Gene 574 (2015) 345–351

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

ARTICLE INFO

Article history:
Received 31 July 2015
Accepted 10 August 2015
Available online 11 August 2015

Keywords:
Human ancestry
Mitochondrial DNA
Haplotype
Afro-Colombian
Africa
Trans-Atlantic slave voyages
Bayes' rule
Combined evidence

ABSTRACT

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.

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

Abbreviations: mtDNA, mitochondrial DNA; Y-DNA, Y chromosomal DNA.

* Corresponding author at: Engineered Biosystems Building, Georgia Institute of Technology, 950 Atlantic Drive, Atlanta, GA 30332, USA.
E-mail address: King.jordan@biology.gatech.edu (I.K. Jordan).

http://dx.doi.org/10.1016/j.gene.2015.08.015
0378-1119/

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).

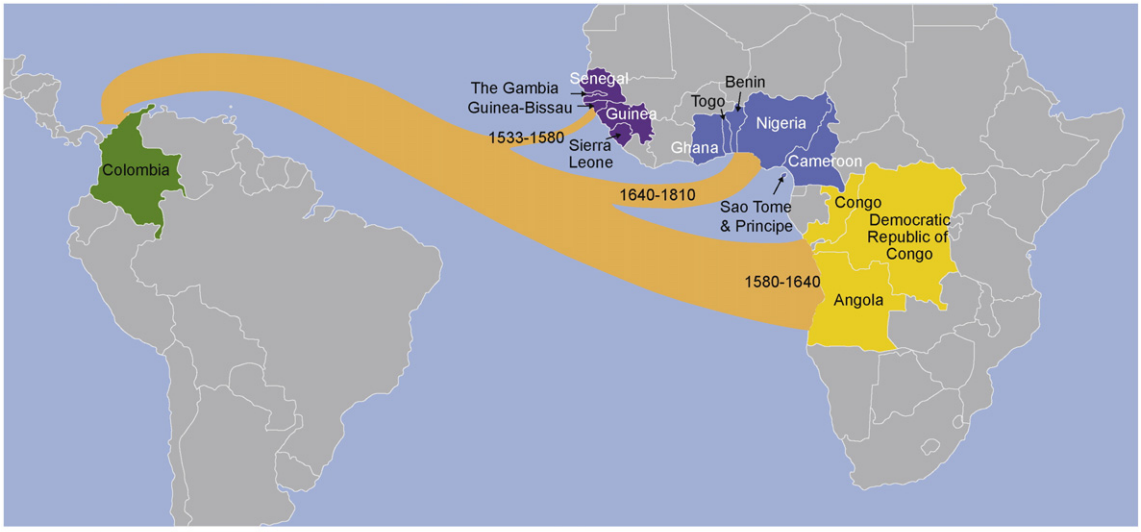
Increasingly, historians and genealogists are recognizing the utility of a synthetic approach to the study of human ancestry that combines information gleaned from historical records with results based on the analysis of genetic markers. This combined evidence approach could provide for substantially 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.

which is consistent with findings that Africans in the Americas do not show evidence of assortative mating with respect to tribal or regional origins (Zakharia et al., 2009). The method also assumes relatively low levels of migration within the African regions since the time of the trans-Atlantic slave trade.

Regional distributions of mtDNA haplotypes for countries corresponding to the three Afro-Colombian ancestral regions were taken from a recently compiled database of African mtDNA haplotypes (Stefflova et al., 2011). For this data set, results from 39 different studies were combined to yield a total of 128 mtDNA haplotypes from 34 countries characterized from 13,783 individuals. Individual mtDNA haplotypes were drawn from this data set in order to simulate modern Afro-Colombian populations as described below. A dataset of all the individual mtDNA haplotypes used for this study, along with their regional origins and additional demographic information, is provided as Supplementary Table 1.

2.2. Simulation of in silico Afro-Colombian populations

Monte Carlo simulation was used to create *in silico* Afro-Colombian populations by randomly drawing mtDNA haplotypes, from the dataset described in Section 2.1 (Supplementary Table 1), in proportion to the relative frequencies of the three African ancestral populations $P(Region)$: $P(W) = 0.011$, $P(WC) = 0.366$, $P(SW) = 0.623$. Each randomly simulated population consisted of 1000 individuals, and 1000 randomized populations were created for the subsequent ancestry inference and evaluation steps.

