The continent of Africa is the source of all anatomically modern humans that dispersed across the planet during the past 100,000 years. As such, African populations are characterized by high genetic diversity and low levels of linkage disequilibrium (LD) among loci, as compared to populations from other continents. African populations also possess a number of genetic adaptations that have evolved in response to the diverse climates, diets, geographic environments, and infectious agents that characterize the African continent. Recently, Tishkoff et al. (2009) performed a genome-wide analysis of substructure based on DNA from 2432 Africans from 121 geographically diverse populations. The authors analyzed patterns of variation at 1327 nuclear microsatellite and insertion/deletion markers and identified 14 ancestral population clusters that correlate well with self-described ethnicity and shared cultural or linguistic properties. The results suggest that African populations may have maintained a large and subdivided population structure throughout much of their evolutionary history. In this chapter, we synthesize recent work documenting evidence of African population structure and discuss the implications for inferences about evolutionary history in both African populations and anatomically modern humans as a whole.

Africa is a continent of considerable genetic, linguistic, cultural, and phenotypic diversity. It contains more than 2000 distinct ethno-linguistic groups, speaking languages that constitute nearly a third of the world’s languages (http://www.ethnologue.com/). The populations within Africa practice a wide range of subsistence patterns, including various modes of agriculture, pastoralism, and hunting and gathering. Africans also live in climates that range from the world’s largest desert (the Sahara) and second largest tropical rainforest (the Congo Basin) to savanna, swamps, and mountain highlands. This dramatic range in culture, geography, and diet has given rise to a complex history across the African continent, characterized by high levels of both genetic and phenotypic variation.

Africa is also the source of all modern humans, making its populations the oldest and most genetically diverse among the world’s human populations. According to the Recent African Origin (RAO) model, anatomically modern humans originated in Africa and then migrated to all other regions of the globe within the past ~100,000 years (Tishkoff and Verrelli 2003). The transition to modern humans within Africa was not sudden. Rather, the paleobiological record indicates an irregular mosaic of modern, archaic, and regional traits occurring over a substantial period of time and across a broad geographic range (McBrearty and Brooks 2000). The earliest known suite of morphological traits associated with modern humans appears in fossil remains from Ethiopia that are dated to ~150–190 thousand years ago (kya) (White et al. 2003; McDougall et al. 2005). However, this finding does not preclude the existence of modern morphological traits in other African regions before 100 kya; paleobiological specimens from other regions may be less preserved and thus less informative than those discovered in the arid climate of Ethiopia, and extensive archaeological investigations have yet to be conducted across all of Africa (Reed and Tishkoff 2006). A more modern suite of traits appears in East Africa and Southwest Asia ~90 kya, followed by a rapid spread of modern humans throughout the rest of Africa and Eurasia within the past 40,000–80,000 years (Macaulay et al. 2005).

Patterns of genetic variation in modern African populations are shaped by demographic forces that influence variation on a genome-wide scale, such as ancient migration events and fluctuations in population size, and by evolutionary forces that influence individual loci, such as natural selection and mutation. The Bantu expansion is one example that dramatically illustrates the impact of migration on extant patterns of African genetic variation. Within the past ~4000 years, Bantu speakers from West Africa practicing agricultural subsistence migrated throughout sub-Saharan Africa and subsequently admixed with indigenous populations (Ehret 1998; Tishkoff et al. 2009). This expansion greatly influenced genome-wide patterns of genetic variation in modern African populations, an impact that can be observed readily in studies of mitochondrial loci (Soodyall et al. 1996; Behar et al. 2008; Quintana-Murci et al. 2008; Castri et al. 2009), Y-chromosome DNA (Poloni et al. 1997; Hammer et al. 2001), or both (Passarino et al. 1998; Wood et al. 2005; Tishkoff et al. 2007, 2009; Pilkington et al. 2008; Coelho et al. 2009; de Filippo et al. 2009). A classic example of positive natural selection in African populations involves a single-base mutation upstream of FY, the gene encoding the Duffy blood group system. Individuals homozygous for the Duffy-O mutation do not express FY in their bone marrow and are resistant to...
malaria caused by the parasite *Plasmodium vivax* (Miller et al. 1978; Barnwell et al. 1989). The mutation exists at high frequencies in populations from sub-Saharan Africa but is virtually nonexistent elsewhere (Mourant 1976). Because of this unusual geographic distribution, it has long been hypothesized that Duffy-O was subject to selection pressure caused by the presence of either *P. vivax* or some similarly harmful pathogen in prehistoric Africa. Indeed, sequencing studies confirmed that the genetic signature at the *FY* locus is consistent with African-specific positive selection (Hamblin and Di Rienzo 2000; Hamblin et al. 2002).

