Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Ancestry inference for two admixed populations African American and African Caribbean

Mingjia Yao [1] **and Wenlan Pan** [1]

[1] Department of Biostatistics,Fielding School of Public Health, UCLA, Los Angeles, California, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Genetic diversity demonstrates that the African continent is the birthplace of modern humans, and there exists significant genetic variance between West and East Africa populations. Therefore, the research question is whether the ancestors of African Americans in the southwest US (ASW) and African Caribbean in Barbados (ACB) are genetically closer to the population of West Africa or East Africa.
Results: The African American and African Caribbean populations are genetically closest to the populations surround the Gulf of Guinea in west Africa.
Availability: Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text
Contact: wenlanpan@g.ucla.edu mingjia96@ucla.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Investigating the ancestry inference and population structures play an essential role in current research since their understanding helps several further studies. For instance, they can benefit inheritance tracing, disease susceptibility gene discovery, and the genome-wide association study (GWAS) that require control for the population stratification (Benn-Torres *et al.*, 2008). Genetic variations, which are the differences in gene frequencies among individuals and populations, are utilized to conduct the ancestry inference. Individuals within the same population are more likely to share the same genetic variations since they experienced coevolution (Nielsen, 2004). The single nucleotide polymorphism (SNP) is one of the genetic variations that commonly used. Ancestry informative markers (AIMs) containing ancestry informative SNPs (AISNPs) that display significant variation in different populations mainly contribute to perform ancestry inference (Benn-Torres *et al.*, 2008).

A more extended history of genetic evolution is always associated with more gene diversity. Rich patterns of gene diversity provide evidence that African ancestry experienced the most extended evolutionary history, and the African continent is the place where modern humans originated (Choudhury *et al.*, 2020). The more genetic variation of African genomes than any other continents leads to the significant genetic difference among its subpopulations such as West and East African (Chaichoompu *et al.*, 2020). Besides subpopulations in the African continent, two admixed populations are also derived from Africa, and they are the African American and the African Caribbean. They are all significant components in their regions as African Americans account for 13.4% of the United States population (Bureau, 2011) and African Caribbean compromise 95.5% of Barbadian population (Central-Intelligence-Agency, 2020). They are mosaic populations with African ancestry and other ancestries such as those of European and South Asian. One research focused on the admixture components of these two populations from three super populations African, European, and Asian (Murray *et al.*, 2010). However, not many studies investigate the genetic relationship between these two admixed populations and the subpopulations in the African continent.

We would like to analyze the genetic variation data (SNPs) to trace the complex dispersal of modern humans out of Africa and their population expansion. Therefore, the research question of this study is the ancestors of African Americans in the southwest US and the African Caribbean in Barbados are genetically closer to the population of West Africa or East Africa. Two dimensional reduction methods principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were applied to analyze the AISNPs from the 1000 Genomes Project phase 3. Based on the plots obtained from PCA and t-SNE, geographic relationships between the populations can be determined, and the ancestors of African Americans in the southwest US (ASW) and the African Caribbean in Barbados (ACB) can be inferred.