2.3. Assessment of posterior probabilities for Afro-Colombian ancestry assignment

Bayes' rule was applied to combine evidence from (i) historical records and (ii) genetic data in order to assign the most likely African ancestral origins for Afro-Colombian individuals using the posterior probability. For any given Afro-Colombian individual who has their mtDNA haplotype characterized, their African ancestral origin can be assigned as the posterior probability of coming from one of the three ancestral regions given that particular haplotype: $P(Region|Haplotype)$. Bayes' rule can be used to infer this posterior probability based on the relative proportion of mtDNA haplotypes of that kind coming from the same region (i.e., the conditional probability): $P(Haplotype|Region)$. To combine the historical data using the Bayes' rule approach, the relative contributions of African ancestral regions to the modern Afro-Colombian population are considered as prior probabilities: $P(Region)$. $P(Region)$ is defined as the relative fraction of Afro-Colombians from any given region based on historical records (see 2.2). Finally, the overall probability of seeing the mtDNA haplotype in the entire data set, $P(Haplotype)$, is used to normalize the Bayesian ancestry inference. $P(Haplotype)$ is defined as the overall fraction of any given haplotype in the entire dataset. Thus, for any given Afro-Colombian individual with a mtDNA haplotype, and for any of the three ancestral regions, the posterior probability of the ancestry assignment can be found as:

$$P(Region|Haplotype) = \frac{P(Haplotype|Region)P(Region)}{P(Haplotype)}.$$

An illustration of this approach for an example of a single mtDNA haplotype ($L2c$) and one ancestral region (W) is shown here:

$$P(W|L2c) = \frac{P(L2c|W)P(W)}{P(L2c)} = \frac{P(L2c|W)P(W)}{\sum P(L2c|Region)P(Region)}$$

$$= \frac{P(L2c|W)P(W)}{P(L2c|W)P(W) + P(L2c|WC)P(WC) + P(L2c|SW)P(SW)}$$

$$= \frac{0.179 * 0.011}{0.179 * 0.011 + 0.024 * 0.366 + 0.009 * 0.623} = 0.113.$$

2.4. Evaluation of the combined evidence Bayesian ancestry assignments

The (i) relative accuracy (A) and (ii) error (E) levels of the Bayesian ancestry assignment method for Afro-Colombian individuals were calculated by comparing individuals' predicted ancestries, based on the described application of Bayes' rule, and their known ancestries based on the simulated populations (described in Section 2.2).

For each simulated Afro-Colombian population of 1000 individuals, the relative accuracy (A) of the ancestry assignment is calculated as the fraction of individuals with correctly predicted ancestries (C), with a scaling factor applied such that random accuracy is equal to 50% (as opposed to 33% expected when inferring ancestry across three regions). The fraction of individuals with correctly predicted ancestries (C) is taken as the fraction of simulated individuals whose highest posterior probability ancestry assignment corresponds to the population from which they were simulated. Accuracy (A) = $C/1000$, and the relative accuracy $= \frac{A \times 0.5}{33}$ if $A \leq 33$ or $\frac{(A-33) \times 0.5}{67} + 0.5$ if $A \geq 33$. The error (E) of the ancestry inference for each simulated Afro-Colombian population is computed as the sum of the root mean squared difference between individuals predicted ancestries and their known ancestries across the three regions:

$$E = \sum_{i=1}^{1000} \left(\sqrt{\left(\sum \left(P(Region)_i^{Bayes} - P(Region)_i^{Known} \right)^2 \right) / 3} \right)$$

where $P(Region)_i^{Bayes}$ is the posterior probability that the ancestry of individual i corresponds to that region and $P(Region)_i^{Known}$ is the frequency individual i haplotype in that region.