One of the key demographic forces influencing genome-wide patterns of genetic variation is population structure, i.e., population subdivision, migration, and subsequent admixture. Ancient population structure is a neutral process that can mimic patterns of genetic variation expected under balancing selection, because genetic drift affects allele frequencies in subdivided populations independently and not in the population as a whole. Balancing selection is loosely defined as any locus-specific process that maintains variation within a population (Harris and Meyer 2006); this is in contrast to positive selection, which reduces levels of variation at sites linked to a selected variant. Examples of balancing selection include overdominant selection, where heterozygous genotypes have higher fitness than any of the corresponding homozygous genotypes, and frequency-dependent selection, where the fitness of a particular genotype fluctuates as a function of its frequency in a population. Because loci under balancing selection tend to exhibit an excess of variants at intermediate frequencies, coalescence times for such loci are expected to be significantly longer than those for neutral loci (Slatkin 2000; Navarro and Barton 2002). This property forms the main principle underlying many statistical tests designed to detect balancing selection, such as Tajima’s D statistic (Tajima 1989).

In a structured population, neutral polymorphisms can randomly drift to fixation in some subpopulations but be lost from others, so that the overall population maintains variation longer than expected by chance (Schierup et al. 2000; Muirhead 2001). Therefore, it is possible to reject a model of neutral evolution in favor of balancing selection when in fact the study populations are actually subdivided (Simonsen et al. 1995). Whereas demographic processes such as population structure impact the entire genome, natural selection affects patterns of genetic variation only at localized regions. Thus, it is theoretically possible to disentangle the effects of population structure and balancing selection by examining genome-wide patterns of variation. If a number of loci show similar patterns such as unusually long coalescence times, balancing selection need not be invoked and underlying substructure may, in fact, be a better model.

In this chapter, we synthesize recent work documenting evidence of substructure in African populations, splitting the discussion into studies of single loci and those of genome-wide patterns of variation. Several studies of genetic variation suggest that ancestral populations were geographically structured before the migration of modern humans out of Africa. A model of ancient subdivision within Africa is consistent with observations of divergent LD patterns among African populations (Tishkoff et al. 1996; Tarazona-Santos and Tishkoff 2005), because the stochastic effects of genetic drift can theoretically result in a set of alleles being positively associated in one population but negatively associated in another. Substructure in Africa is likely due to ethnicity, language, and geography as well as technological, ecological, and climatic factors. Such factors may have contributed to population expansions, contractions, fragmentations, and dispersals during recent human evolution in Africa (Mellars 2006; Hassan et al. 2008).