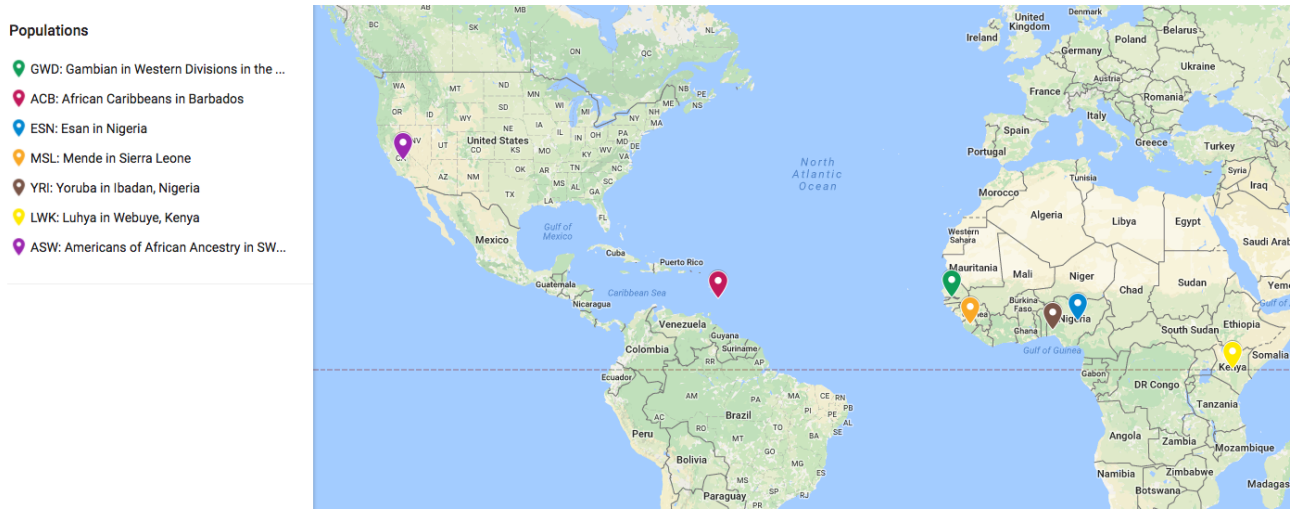

Fig. 1. Georaphical locations of the subpopulations of Africa.

Table 1. Population composition of the super populations.

AMR	EAS	EUR	SAS	AFR
347	504	503	489	661

AMR: American, EAS: East Asian, SAS: South Asian, AFR: African

Table 2. Population composition of the subpopulations in Africa.

ACB	ASW	GWD	MSL	ESN	YRI	LWK
96	61	113	85	99	108	99

ACB: African Caribbeans in Barbados, ASW: Americans of African Ancestry in SW USA, GWD: Gambian in Western Divisions in the Gambia, MSL: Mende in Sierra Leone, ESN: Esan in Nigeria, YRI: Yoruba in Ibadan, Nigeria, LWK:Luhya in Webuye, Lenya

2 Methods

2.1 Samples

We used the data from the 1000 Genomes Project phase 3, which provides a global reference for human genetic variation (mostly SNPs) (Consortium *et al.*, 2015). This database contains 2,504 individuals covering five super populations - African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). Also, it includes 661 samples from Africa that we mainly focused on - subpopulations from West Africa "Gambian in Western Divisions in the Gambia" (GWD), "Mende in Sierra Leone" (MSL), "Esan in Nigeria" (ESN), "Yoruba in Ibadan, Nigeria" (YRI) and one subpopulation from East Africa "Luhya in Webuye, Lenya" (LWK). Additionally, two admixed populations of our interests "African Caribbeans in Barbados" (ACB) and "Americans of African Ancestry in SW USA" (ASW). Tabel 1 and Tabel 2 shows the population composition of the dataset, and Figure 1 demonstrates the geographical locations of the subpopulations of Africa.

2.2 SNP genotypes

We utilized several published ancestry informative SNPs (AISNPs) exhibiting substantially different frequencies between diverse populations since they are distinguishable for populations. Firstly, we merged two datasets, which were 55 AISNPs from Kidd *et al.*, 2014 and 128 AISNPs from Kosoy *et al.*, 2009. After removing the duplication, this dataset consisting of 170 AISNPs was used as a trial with limited SNPs. The second dataset, including 2,302 AISNPs, is the primary dataset (Mikblack, 2017). Besides, a dataset covering the results of PCAs from a previous study was employed for the comparison with our results (Chrisporras, 2019).

