Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

A combined evidence Bayesian method for human ancestry inference applied to Afro-Colombians

Lavanya Rishishwar [a,b,c], Andrew B. Conley [a], Brani Vidakovic [d], I. King Jordan [a,b,c,*]

[a] School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA
[b] PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia
[c] BIOS Centro de Bioinformática y Biología Computacional, Manizales, Caldas, Colombia
[d] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

A B S T R A C T

Uniparental genetic markers, mitochondrial DNA (mtDNA) and Y chromosomal DNA, are widely used for the inference of human ancestry. However, the resolution of ancestral origins based on mtDNA haplotypes is limited by the fact that such haplotypes are often found to be distributed across wide geographical regions. We have addressed this issue here by combining two sources of ancestry information that have typically been considered separately: historical records regarding population origins and genetic information on mtDNA haplotypes. To combine these distinct data sources, we applied a Bayesian approach that considers historical records, in the form of prior probabilities, together with data on the geographical distribution of mtDNA haplotypes, formulated as likelihoods, to yield ancestry assignments from posterior probabilities. This combined evidence Bayesian approach to ancestry assignment was evaluated for its ability to accurately assign sub-continental African ancestral origins to Afro-Colombians based on their mtDNA haplotypes. We demonstrate that the incorporation of historical prior probabilities via this analytical framework can provide for substantially increased resolution in sub-continental African ancestry assignment for members of this population. In addition, a personalized approach to ancestry assignment that involves the tuning of priors to individual mtDNA haplotypes yields even greater resolution for individual ancestry assignment. Despite the fact that Colombia has a large population of Afro-descendants, the ancestry of this community has been understudied relative to populations with primarily European and Native American ancestry. Thus, the application of the kind of combined evidence approach developed here to the study of ancestry in the Afro-Colombian population has the potential to be impactful. The formal Bayesian analytical framework we propose for combining historical and genetic information also has the potential to be widely applied across various global populations and for different genetic markers.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

The desire to trace one's family ancestry and origins, i.e., the study of genealogy, is an ancient human impulse (Potter-Phillips, 1999). People have sought to uncover their family lineages via the interrogation of written historical records for millennia. Early examples of genealogy based on written historical records include the documentation of pharaonic dynasties in Egypt, the interrogation of epic poems in Greece and Biblical accounts of Christ's descent from Abraham. The field of genealogy was revolutionized within the last 25 years by the application of genetic, DNA marker-based methods to genealogical investigations (Fitzpatrick and Yeiser, 2005; Aulicino, 2013). DNA sequences have the potential to provide accurate, unbiased and sensitive markers for

the discernment of relationships among family members and for the assignment of individual ancestral origins. Genetic approaches to genealogy have been particularly attractive to communities of Afro-descendants in the Americas, who have often lacked access to the same level of detailed historical records that are available to other immigrant populations (Gates Jr., 2010).

To date, genetic genealogy has been dominated by studies of mitochondrial DNA (mtDNA) and Y-DNA sequences (haplotypes) (Cann et al., 1987; Stumpf and Goldstein, 2001; Jobling and Tyler-Smith, 2003; Pakendorf and Stoneking, 2005). Studies of mtDNA and Y-DNA haplotypes afford a number of advantages for genetic genealogy: they are sex-specific markers that allow for the distinct characterization of female (mtDNA) and male (Y-DNA) lineages, they do not recombine allowing for straightforward and tractable delineation of ancestral lineages and relationships among groups of lineages, and they show geographical differentiation providing for localization of ancient ancestral origins. The large databases of mtDNA and Y-DNA haplotypes that have accumulated over the years have provided for steadily increasing

---

resolution for ancestry assignment (Shriver and Kittles, 2004; Congiu et al., 2012).

Nevertheless, the use of these uniparental markers for genealogical studies also has important limitations. Since these markers capture single – female or male – unbroken ancestral lineages, they only represent a tiny fraction of any individual's genetic ancestry. Indeed, it has recently been shown that levels of continental ancestry, based on analysis of autosomal DNA sequences, can vary widely for individuals with the same mtDNA haplotype (Emery et al., 2015). Another unresolved issue with the use of such markers relates to the level of resolution that they afford for localized ancestry assignment. While they do show high levels of continental differentiation, uniparental markers can be broadly distributed across different sub-continental geographic regions. Thus, it may not be possible to unambiguously localize ancestral origins using such markers. This has been shown to be the case for African–Americans (Salas et al., 2005; Stefflova et al., 2011). Despite claims to be able to trace individual's ancestry to precise locations in Africa using mtDNA analysis, it was shown that mtDNA haplotypes in many cases can only be assigned to broad geographic regions in the continent (Salas et al., 2004).

Increasingly, historians and genealogists are recognizing the utility of a synthetic approach to the study of human ancestry that combines information gleaned from historical records with results based on the analysis of genetic markers. This combined evidence approach could provide for substantially increased resolution in ancestry localization for cases where genetic approaches only yield broad geographic assignments. Historical information could also be combined with genetic information at the population level to increase confidence in genetic-based ancestry assignments for individuals who lack access to reliable historical records. However, there currently exists no formal analytical framework for the integration of historical and genetic data in the study of genealogy. Here, we present a Bayesian analytical approach for the combination of population-level historical records with genetic marker data for the assignment of individual ancestry. We show that this combined evidence approach provides for substantially increased

resolution over a genetics-only approach and demonstrate the utility of tuning historical information to distinct genetic profiles.

