojibwa, and Chipewyan) contribute relatively little to the lack of fit despite the fact that they have among the highest levels of ancestry from other regions, especially from Europe. This result extends that from the model fitting results for Africa above in showing that, to the extent that 1) inter-regional admixture occurs from populations with relatively high gene diversity (e.g., Europe) into populations with relatively low gene diversity (e.g., the Americas), and 2) the low-diversity populations were already more diverse that other populations in their region, then the admixture has minimal impact on tree structure and the fit of the tree.

Finally, to examine the impact of localized gene flow on patterns of regional diversity, we constructed scatter plots of population-pairwise gene diversity vs. geographic distance, shown

in the top row of Figure 8. Persistent localized gene flow will produce a positive correlation between these variables. If the populations span a large geographic range, the correlation will tend to level out at high geographic distances. As the plots show, in all regions, the correlation is positive. Each correlation is statistically significant at the 0.05 level.

The Mantel correlograms on the second row permit us to explore the correlations more closely. They show evidence of an asymptote at higher geographic distances in Central West Africa, Oceania, and the Americas. In other regions, the steady rise in gene diversity continues across the full geographic range. While it is tempting to interpret the plots as evidence of localized gene flow, as we showed in Figure 3, geographically-patterned, tree-like range expansions can also produce positive correlations. To distinguish between these possibilities, it is essential to control for genetic structure that may have arisen through range expansion.

The partial Mantel correlograms on the bottom row of Figure 8 provide this control by holding constant the gene diversities estimated using the GHM method for the region-only trees. If the Mantel correlograms patterns persist after controlling for genetic structure, the correlations may be attributed to localized gene flow; if they dissipate, they are instead the product of tree-like structure in each region. Compared to the correlogram patterns, the partial correlation patterns flatten out in Central West Africa, Central South Asia, and the Americas, indicating that tree-like structure, not localized gene flow is responsible for the correlation between gene diversity and geographic distance. In contrast, the Mantel correlogram patterns persist in the partial correlograms in East Africa and Oceania, and in first distance class in the Middle-East Europe and East Asian regions. This persistence indicates that localized gene flow has shaped

variation in these regions. The persistence across the full geographic range in East Africa and Oceania may reflect the low average geographic distance between populations in these two regions, at only 231 km in East Africa, and 280 km in Oceania.

DISCUSSION

Our analyses of 248 widespread populations confirm findings from previous analyses of the HGDP-CEPH sample of a strong negative correlation between gene diversity within populations and geographic distance from East Africa. However, our analyses also showed that gene diversity is uncorrelated with geographic distance from East Africa within each region. On its own, this finding potentially calls into question the hypothesis that the initial peopling of the globe occurred though an out-of-Africa SFE process. The finding also demonstrates the need for caution in inferring complex evolutionary processes simply from the magnitude of the correlation of a linear model.

However, we also documented a layered pattern of *between-population* variation (Fig. 3) that also confirms findings from previous analyses of the HGDP-CEPH sample (Hunley et al., 2009; DeGiorgio et al., 2011). The layers of variation correspond closely to the structure of the GPT (Fig. 2). This correspondence is clearly consistent with a tree-like out-of-Africa SFE process *at the global level*, and it is not consistent with the structured admixture process envisioned by Pickrell and Reich (2014).

Our analyses also demonstrate that recent inter-regional admixture has affected the structure of regional population trees (see also Pickrell and Pritchard, 2012). This admixture contributes to the lack of correlation between gene diversity and distance from East Africa within regions.

Though we concentrate our analyses on the Americas, we find that many populations in all regions are admixed; the average level of inter-regional ancestry across all 248 populations was greater than 10% for all values of *K* above 5. In fact, our analyses may understate the importance of admixture in recent human evolution because many of the populations in the original HGDP-CEPH sample (which comprise 21% of the 248-population sample) were intentionally chosen to represent un-admixed descendants of the initial inhabitants of a given location (Cavalli-Sforza, 2005). Other evolutionary processes that may contribute to the absence of a correlation between gene diversity and geographic distance from East Africa within regions include non-uniform movement across a landscape, e.g., preferential dispersal along coasts (Wang et al., 2007), and secondary population expansions that occurred after initial peopling.

Our Mantel correlogram analyses showed that spatially-patterned variation that is frequently attributed to localized gene flow (Relethford, 2004, 2009; Handley et al., 2007), is in fact the product of tree-like range expansions in Central West Africa, Central South Asia, and the Americas. The range expansions may have occurred during initial peopling (though probably not from a location that is compatible with a simple out-of-East-Africa SFE process), or afterwards, for example, during the spread of agriculture. Our findings are consistent with those from previous studies that demonstrated that both localized gene flow and SFE processes can produce a positive correlation between geographic and genetic distances, a positive correlation between gene diversity (between populations) vs. geographic distances, and geographic gradients to PCA factors (Ramachandran et al., 2005; Hunley et al., 2009; Hunley and Healy, 2011; Creanza et al., 2015). These findings emphasize the importance of taking into account the

potentially confounding effects of geographically-patterned genetic structure in tests of localized gene flow (Meirmans, 2012). We also note that it would take a great deal of time for localized gene flow to affect patterns of diversity at large geographic scales (Wilkins and Marlowe, 2006). In fact, it would take tens of thousands of generations to reach a steady state isolation-by-distance pattern that would completely erase genetic signatures of earlier processes. This fact may explain why region-only trees still fit the data well even in regions where we detected signals of localized gene flow; in these regions, both tree-like range expansions and localized gene flow have shaped patterns of variation. Even in the regions where partial Mantel correlograms revealed an absence of evidence for localized gene flow, we are not claiming that it has been unimportant, but only that its effects are not always manifest at the large geographic scales analyzed here.