Particularly important to the discussion of African population structure are the timescales on which substructure has occurred. Figure 1 depicts two extreme models of population structure within Africa that are analogs of global human-origin models. The first model is analogous to the multiregional model of human origins (Wolpoff 1996; Wolpoff et al. 2000). Under this model, hominid populations existed in relative isolation across Africa for most of their histories and evolved independently into anatomically modern humans (AMH). Low levels of gene flow across structured populations may have permitted AMH to evolve independently in more than one African region, a scenario that is more feasible on the African continent than it is on a global scale. This model predicts that, in African populations, a significant number of loci have ancient times to the most recent common ancestor (tMRCA), i.e., tMRCA on the order of millions of years, which is much older than the expected tMRCA for a neutrally evolving autosomal locus of ~800,000 years (Harding 1999). The second model is analogous to the recent-origin model of modern humans (Stringer and Andrews 1988; Stringer 2002). Under this model, archaic human populations across Africa were com-

![Figure 1](image-url). Extreme models of population structure within Africa. (Vertical lines) African lineages, (black lines) lineages that have survived to the present day, (gray lines) lineages that have not survived. (A) The first model is analogous to the multiregional model of global human origins. Under this model, hominid populations existed in relative isolation across Africa for most of their histories and evolved independently into anatomically modern humans (AMH). Horizontal arrows indicate low levels of migration between the long-structured populations. (B) The second model is analogous to the recent-origin model of global human origins. Under this model, archaic hominids across Africa were completely replaced with AMH, which then became structured only recently.
that identifies groups of individuals with similar allele frequency profiles while avoiding a priori population classifications. The STRUCTURE algorithm estimates the shared population ancestry of individuals based solely on their genotypes, assuming both Hardy–Weinberg equilibrium and linkage equilibrium in ancestral populations. It is based on a simple admixture model in which \( K \) theoretical ancestral populations gave rise to the individuals under analysis (Weiss and Long 2009). The algorithm places individuals into \( K \) clusters, where \( K \) is chosen in advance but can be varied across independent runs. Each sample in the data set under analysis is subsequently assigned an ancestry proportion from each of the \( K \) clusters (Friedlaender et al. 2008).

The output from STRUCTURE includes statistical support for ancestry assignments, but it is important to note that the results are fundamentally statistical in nature. The more markers and samples in a particular data set, the more fine-scale variation that STRUCTURE can infer, but this does not necessarily correspond to more complex patterns of admixture from real parental populations. The original publication describing the STRUCTURE software (Pritchard et al. 2000) provides the following example to illustrate this point: Suppose a single population has allele frequencies that vary continuously across a large geographic area but is sampled at only \( K \) distinct locations. Provided the allele frequencies at the sampling locations are different enough, STRUCTURE may infer the presence of \( K \) populations in the data set, even though there is a single biological population underlying all \( K \) samples.

STRUCTURE excels at illuminating patterns of variation within and between samples, but it was not designed to make inferences about evolutionary processes that generated those patterns of variation. Such inferences require demographic modeling and hypothesis testing, as described below.

**PHYLOGENETIC ANALYSES**

Before the availability of genome-wide data sets such as those analyzed by Tishkoff et al. (2009) and Patin et al. (2009), population structure in Africa was inferred by examining patterns of genetic variation at single loci. Such studies tended to be phylogenetic in nature: The authors generated trees inferred from the genetic variation and haplotypes that they observed and then estimated tMRCAs for their samples. Unusually long tMRCAs, caused by the presence of one or more highly divergent lineages, suggested that the sampled populations may have been structured throughout their histories. Table 1 lists single-locus studies in which the authors inferred tMRCAs older than 1 million years. The authors of most studies in the table ruled out the effects of balancing selection and thus attributed the long tMRCAs to ancient substructure within Africa.