For any given set of historical prior probabilities, the process of Bayesian ancestry inference followed by calculation of the relative accuracy (A) and error (E) was repeated for the 1000 randomly simulated populations to compute accuracy (A) and error (E) distributions. This process was repeated across a grid search space covering all of the possible historical prior probabilities in a stepwise manner; at each prior probability set in the grid, the posterior probability ancestry assignment values were re-calculated followed by accuracy (A) and error (E) calculations (Fig. 2). Accuracy (A) and error (E) levels were calculated for each population, and the distributions of these values over 1000 populations were compared between different prior spaces using the Student's t-test.

2.5. Sensitivity of mtDNA haplotypes to changes in historical probability priors

The effect of changing historical prior probability sets on haplotype-based ancestry inference, as measured by the posterior probability $P(Region|Haplotype)$, was computed via Manhattan distances between posterior probability vectors for the three regions:

$$D_{Hap} = \sum_{Regions} \left| P(Region|Hap)_{prior\ set1} - P(Region|Hap)_{prior\ set2} \right|.$$

The relative breadth, or conversely the regional-specificity, of mtDNA haplotype distributions across the three African regions evaluated here was measured using Shannon's Entropy (H), where:

$$H_i = -\sum_{j \in all\ Regions} P(Hap_i|Region_j) \times \log_2 \left(P(Hap_i|Region_j) \right).$$

The use of Shannon's entropy in this way allowed us to evaluate the relationship of the evenness of historical prior probabilities with respect to accuracy of ancestry assignment.