We wish to address the fact that the clustering of people by region in model-based cluster analysis (e.g., Figure 4) has been attributed to the existence of biological races, both in academic (Sesardic, 2010) and popular literature (Wade, 2014). These racial interpretations fail to take into account the following facts. First, unsupervised cluster analyses provide no direct information about evolutionary process because they fail to account for the pattern of diversity *between* populations. So, for example, these analyses fail to capture the important fact that African populations are not monophyletic, as is clearly demonstrated in the GPT in Figure 2. We postulate that the only reason that African individuals cluster together in unsupervised cluster analysis is that they do not cluster with individuals from other regions, and not because they themselves are a cohesive genetic group. Second, regions are successively nested inside of one another in the GPT. This nested pattern indicates that regional, i.e., putative racial, groups did

not evolve independently of one another, and, therefore, that any taxonomic structure that may exist in our species must consist of successively nested groupings of regions. Third, unsupervised cluster analyses show that the populations near the borders of regions contain substantial ancestry from regions on the other side of the border. This pattern of substantial shared ancestry across regions persisted at all values of *K* that we analyzed (up to *K* = 35), and it is not consistent with the notion of independently evolving biological races. Based on these findings, we reaffirm our previous conclusion that evolution has not produced clusters of genetic variation that correspond to conventional racial groupings (Hunley et al., 2009, 2015), and we categorically reject racial interpretations of cluster analyses of human genetic diversity.

Finally, noting the ubiquity of post-initial-colonization population movements during human evolution, Pickrell and Reich (2014) advocate large-scale analyses of ancient DNA as a means of reconstructing the details of the evolutionary process. While analyses of ancient DNA have and will continue to provide important insights into these processes, our analyses demonstrate that many of these details can be reconstructed by careful analysis of the joint pattern of diversity within and between populations in extant humans.

ACKNOWLEDGMENTS The authors thank Jessica Gross for comments on the manuscript.

Received 27 April, 2016; accepted 26 May 2016.

REFERENCES

Alves, I, M. Arenas, M. Currat, et al. 2016. Long-Distance Dispersal Shaped Patterns of Human Genetic Diversity in Eurasia. *Mol. Biol. Evol.* 33:946–958.

Cann, H. M., Toma C de, L. Cazes, et al. 2002. A human genome diversity cell line panel. *Science* 296:261–262.

Cavalli-Sforza, L. L. 2005. The Human Genome Diversity Project: Past, present and future. *Nat. Rev. Genet.* 6:333–340.

Cavalli-Sforza, L. L. and A. W. Edwards. 1967. Phylogenetic analysis. Models and estimation procedures. *Am. J. Hum. Genet.* 19:233–257.

Cavalli-Sforza, L. L. and A. Piazza. 1975. Analysis of evolution: Evolutionary rates, independence and treeness. *Theor. Popul. Biol.* 8:127–165.

Creanza, N., M. Ruhlen, T. Pemberton, et al. 2015. A comparison of worldwide phonemic and genetic variation in human populations. *Proc. Natl. Acad. Sci.* 112:1265-1272.

DeGiorgio, M., J. H. Degnan, N. A. Rosenberg. 2011. Coalescence-time distributions in a serial founder model of human evolutionary history. *Genetics* 189:579–593.

DeGiorgio, M., M. Jakobsson, N. A. Rosenberg. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc. Natl. Acad. Sci.* 106:16057–16062.

Fitch, W. M. and E. Margoliash. 1967. A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochem. Genet.* 1:65–71.

Friedlaender, J. S., F. R. Friedlaender, F. A. Reed, et al. 2008. The genetic structure of Pacific Islanders. *PLoS Genet.* 4:e19.

Goldstein, D. B., A. Ruiz Linares, L. L. Cavalli-Sforza, et al. 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. U.S.A.* 92:6723–6727.

Handley, L. J. L., A. Manica, J. Goudet, et al. 2007. Going the distance: human population genetics in a clinal world. *Trends Genet.* 23:432–439.

Harpending, H. and A. Rogers. 2000. Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* 1:361–385.

Hintze, J. L. and R. D. Nelson. 1998. Violin Plots: A Box Plot-Density Trace Synergism. *Am. Stat.* 52:181–184.

Hunley, K. and M. Healy. 2011. The impact of founder effects, gene flow, and European admixture on Native American genetic diversity. *Am. J. Phys. Anthropol.* 146:530–538.