Although interesting and informative, phylogenetic analyses of single loci are at best only suggestive of ancient population structure. This is because ancient tMRCAs may simply represent the tail of a genome-wide distribution that is expected due to the stochastic nature of evolution acting
at individual loci and thus may not be indicative of admixture between AMH and ancient hominid species. A study by Fagundes et al. (2007) illustrates this idea most effectively. The authors sequenced 500 bp at each of 50 unlinked, noncoding, autosomal loci in samples from 10 Africans, eight Asians, and 12 Native American individuals. With the data, the authors were able to test different models of modern human origins and found that a recent African replacement model best explained the patterns of genetic variation. Under their model, AMH arose ~141 kya in Africa from a founder population of ~12,800 effective individuals and then went on to replace all other hominid species. The authors also computed empirical distributions for tMRCAs, and they found that coalescence times exceeding several million years are expected under the recent African replacement model due to the stochasticity of lineages that might pass through an AMH bottleneck. Fagundes et al. (2007) concluded that a complete replacement of archaic humans during the evolution of AMH could have resulted in ancient tMRCAs for some loci, making it unnecessary to invoke admixture with archaic humans to explain such observations. The model relies on a large population size ancestral to AMH within Africa, an assumption that is consistent with findings from prior studies (Tishkoff and Williams 2002). Although it is true that the sequence data from Fagundes et al. (2007) represent short genomic regions and a sparse sampling of human variation (Garrigan and Hammer 2008), the results nevertheless show that it is possible to obtain extremely long tMRCAs under a model of recent African replacement.

A simulation study published by Wall (2000) formalizes the need for genome-wide data sets in this context. The study expands a model of population subdivision originally developed by Nordborg (2000) to show that ~50–100 unlinked, neutrally evolving, fully resequenced loci are necessary to have sufficient power for detecting ancient contributions in the human genome as the assumed separation time increases and decreases as the assumed time to admixture increases. The results should apply to detecting ancient substructure in African populations as well, implying that numerous loci are needed to make definitive statements about ancient population structure within Africa.

### Table 1. Phylogenetic Studies that Infer Ancient tMRCAs for African Populations

<table>
<thead>
<tr>
<th>Gene/locus</th>
<th>Reference</th>
<th>Sample size</th>
<th>Length</th>
<th>Estimated tMRCAs (millions of years)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>Zhao et al. (2000)</td>
<td>20</td>
<td>6</td>
<td>10 kb</td>
</tr>
<tr>
<td>CD209</td>
<td>Barreiro et al. (2005)</td>
<td>41</td>
<td>9</td>
<td>5.5 kb</td>
</tr>
<tr>
<td>CMAH (intronic)</td>
<td>Hayakawa et al. (2006)</td>
<td>13</td>
<td>3</td>
<td>7.3 kb</td>
</tr>
<tr>
<td>ASAH</td>
<td>Kim and Satta (2008)</td>
<td>15</td>
<td>3</td>
<td>11 kb</td>
</tr>
<tr>
<td>PDHA1</td>
<td>Harris and Hey (1999)</td>
<td>16 males</td>
<td>4</td>
<td>4.2 kb</td>
</tr>
<tr>
<td>Intergenic</td>
<td>Shimada et al. (2007)</td>
<td>95 males</td>
<td>6</td>
<td>10.1 kb</td>
</tr>
<tr>
<td>Intergenic</td>
<td>Garrigan et al. (2005a)</td>
<td>10 males</td>
<td>6</td>
<td>7.5 kb spanning 17.5 kb</td>
</tr>
</tbody>
</table>

**MODELING DEMOGRAPHIC HISTORY**

Although the STRUCTURE program analyzes patterns of variation to detect the presence of population structure, statistical hypothesis testing is necessary to model the demographic and evolutionary processes that may have led to structured human populations. This kind of statistical modeling has typically been done using coalescent theory (Hudson 1991) as implemented in the GENETREE software package (Griffiths and Tavaré 1994). The GENETREE algorithm estimates the tMRCAs for a sample of sequences under assumptions of random mating and either constant population size or a recent population expansion. This approach has been used in a variety of studies during the past decade (Harris and Hey 1999; Jaruzelska et al. 1999; Yu et al. 2002; Barreiro et al. 2005; Hayakawa et al. 2006; Yotova et al. 2007). To provide statistical evidence of population structure, some authors have also tested assumptions of panmixia directly using custom-designed coalescent simulations (Garrigan et al. 2005a; Kim and Satta 2008). In particular, Kim and Satta (2008) tested 22 sets of demographic parameters to examine models of panmixia, bottlenecks, expansions, and population structure (both ancient and recent) for an 11-kb region of the ASAH gene and found that a model of ancient structure in Africa best explained their data.