2.3 Dimensional reduction

In the study, two different dimensional reduction approaches were utilized to determine the origin of the two admixed populations ACB and ASW. The first one is the most commonly used principal component analysis (PCA), which is a linear mathematical technique. It seeks to preserve the global structure of the data by rotating the vectors to keep variance. The second one is t-Distributed Stochastic Neighbor Embedding (t-SNE), which is a non-linear probabilistic technique. It tries to maintain local similarities by prioritizing neighboring points. These two dimensional reduction methods were implemented by prcomp and rtsne function in r.

Applied PCA and t-SNE, we first compared the ACB and ASW populations with five super populations to confirm their closest super population. Then we compare them with the other five subpopulations in Africa to have a closer look at their migration from the African continent. Finally, we contrasted the admixed ACB and ASW populations with subpopulations in Europe and America, respectively, to determine their contributions to these two admixed populations.

3 Results and discussion

3.1 ACB, ASW and four super populations (AFR, AMR, EUR, ASA)

From the PCA plot, we can see the ACB and ASW populations are at the edge of the AFR superpopulation cluster (Figure 2). We can also see the AMR super population is the most dispersed one. From the t-SNE plot, one thing interesting is the AMR population was divided into two clusters (Figure 3). The ACB and ASW populations are still at the edge of the AFR super population cluster, and close to one of the two AMR clusters.

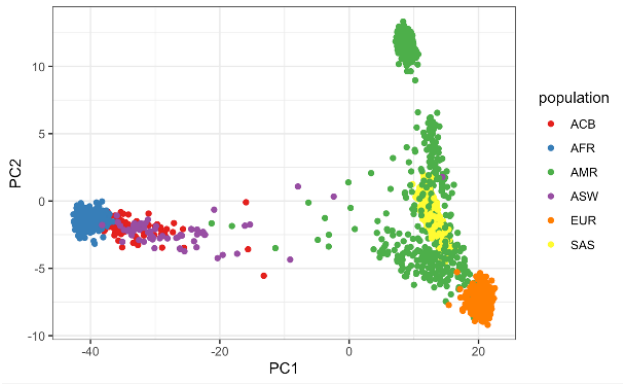

Fig. 2. PCA plot of ACB, ASW and four super populations.

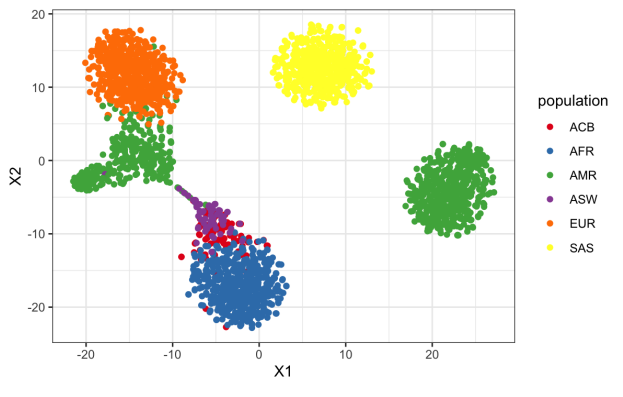

Fig. 3. t-SNE plot of ACB, ASW and four super populations.

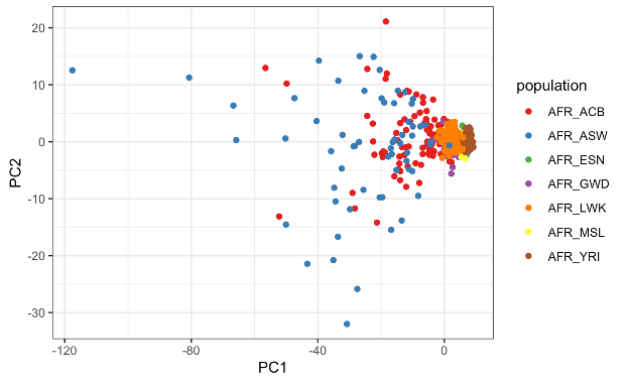

Fig. 4. PCA plot of ACB, ASW and five African subpopulations.