Hunley, K. L., Cabana, G. S., J. C. Long. 2015. The apportionment of human diversity revisited. *Am. J. Phys. Anthropol*. :n/a–n/a.

Hunley, K. L., M. E. Healy, J. C. Long. 2009. The global pattern of gene identity variation reveals a history of long-range migrations, bottlenecks, and local mate exchange: Implications for biological race. *Am. J. Phys. Anthropol.* 139:35–46.

Jakobsson, M., S. W. Scholz, P. Scheet, et al. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003.

Kimura, M. and J. F. Crow. 1964. The number of alleles that can be maintained in a finite population. *Genetics* 49:725–738.

Kopelman, N. M., L. Stone, C. Wang, et al. 2009. Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations. *BMC Genet.* 10:80.

Legendre, P. and L. F. J Legendre. 2012. Numerical Ecology, Volume 24, Third Edition. 3 edition. Amsterdam: Elsevier.

Li, J. Z., D. M. Absher, H. Tang, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319:1100–1104.

Long, J. C. and R. A Kittles. 2003. Human genetic diversity and the nonexistence of biological races. *Hum. Biol.* 75:449–471.

Matesanz, S., T. E. Gimeno, M. de la Cruz, et al. 2011. Competition may explain the fine-scale spatial patterns and genetic structure of two co-occurring plant congeners. *J. Ecol.* 99:838–848.

Meirmans, P. G. 2012. The trouble with isolation by distance. *Mol. Ecol.* 21:2839–2846.

Nei, M. 1987. Molecular Evolutionary Genetics. Reprint edition. New York: Columbia University Press.

Novembre, J., T. Johnson, K. Bryc, et al. 2008. Genes mirror geography within Europe. *Nature* 456:98–101.

Oden, N. and R. R. Sokal. 1986. Directional autocorrelation: An extension of spatial correlograms to two dimensions. *Syst. Biol.* 35:608–617.

Ohta, T. and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet. Res.* 22:201–204.

Pemberton, T. J., M. DeGiorgio, N. A. Rosenberg. 2013. Population structure in a comprehensive genomic data set on human microsatellite variation. *G3* 3:891–907.

Pemberton, T. J., F.-Y. Li, E. K. Hanson, et al. 2012. Impact of restricted marital practices on genetic variation in an endogamous Gujarati group. *Am. J. Phys. Anthropol.* 149:92–103.

Pickrell, J. K. and J. K. Pritchard. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8:e1002967.

Pickrell, J. K. and D. Reich. 2014. Toward a new history and geography of human genes informed by ancient DNA. *Trends Genet.* 30:377–389.

Pritchard, J. K., M. Stephens, P. Donnelly. 2000. Inference of population structure Using multilocus genotype data. *Genetics* 155:945–959.

Prugnolle, F., A. Manica, F. Balloux. 2005. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* 15:R159–R160.

Ramachandran, S., O. Deshpande, C. C. Roseman, et al. 2005. 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.* 102:15942–15947.

Relethford, J. 2004. Global patterns of isolation by distance based on genetic and morphological data. *Hum. Biol.* 76:499–513.

Relethford, J. H. 2009. Race and global patterns of phenotypic variation. *Am. J. Phys. Anthropol.* 139:16–22.

Reynolds, J., B. S. Weir, C. C. Cockerham. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* 105:767–779.

Rogers, A. R. and L. B. Jorde. 1996. Ascertainment bias in estimates of average heterozygosity. *Am. J. Hum. Genet.* 58:1033–1041.

Rosenberg, N. A., S. Mahajan, C. Gonzalez-Quevedo, et al. 2006. Low Levels of Genetic Divergence across Geographically and Linguistically Diverse Populations from India. *PLoS Genet.* 2:e215.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, et al. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.* 1:e70.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, et al. 2002. Genetic structure of human populations. *Science* 298:2381–2385.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.

Sesardic, N. 2010. Race: a social destruction of a biological concept. *Biol. Philos.* 25:143–162.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457–462.

Tishkoff, S. A. and K. K. Kidd. 2004. Implications of biogeography of human populations for "race" and medicine. *Nat. Genet.* 36:S21–S27.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, et al. 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035–1044.

Urbanek, M., D. Goldman, J. C. Long. 1996. The apportionment of dinucleotide repeat diversity in Native Americans and Europeans: a new approach to measuring gene identity reveals asymmetric patterns of divergence. *Mol. Biol. Evol.* 13:943–953.

Wade, N. 2014. A Troublesome Inheritance: Genes, Race and Human History. Second Printing edition. Penguin Press.

Wang, S, C. M. Lewis Jr., M. Jakobsson, et al. 2007. Genetic variation and population structure in Native Americans. *PLoS Genet.* 3:e185.

Wang, S., N. Ray, W. Rojas, et al. 2008. Geographic patterns of genome admixture in Latin American mestizos. *PLoS Genet.* 4:e1000037.

Weir, B. S. and C. C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38:1358.