It is sometimes difficult to formally model demographic scenarios because the number of parameters that must be estimated becomes prohibitively large as the models become more complex. Approximate Bayesian computation, or ABC, is one approach that allows for flexible yet statistically sound comparisons of different demographic models (Beaumont et al. 2002). ABC is Bayesian in the sense that it estimates posterior probability distributions for the parameters of interest, given a demographic model and a set of prior distributions. It is approximate in two possible ways, depending on the implementation. First, because exact posterior distributions are often too complicated to
calculate explicitly, ABC constructs approximate distributions numerically using stochastic simulation methods such as rejection algorithms, importance sampling, and Markov chain Monte Carlo (summarized in Marjoram and Tavaré 2006). A stochastic method simulates a data set based on the given demographic model and a parameter value sampled from the prior distribution. If the simulated data are sufficiently close to the actual data, the parameter value is stored. After simulating many data sets in this way, a posterior distribution can be built from frequencies of the stored parameter values. Second, ABC is approximate in the sense that comparisons between simulated and real data sets are often quantified using summary statistics such as the number of segregating sites. Many researchers have developed custom ABC implementations to analyze their data, although user-friendly software packages are becoming more common (see, e.g., Cornuet et al. 2008; Lopes et al. 2009).

Cox et al. (2008) used ABC to show that the pattern of variation at the X-linked pseudogene RRM2P4 is best explained by population structure dating to 2.33 million years and rooted in East Asia. Building on previous work (Garrigan et al. 2005b), the authors sequenced 2.4 kb of the pseudogene in 131 Africans and 122 non-Africans and extended the region by sequencing two additional nearby loci, for a total 5.6 kb of resequencing data that spans 16.5 kb of the X chromosome. With the data, the authors explicitly tested whether an RAO model could explain both the ancient tMRCA that they observed and the basal lineage rooted in East Asia. They found that although an RAO model may conceivably explain the extremely long tMRCA, a model of ancient admixture best explains the East Asian root for the genealogy. The authors thus concluded that RRM2P4 may be a remnant of admixture between AMH and an archaic hominid population in Asia such as H. erectus. ABC has also been recently used to compare several demographic scenarios that may have produced the patterns of genetic variation seen in contemporary African Pygmy populations (Patin et al. 2009; Verdu et al. 2009). In particular, Patin et al. (2009) used an ABC approach to estimate separation times and levels of gene flow between Western and Eastern Pygmy populations. Notably, this study used ABC in conjunction with a genome-wide data set consisting of 24 unlinked noncoding nucleotide sequences across the autosomes, sex chromosomes, and mitochondrial genome. The most likely model explaining the data involves a split ~60 kya between an ancestral Pygmy population and a population ancestral to modern-day African farmers, followed by a split between the Western and Eastern Pygmy populations ~20 kya.

A statistical framework that infers demographic parameters specifically related to population structure is the isolation with migration model, implemented in the software package IM (Nielsen and Wakeley 2001). The IM algorithm computes marginal Bayesian posterior probabilities for a suite of parameters including population sizes, migration rates, and divergence times, allowing inference of population structure both with and without the effects of migration. This approach was used by Shimada et al. (2007) to analyze 10.1 kb of resequencing data from a noncoding region on chromosome Xp11.22. The authors sequenced the locus in a panel of 672 males from 52 worldwide populations and found a surprisingly divergent haplotype distributed at low frequencies throughout Africa, Europe, and Asia. Using IM, they estimated a tMRCA of 5230 years for the haplotype, but a tMRCA of more than 1.4 million years between the haplotype and all other sequences in their sample, suggestive of archaic population structure in Africa. Although the IM approach is particularly sophisticated in its ability to test models of population structure using multilocus data sets (Hey and Nielsen 2004), it does not currently account for recombination within loci.