3.2 ACB, ASW and five African subpopulations (ESN, GWD, LWK, MSL, YRI)

From the PCA plot, the only apparent thing we can see is the ACB and ASW populations are very dispersed on the plot, and all the other 5 African populations are gathered very closely (Figure 4). We cannot see a clear geographical pattern on the plot.

We found that t-SNE method gives a clearer geographical pattern than the PCA method (Figure 5). The clusters from top to the bottom of the plot indicated the populations from East to West Africa. Along the X2 axis, we can see the populations: LWK, which is sampled from Kenya in the east Africa, is located at the top of the plot. Then, as the population go west to the Gulf of Guinea, the ESN and YRI populations from Nigeria are below the LWK population. Then we go to the west Africa, we can see the MSL population from Sierra Leone below the two Nigeria populations. Finally, the GWD population from Gambia, which is located at the west end of Africa, is at the bottom of the plot. Therefore, we may assume that the X2 axis represents the geographical pattern of Africa population from east to west. We may conclude that the ACB and ASW populations are genetically closest to these three populations surrounding the Gulf of Guinea in West Africa.

However, it is also apparent that the African American populations are actually distinct from any of those five African populations. This may be because of the fact that we don't have enough populations' data or maybe because the African American populations underwent an admixture process after they left West Africa hundreds of years ago.

Then, we compared the ACB and ASW populations to EUR super population and AMR super population respectively, try to understand how these African American populations were formed.

Fig. 5. t-SNE plot of ACB, ASW and four super populations.

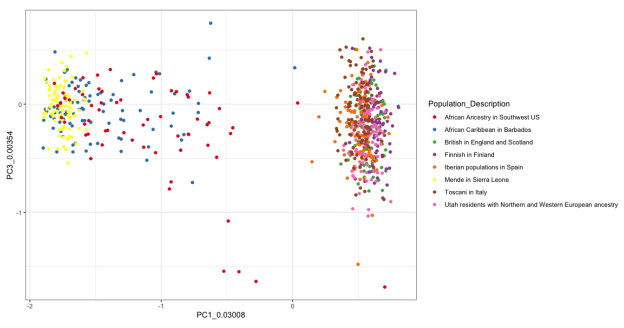

Fig. 6. PCA plot of ACB, ASW, MSL and European subpopulations.

3.3 ACB, ASW, MSL and European subpopulations

We extracted the MSL population which is assumed to be genetically closest to the two African American populations, together with these European populations to make the PCA and t-SNE plots (Figure 6 and Figure 7). We can see that, on both plots, the European populations are very far away from the African American populations. It means the European populations hardly made any contribution to the development of the new African American populations.

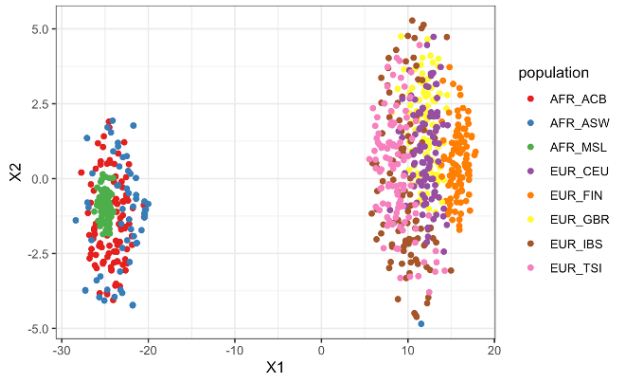

Fig. 7. t-SNE plot of ACB, ASW, MSL and European subpopulations.

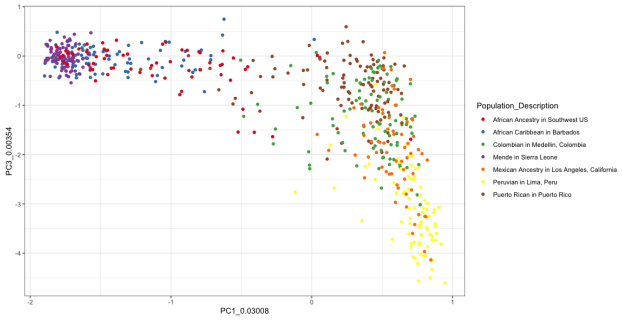

Fig. 8. PCA plot of ACB, ASW, MSL and American subpopulations.