Wilkins, J. F. and F. W. Marlowe. 2006. Sex-biased migration in humans: what should we expect from genetic data? *BioEssays* 28:290–300.

FIGURE LEGENDS

Figure 1. Gene diversity within populations vs. waypoint geographic distances from Addis Ababa, Ethiopia. Populations are colored by geographic region.

Figure 2. Global population tree (GPT) constructed from Nei's minimum genetic distances using the Fitch-Margoliash method. The scale at the bottom is gene diversity. Color coding is the same as in Figure 1.

Figure 3. Gene diversity between populations vs. waypoint geographic distances. Dashes show the gene diversity between populations in different regions. Circles show gene diversity between populations in the same region. The dark gray color shows the gene diversity between

the San and !Xun/Khoe vs. the other 246 populations. Otherwise, colors are the same as in Figure 1.

Figure 4. Plot of ancestry estimates for 5,415 individuals using an unsupervised cluster analysis at *K* = 7. Individuals are arranged by population in the order of their location in the GPT in Figure 2. Regional boundaries are marked with vertical black lines.

Figure 5. Four types of regional trees for the Americas. White slices of the pie charts show the proportion of Native American ancestry. Colored slices show the proportion of non-Native American ancestry determined using unsupervised cluster analysis at K = 7. These latter pie colors correspond to those from Figure 4. (A) Island model. (B) Sub-tree, pruned from the global population tree, with topology and branch lengths intact. (C) Region-only tree, constructed from genetic distances from only the Native American populations; its topology and branch lengths are unconstrained by genetic distances between Native American and non-Native American populations. (D) Un-admixed tree, constructed from genetic distances from only the Native American populations after removing nine populations with more than 10% non-Native ancestry. Population names are colored by geographic sub-region: dark blue – northern North America; light blue – northern Mexico; green – southern Mexico and northern Central Americas; purple – southern Central America and northern South America; black – western South America; red – Colombia and southeastern South America. Asterisks mark the Arhuaco, Wayuu, and Kaingang populations.

Figure 6. Results of GHM model fitting method. The bars show the quantity Λ/degrees of freedom. Lower values represent better fit.

Figure 7. Violin plots comparing residual deviations in fit of the trees in the Americas. Narrower ranges along the x-axis indicate better fit. Circles within each plot show median residual value. The top four plots show the distribution of residuals for the island model, subtree, region-only tree, and un-admixed tree respectively. The remaining plots compare the residuals for nine populations with more than 10% non-Native ancestry in order of the amount of admixture. For each of these populations, the residuals from the sub-tree are on the top, and the residuals from the region-only tree are on the bottom.

Figure 8. Gene diversity vs. geographic distance (top row), Mantel correlograms (middle row) and partial Mantel correlograms (bottom row). The partial Mantel correlograms control for the GHM-estimated gene diversities from the region-only trees. Filled circles represent statistically significant correlations at multiple-test-corrected p-values. The Middle Eastern and European populations are combined into a single region. Colors correspond to those from Figure 1.