CONCLUSIONS

Traditionally, inferences about population structure in Africa have relied on analyses of single loci. Because population structure can affect genome-wide patterns of variation, an unusually old tMRCA at a single locus is only suggestive of ancient substructure. Extremely old tMRCAs are expected in recent replacement models that do not involve ancient population structure and may therefore represent the tails of the distribution for genome-wide tMRCAs. For this reason, large data sets consisting of hundreds of resequenced loci are needed to make definitive inferences about ancient population structure in Africa. Such data sets will allow researchers to model demographic scenarios leading to structured human populations in a statistically sound way. In particular, ABC and/or IM can be used with more genome-wide data sets to formally test whether current data are consistent with a model of ancient African substructure (Fig. 1A), a recent replacement model (Figure 1B), or demographic models intermediate between the two.

Genome-wide data sets of single-nucleotide polymorphism (SNP) genotypes, copy-number variants, and other forms of structural variation should also prove informative for elucidating patterns of genetic variation in African populations. For example, Bryc et al. (2010) genotyped 500,000 SNPs in 203 individuals from 12 West African populations and found that population structure in this region of Africa reflected both linguistic and geographic variation. It is important to note, however, that currently available SNP platforms such as the Illumina Human1M-Duo or Affymetrix Genome-Wide Human SNP Array 6.0 are derived from SNPs identified predominantly in non-African populations. This introduces an ascertainment bias when studying genetic variation in African populations. Resequencing studies of ethnically diverse African populations will identify African-specific SNPs that can subsequently be used to create SNP platforms that are more informative for analyses of African populations. Recent advances in sequencing technologies—in particular, next-generation sequencing, exome sequencing, and whole-genome sequencing—will also be extremely valuable for understanding fine-scale patterns of variation in African populations. These technologies will provide complete sequence information at multiple loci, allowing genome-wide inferences to be made about substructure in Africa. To date, the whole genomes of three African men and one
African woman have been sequenced: one Yoruban male (Bentley et al. 2008) and one Yoruban female (Drnacanac et al. 2010), an indigenous hunter-gatherer from the Kalahari Desert, and a Bantu individual from southern Africa (Schuster et al. 2010). As the cost of sequencing technologies continues to decrease, it will become possible to conduct population-level analyses of ethnically diverse groups of African populations.

Finally, a thorough understanding of African evolutionary history will require sampling across a broad range of African populations. Much of what is currently known about African genetic diversity is inferred from a limited number of the ∼2000 linguistically distinct ethnic groups in Africa. Extensive sampling of ethnically diverse African populations will be critical for testing models of the origin and dispersal of modern humans both within and outside of Africa. As the data sets become more diverse in terms of both population samples and genotyped loci, fine-scale inferences about African demographic history will become feasible.

ACKNOWLEDGMENTS

C.A.L. is supported by an National Institutes of Health (NIH) IRACDA postdoctoral fellowship (PAR-06-470). S.A.T. is supported by National Science Foundation R01GM076637 and 1R01GM083606-01, and an NIH Pioneer 1-DP1-OD-006445-01 award.

REFERENCES


GENETIC STRUCTURE IN AFRICAN POPULATIONS 401


Genetic Structure in African Populations: Implications for Human Demographic History

C.A. Lambert and S.A. Tishkoff

Cold Spring Harb Symp Quant Biol 2009 74: 395-402 originally published online May 7, 2010
Access the most recent version at doi:10.1101/sqb.2009.74.053

References
This article cites 74 articles, 19 of which can be accessed free at:
http://symposium.cshlp.org/content/74/395.full.html#ref-list-1

License

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to Cold Spring Harbor Symposia on Quantitative Biology go to:
http://symposium.cshlp.org/subscriptions

Copyright © 2009, Cold Spring Harbor Laboratory Press