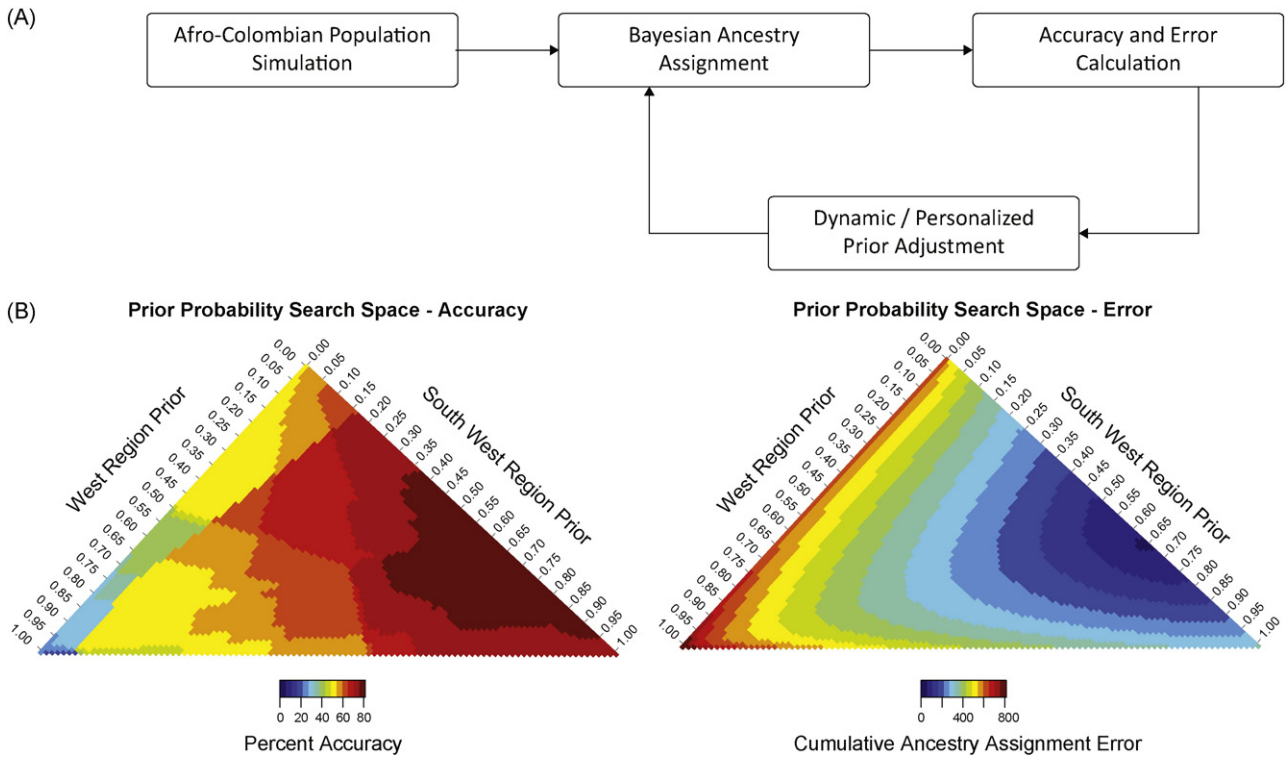


Fig. 2. Simulation and ancestry assignment for AfroColombians. (A) A schematic of the population simulation and Bayesian ancestry assignment approach used here. (B) Results of ancestry assignment accuracy and error across historical prior probability space. Note that while there are historical prior probability values for three geographical regions, the value of the third prior is dependent on the first two priors.

3. Results

3.1. Combined evidence ancestry assignment with Bayes' rule

We proposed an inverse probability approach, based on Bayes' rule, in order to evaluate the utility of combining (i) historical records with (ii) genetic data for making human ancestry inferences. Specifically, a method for assigning the sub-continental African ancestral origins of Afro-Colombians was developed by combining historical records of trans-Atlantic slave voyages (Maya Restrepo, 2005; Rodriguez, 2008) with African mtDNA haplotype distributions (Stefflova et al., 2011). In this approach, the historical records are taken as prior probabilities indicating the probability that an Afro-Colombian comes from one of three possible ancestral regions – West Africa (*W*), West Central Africa (*WC*) or South West Africa (*SW*) – in the absence of any genetic information (Fig. 1). Individuals' mtDNA haplotypes can then be compared to the distribution of mtDNA haplotypes across the three ancestral regions in order to provide additional resolution for ancestry inference. The details of this combined evidence Bayesian method for Afro-Colombian ancestry inference are spelled out in the Material and Methods Section 2.3.

This combined evidence approach to ancestry inference was evaluated by simulating Afro-Colombian populations, with individuals represented by specific mtDNA haplotypes, based on the relative frequencies of individuals expected to have originated from each ancestral region (see Material and Methods Section 2.2). This simulation process yielded randomized sets of individual mtDNA haplotypes with known ancestral origins. The Bayesian method for ancestry assignment was then applied to these simulated mtDNA haplotypes, using different sets of historical prior probabilities, and the ancestry assignments based on the Bayesian approach were compared to the true ancestries based on the simulation (Fig. 2).

3.2. Increased resolution of combined evidence for ancestry assignment

The utility of the Bayesian combined evidence approach to ancestry inference developed here was evaluated by calculating the accuracy and error of predicted ancestries compared to known ancestries based on the simulated Afro-Colombian populations (see Material and Methods Section 2.4). This three-step approach – 1) simulation, 2) ancestry assignment and 3) evaluation (Fig. 2A) – was iterated over the entire space of possible ancestry prior probabilities. The distributions of accuracy and error across prior space are shown in Fig. 2B. The null expectation for Bayesian ancestry assignment was validated by randomly assigning sub-continental ancestries to Afro-Colombian individuals and calculating their accuracy. As expected, random ancestry assignments for simulated populations have relative accuracy values that fluctuate around 50% and high error levels (Fig. 3A, C and D).

Next, we wished to compare the utility of adding historical information to genetic data for making ancestry inferences. To do this, we compared the results of using so-called flat prior probabilities, whereby the prior probabilities of an Afro-Colombian individual coming from any of the three ancestral regions is equal. This is equivalent to not using any historical information in ancestry inference, i.e., relying on genetic information alone. With this approach, the average ancestry assignment accuracy value is 63.2% and the average error level is 203.2 (Fig. 3). While these accuracy and error values are substantially higher than seen for the random ancestry assignment, they are nevertheless lower than what one may desire for a confident sub-continental ancestry inference. This is consistent with previous results calling attention to the fact that broad continental distributions of mtDNA haplotypes make precise sub-continental ancestry inferences based on these data alone highly problematic (Salas et al., 2004, 2005).