FIGURES

Figure 1

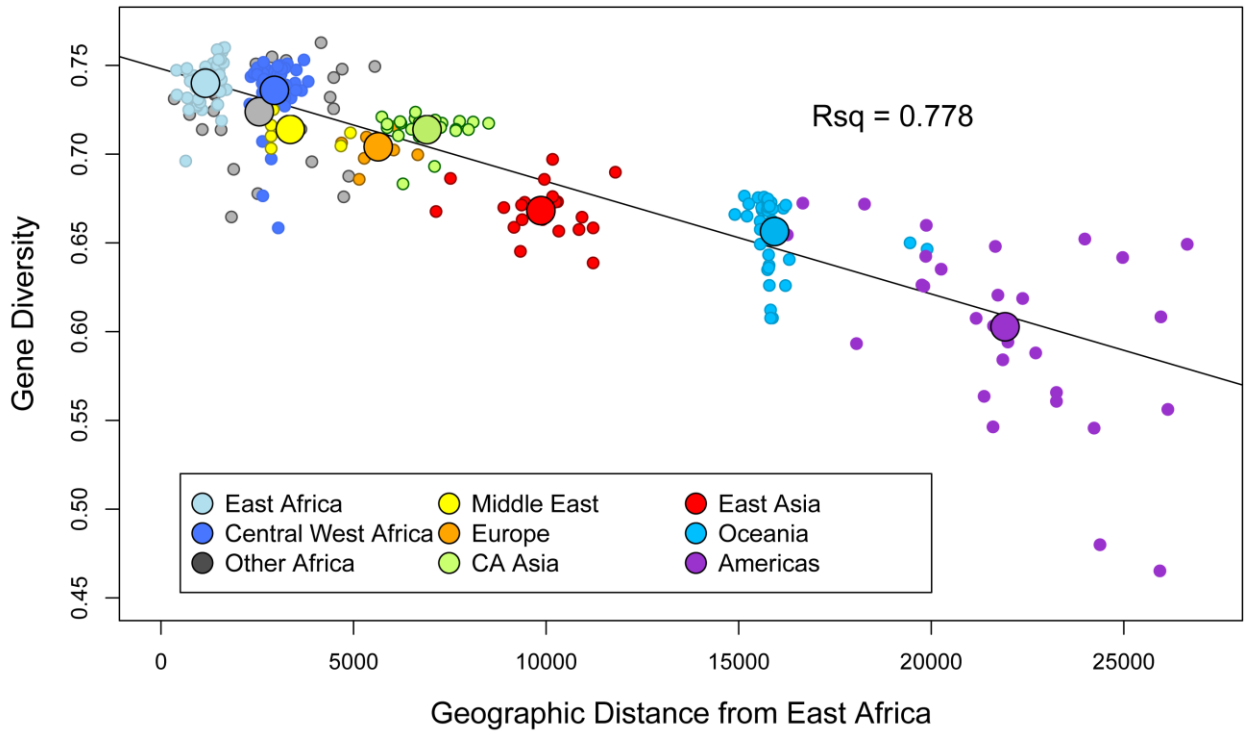

Figure 2

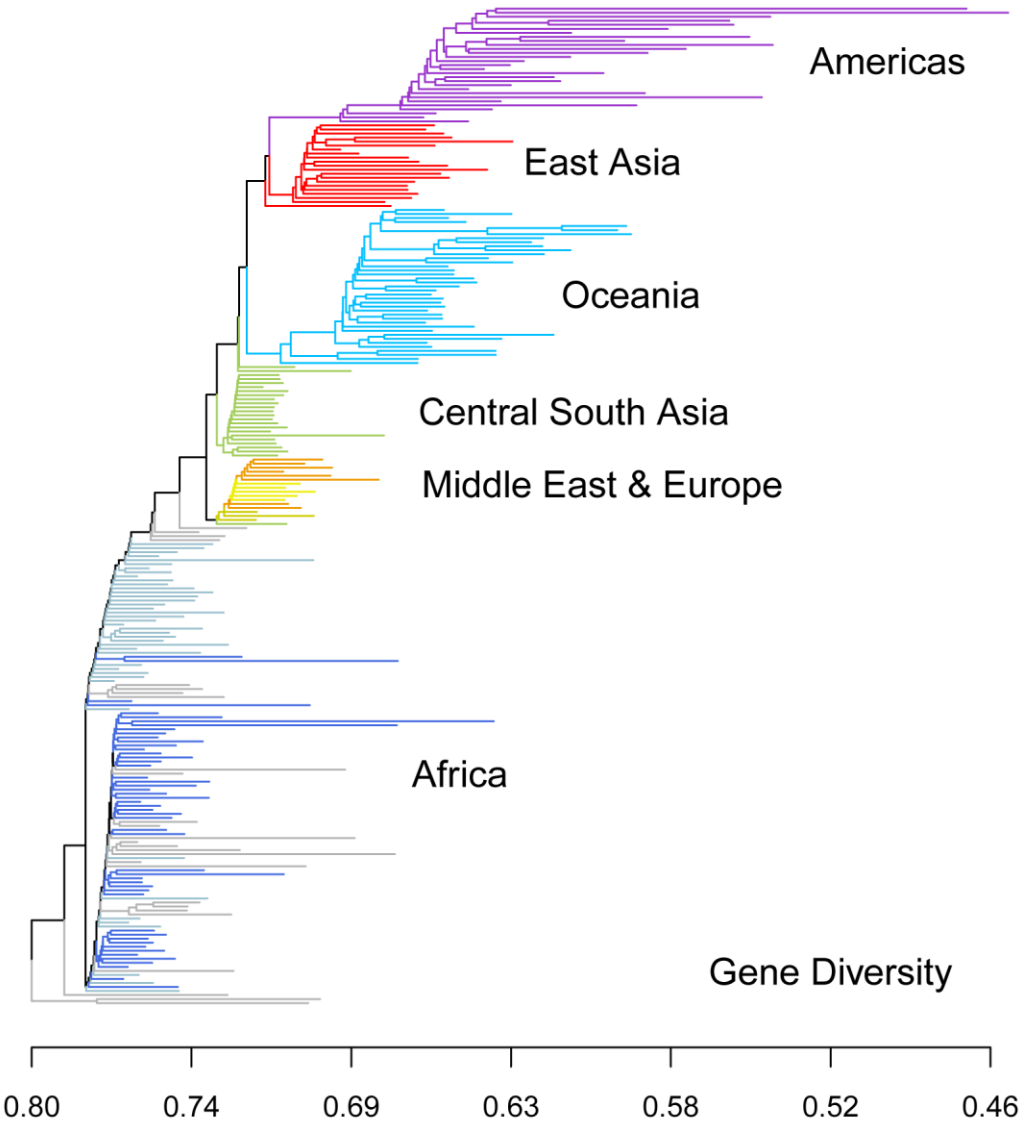