We evaluated the potential of our combined evidence Bayesian framework for the study of African ancestry in the Colombian population. Colombia has an ethnically diverse population with high levels of admixture between African, European and Native American ancestral populations (Bryc et al., 2010; CIA, 2014). There is a large population of ~5 million Afro-descendants in Colombia, making up >10% of the total population. Afro-Colombians include individuals who self-identify as Black (African), Mulatto (Black/African and European) and Zambo (Black/African & Amerindian). Despite a number of studies on the genetic ancestry of Colombians (Carvajal-Carmona et al., 2000, 2003; Bedoya et al., 2006; Wang et al., 2008; Cordoba et al., 2012), there have been few such studies on the Afro-Colombian population. The Bayesian approach applied here combines historical records of trans-Atlantic slave voyages with genetic data on the geographic distribution of mtDNA haplotypes in Africa to provide for increased resolution of ancestry inference in this understudied population.

2. Material and methods

2.1. Historical and molecular anthropological datasets

Historical data on the African ancestral origins of the modern Afro-Colombian population, compiled from records of trans-Atlantic slave voyages, were taken from the literature (Maya Restrepo, 2005; Rodriguez, 2008). The numbers of individuals from the three main regions where Afro-Colombians were found to have originated were recorded: West Africa (W) $n = 6000$, West Central Africa (WC) $n = 340,000$ and South West Africa (SW) $n = 200,000$ (Fig. 1). The modern Afro-Colombian population was assumed to be made up of individuals with ancestries equal to the relative proportions of individuals from these three ancestral regions. This assumes that the regional origin proportions of Afro-Colombians have not changed substantially over time,
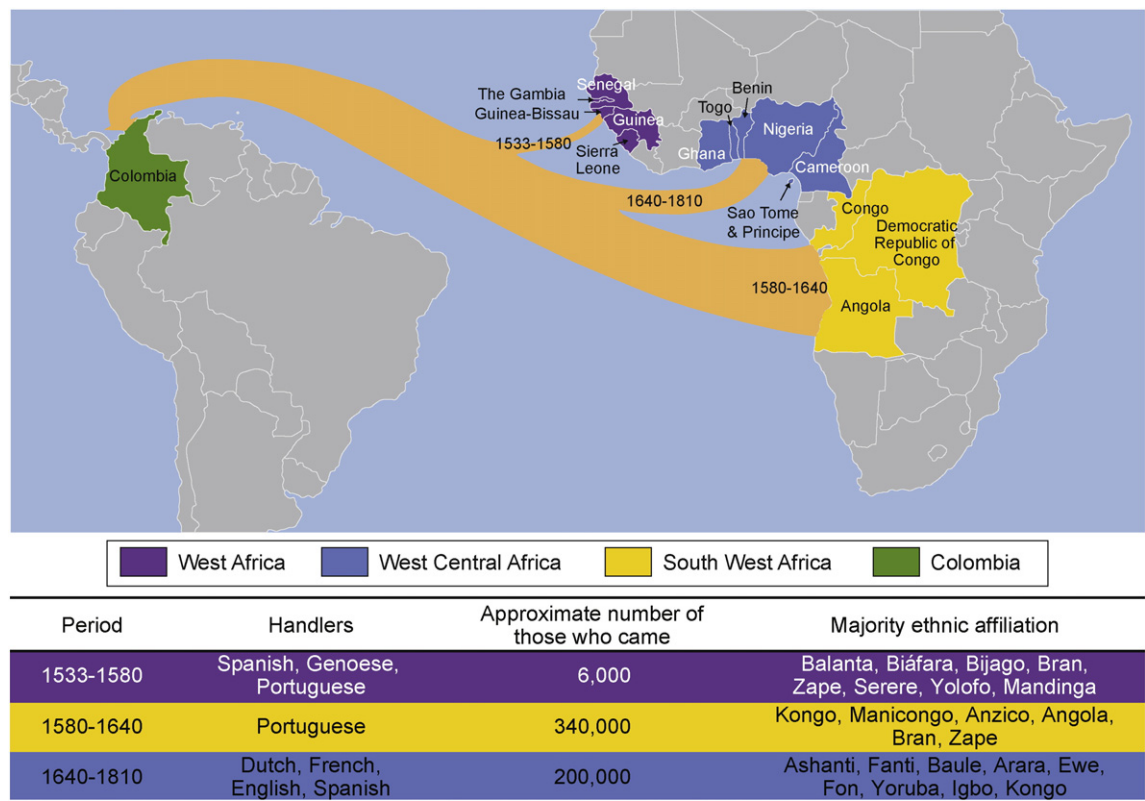

Period	Handlers	Approximate number of those who came	Majority ethnic affiliation
1533-1580	Spanish, Genoese, Portuguese	6,000	Balanta, Biáfara, Bijago, Bran, Zape, Serere, Yolofo, Mandinga
1580-1640	Portuguese	340,000	Kongo, Manicongo, Anzico, Angola, Bran, Zape
1640-1810	Dutch, French, English, Spanish	200,000	Ashanti, Fanti, Baule, Arara, Ewe, Fon, Yoruba, Igbo, Kongo

Fig. 1. African ancestral origins of the Afro-Colombian population. The three primary geographical regions from which Afro-Colombians were taken are shown along with time periods and demographic information.