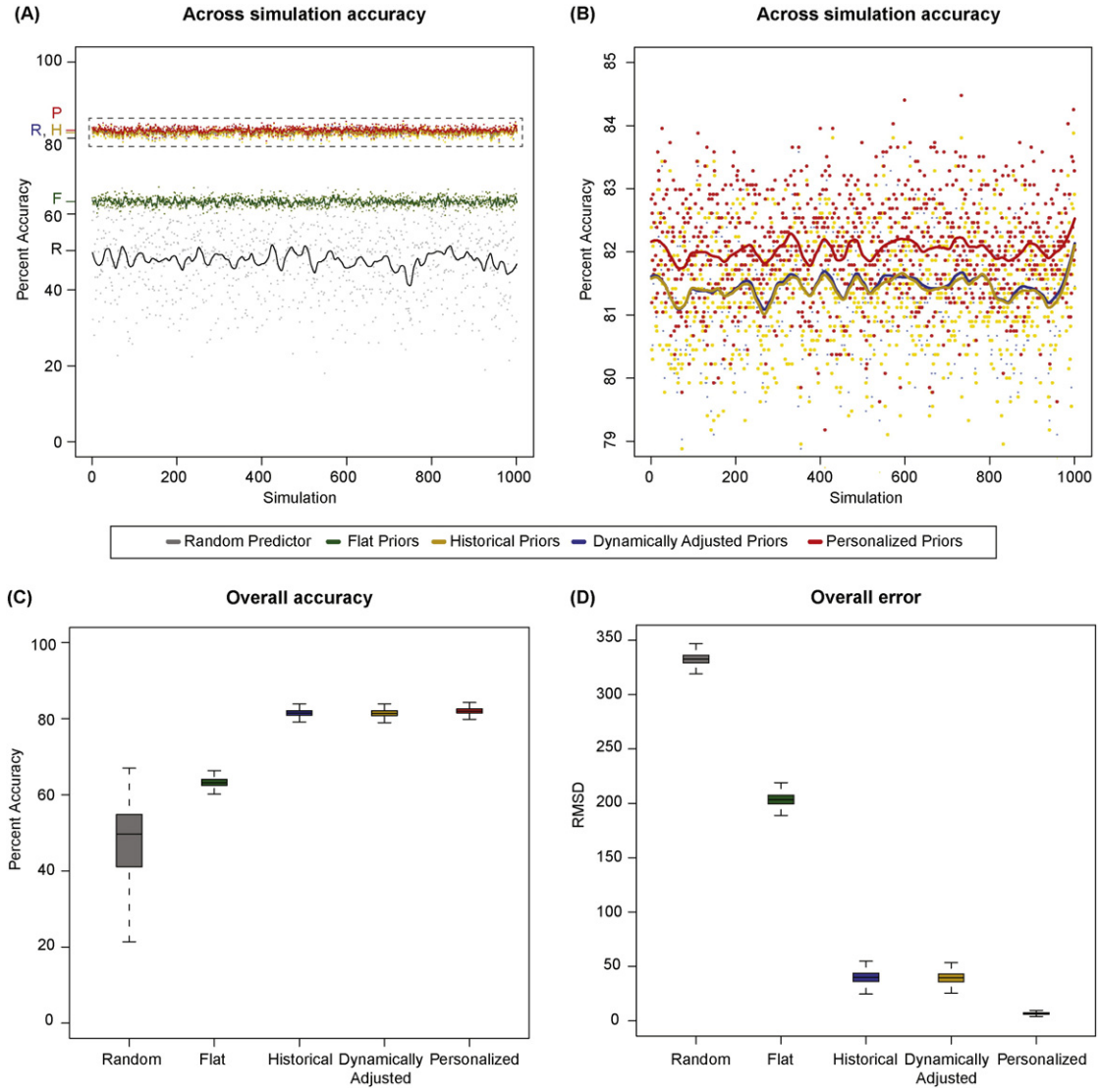


Fig. 3. Accuracy and error of ancestry assignment for different approaches to historical prior probability selection. (A) Accuracy levels are compared for different historical prior probability sets across 1000 simulations. (B) Expansion of the accuracy range from panel A showing results for the historical, dynamic and personalized prior probability sets. (C) Accuracy and (D) error value distributions for the different prior probability sets.