Figure 3

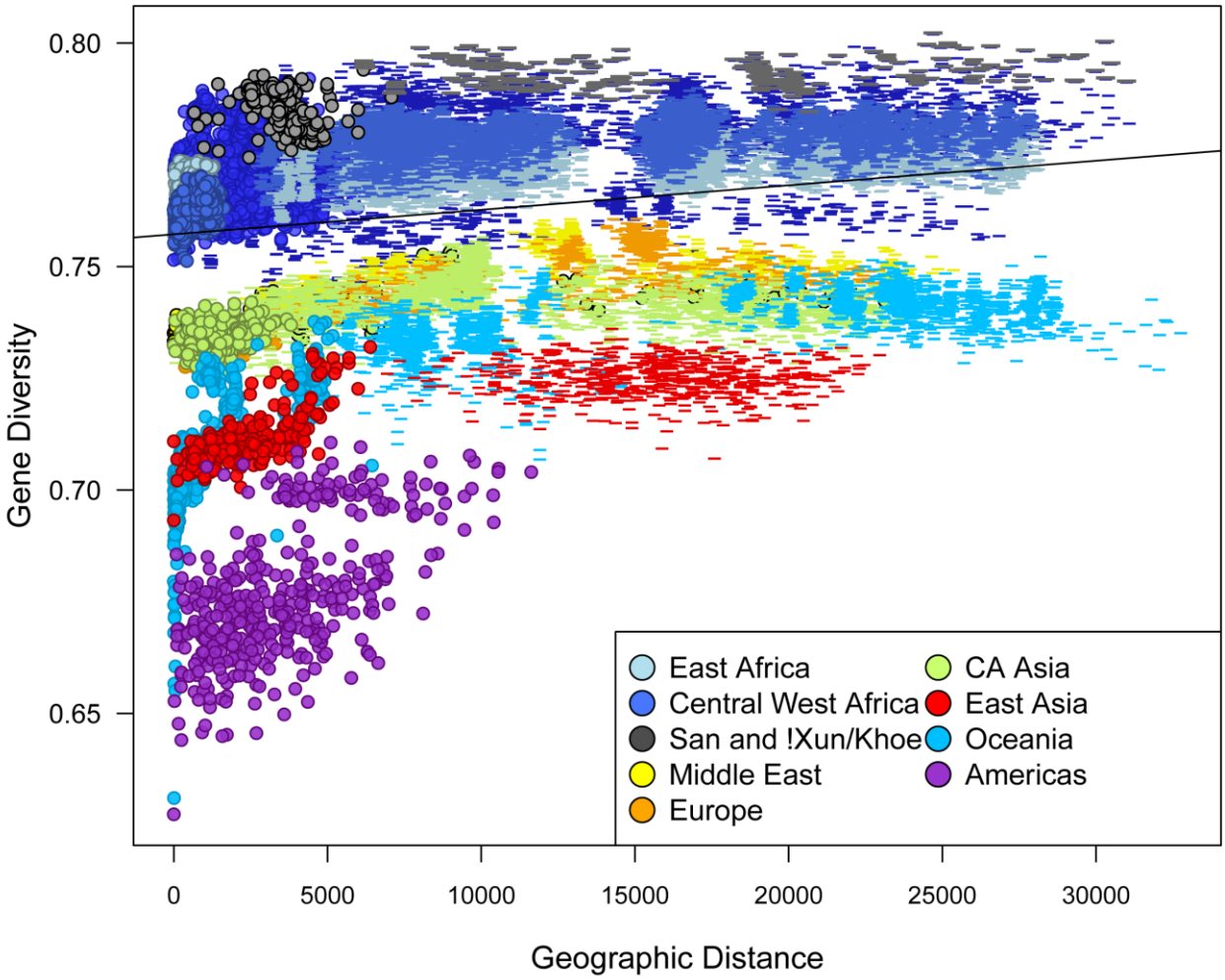

Figure 4