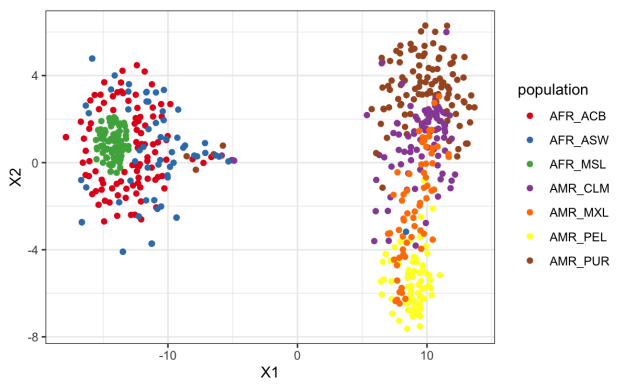

Fig. 9. t-SNE plot of ACB, ASW, MSL and American subpopulations.

3.4 ACB, ASW, MSL and American subpopulations

On the PCA plot, we can see the two African American populations are relatively close to the Puerto Rican and Columbian populations (Figure 8 and Figure 9). We still cannot understand how the African American populations were developed based on this plot, but we can say that they underwent a similar admixture history as these central American populations.

4 Conclusion

From the t-SNE plot we generated, based on the geographical pattern we saw on the plot. We conclude that the ancestors of African Americans in southwest US (ASW) and African Caribbean in Barbados (ACB) populations are genetically closest to these three populations surrounding the Gulf of Guinea in West Africa. However, the ASW and ACB populations are distinct populations from any of the African populations that we compared with, and we found that they underwent a similar admixture history as the Puerto Rican and Columbian populations in America. If we want to have a deeper understanding about how the African American and African Caribbean populations were formed, we may need more detailed data.

Acknowledgements

Text Text Text Text Text Text Text Text. text text text

Funding

This work has been supported by the... Text Text Text Text.

References

Benn-Torres, J. *et al.* (2008). Admixture and population stratification in african caribbean populations. *Annals of human genetics*, **72**(1), 90–98.

Bureau, U. C. (2011). 2010 census shows black population has highest concentration in the south.

Central-Intelligence-Agency (2020). Central america :: Barbados — the world factbook - central intelligence agency. https://www.cia.gov/library/publications/the-world-factbook/geos/bb.html. (Accessed on 12/14/2020).

Chaichoompu, K. *et al.* (2020). A different view on fine-scale population structure in western african populations. *Human genetics*, **139**(1), 45–59.

Choudhury, A. *et al.* (2020). High-depth african genomes inform human migration and health. *Nature*, **586**(7831), 741–748.

Chrisporras (2019). chrisporras/1000genomes-pca. =https://github.com/chrisporras/1000Genomes-PCA. (Accessed on 12/14/2020).

Consortium, . G. P. *et al.* (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68–74.

Kidd, K. K. *et al.* (2014). Progress toward an efficient panel of snps for ancestry inference. *Forensic Science International: Genetics*, **10**, 23–32.

Kosoy, R. *et al.* (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in america. *Human mutation*, **30**(1), 69–78.

Mikblack (2017). mikblack/msg-pca-20171212. https://github.com/mikblack/msg-pca-20171212/tree/ae4d3466daaee6d3b100b87b3c898b7b6cd4abd5. (Accessed on 12/14/2020).

Murray, T. *et al.* (2010). African and non-african admixture components in african americans and an african caribbean population. *Genetic epidemiology*, **34**(6), 561–568.

Nielsen, R. (2004). Population genetic analysis of ascertained snp data. *Human genomics*, **1**(3), 1–7.