Historical records were incorporated into the Bayesian ancestry inference in the form of prior probabilities based on the relative frequencies of Africans recorded to have been taken from those regions to Colombia (Fig. 1). Incorporation of historical priors results in a substantial increase in the average accuracy to 81.4% and a marked decrease in the average error level to 40.2 (Fig. 3). The differences between flat priors, with no historical information included, and historical priors are highly statistically significant (Student's t-test accuracy $t = 3.9 \times 10^2$, $p \approx 0$; error $t = 6.5 \times 10^2$, $p \approx 0$). These results underscore the utility of combining historical and genetic information for ancestry inference.

3.3. Effect of the prior probability space on ancestry assignment

Historical records of trans-Atlantic slave voyages to Colombia point to three eras of forced migration, each of which corresponded to a distinct African region, different colonial perpetrators and markedly divergent numbers of transported individuals (Fig. 1). This process yielded a highly asymmetrical distribution of African sub-continental origins for Afro-Colombians, with a particular under-representation of individuals from West Africa. Accordingly, the historical prior probabilities for the three regions are highly skewed, and this could result in diminishing useful ancestry information provided by mtDNA haplotypes.

To address this possibility, we dynamically adjusted the historical prior probabilities in order to search for the best possible overall prior combination (Fig. 2B). This was done by using a grid-search to calculate the posterior probabilities, i.e., the ancestry assignments, over the entire range of possible prior probabilities. The objective criterion for this grid-search in prior probability space was the lowest possible error rate for ancestry assignment.

We expected that there may be a less skewed overall prior probability space that results in more reliable ancestry inference by virtue of giving additional weight to mtDNA haplotypes that are more regionally-specific. This process did significantly reduce the average error level (Student's $t = 6.5$, $p = 1.0 \times 10^{-10}$), but the difference between the error levels for the historical and dynamic priors is quite small (4% difference, Fig. 3D). In addition, the accuracy level distributions for historical versus dynamic priors are not significantly different (Student's $t = 0.8$, $p = 0.44$). The optimal set of dynamic priors does show a 3.6 fold increase in the prior probability for West Africa, consistent with the idea that the highly skewed historical prior distribution diminishes the ancestry information encoded in regional-specific mtDNA haplotypes. However, the small overall difference in the effect of dynamically

adjusting the historical prior probabilities suggests the possibility that this is only the case for some of the mtDNA haplotypes.

3.4. Personal prior probabilities for ancestry assignment

The utility of combining historical records with genetic information in making ancestry inferences, particularly for inferring African subcontinental ancestry, was initially suggested by results showing that mtDNA haplotypes are broadly distributed across the continent (Salas et al., 2004, 2005). However, the fact that dynamically adjusting the prior probabilities results in a flattening of the prior space, i.e., eliminating some of the skew in the prior probabilities seen based on the historical records, suggests the possibility that more regional-specific mtDNA haplotypes do in fact contain useful ancestry information. Nevertheless, the entire set of mtDNA haplotypes represented in any given Afro-Colombian population will likely differ substantially with respect to the breadth of their distribution across African regions. Thus, it may be the case that using a single set of historical prior probabilities for ancestry inference with mtDNA haplotypes that show widely different degrees of regional-specificity will not yield the best results. Instead, it may be preferable to infer ancestry using optimal prior probability sets that are individually determined for each distinct mtDNA haplotype.

To evaluate this possibility, we dynamically adjusted the historical probabilities individually for each mtDNA haplotype and chose the prior probability set that yielded the lowest error rate for that particular haplotype, i.e., we generated a set of mtDNA haplotype-specific 'personalized priors'. The ancestry inference results for all mtDNA haplotypes were then combined when evaluating the accuracy and error levels for the entire set of simulated populations. Calculating mtDNA-specific prior probability sets in this way, to yield personalized priors, results in the highest accuracy levels and lowest error levels of any of the methods. The improvement in accuracy is highly significant (Student's $t = 15.9$, $P = 7.7 \times 10^{-54}$), albeit quite marginal (0.74% difference, Fig. 3C), compared to the dynamic prior probabilities, whereas the reduction in the error level is both highly significant and far more substantial (96% difference, Student's $t = 2.2 \times 10^{2}$, $P \approx 0$, Fig. 3D).