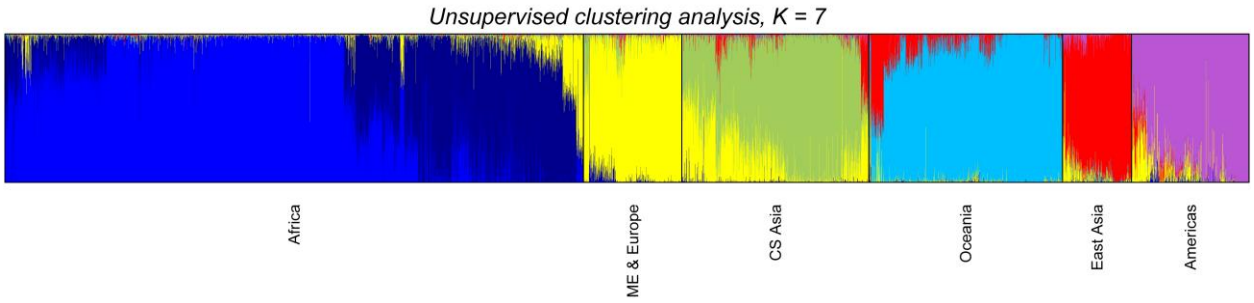

Figure 5

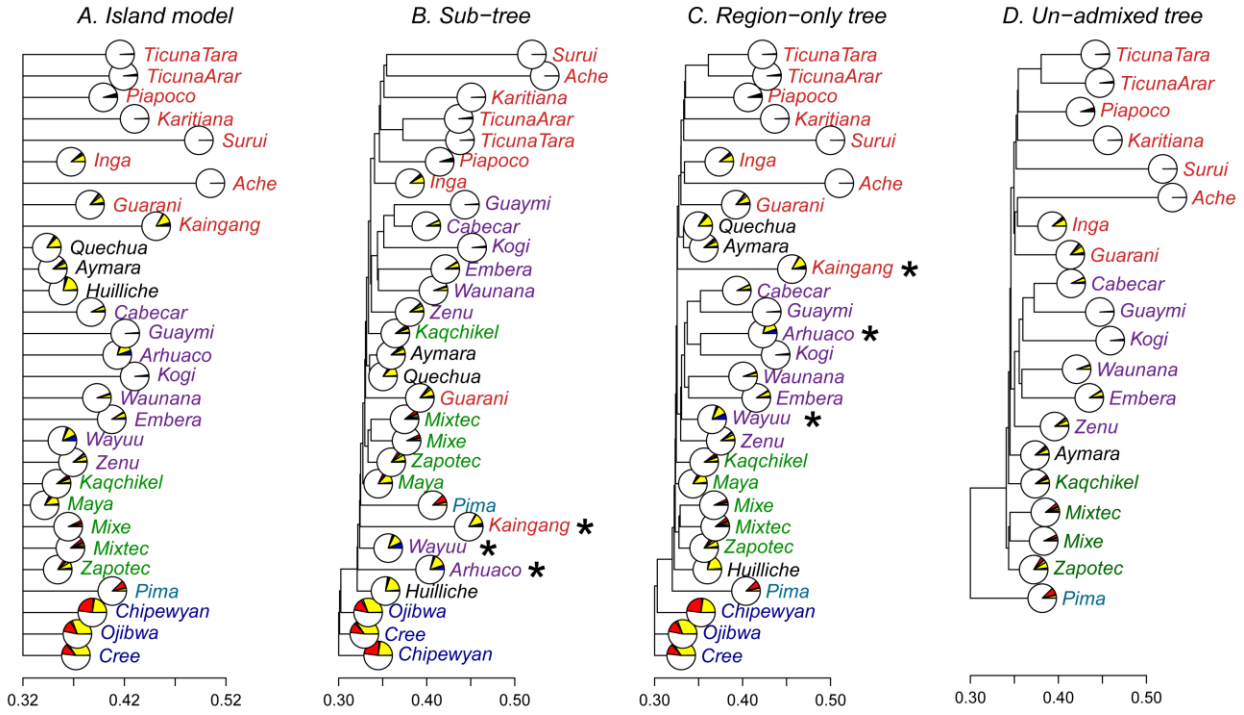

Figure 6

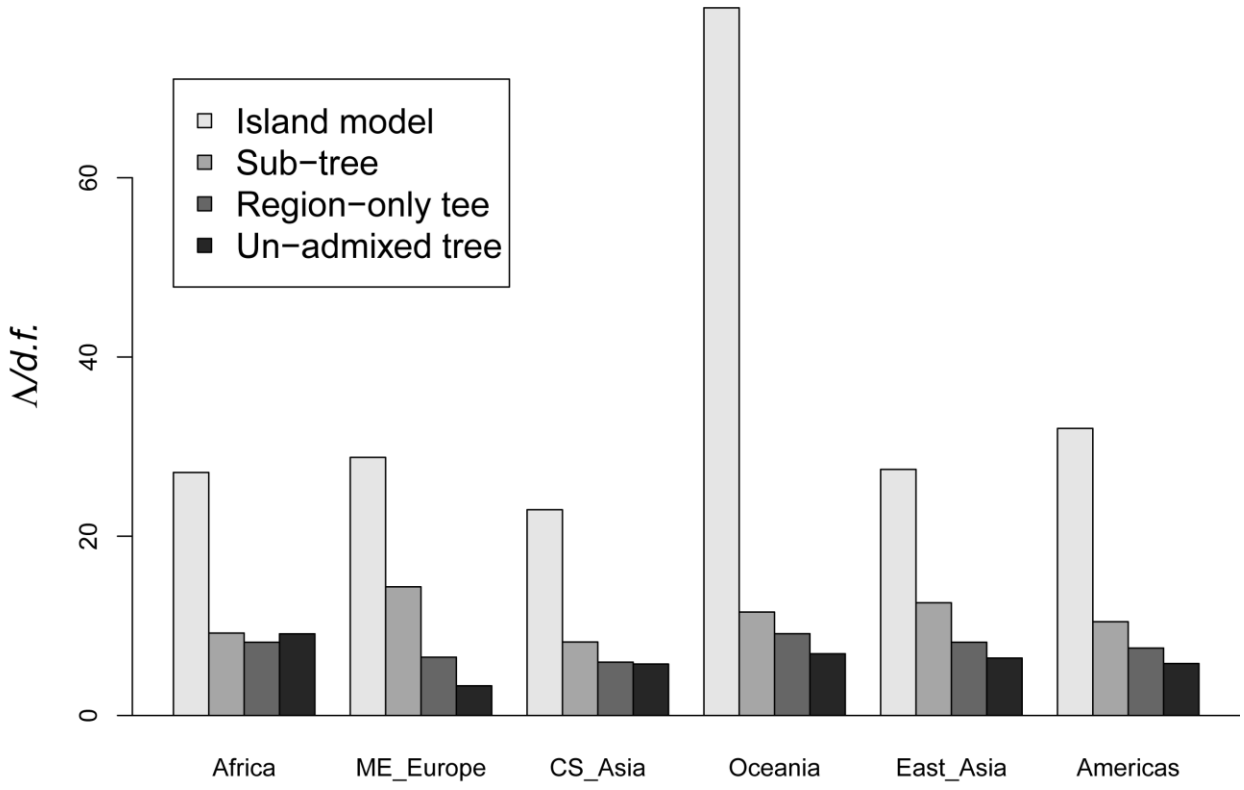

Figure 7

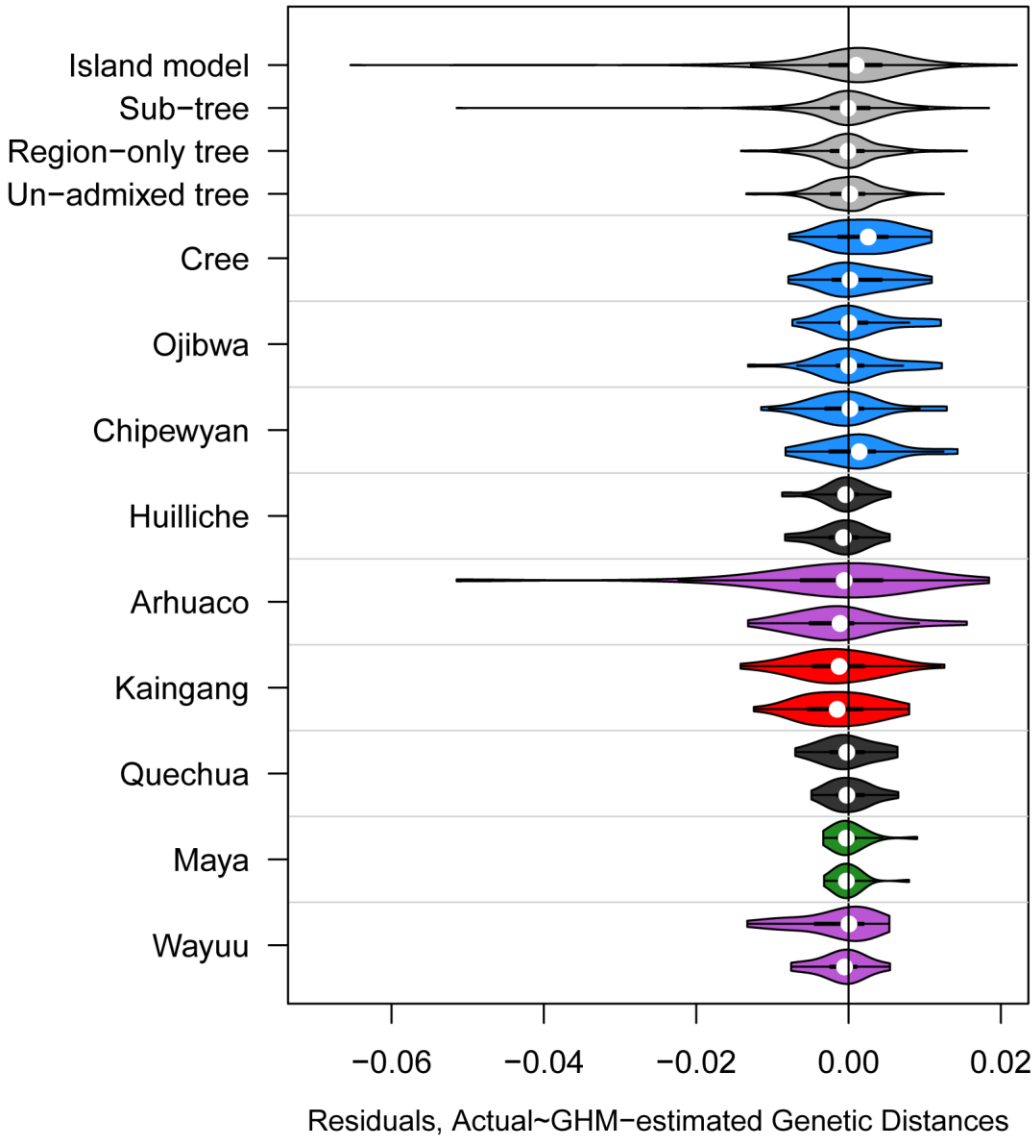

Figure 8