The source of this improved performance, and conversely the marginal effect of dynamically adjusting the historical prior probabilities (see Section 3.3), can be traced to the differences in how individual mtDNA haplotypes respond to changes in priors. Individual haplotypes vary greatly with respect to the extent to which changing the prior probabilities affects ancestry inference. Some mtDNA haplotypes show low levels of change, or virtually no change, in ancestry assignment for different prior probabilities, whereas others show higher levels of change (Fig. 4A). There are three discrete clusters with respect to the extent to which changing the historical prior probabilities influences ancestry inference (Fig. 4B). Each of these clusters of mtDNA haplotypes has distinct levels of haplotype likelihood variation ($P(Haplotype|Region)$) across the three geographical regions, as represented by entropy in Fig. 4C. Low entropy mtDNA haplotypes have skewed distributions, in terms of their relative frequencies across geographic regions, and conversely high entropy haplotypes are more evenly distributed across regions. Ancestry assignment for unevenly distributed mtDNA haplotypes is more determined by the genetic data, whereas ancestry for more evenly distributed haplotypes is more dependent on the historical prior probabilities. These differences in haplotype behavior mean that a single set of prior probabilities, even if optimally assigned, cannot accurately capture ancestry assignment as well as a set of personal prior probabilities individually tailored to distinct mtDNA haplotypes.

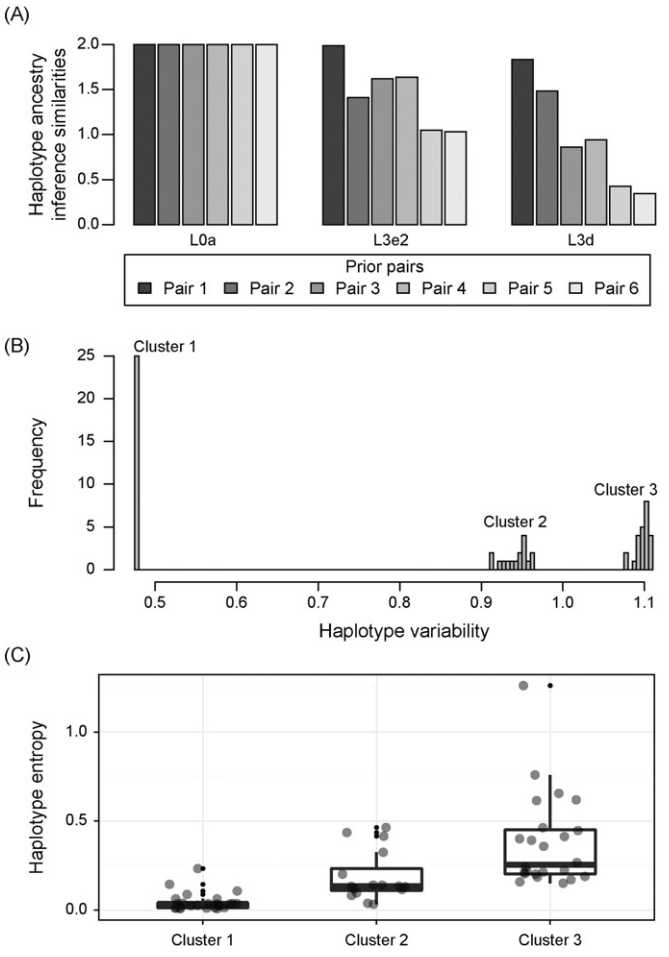


Fig. 4. Changes in mtDNA ancestry assignment across different historical prior probability sets. (A) Examples of the similarity of ancestry assignment between different pairs of prior probabilities is shown for three mtDNA haplotypes. (B) The extent to which mtDNA haplotype ancestry assignments vary across different sets of geographical priors. (C) Entropy levels, based on likelihoods of mtDNA haplotypes across the three geographical regions, for groups mtDNA haplotypes from the three clusters in panel B.

4. Discussion

The ability to pinpoint the ancestral geographical origins of a family lineage is fundamental to genetic genealogy. Accurate assignment of ancestral origins is dependent on the differential geographic distribution of genetic markers. While there are numerous genetic markers (haplotypes) that are clearly differentiated between continents, most ancestral lineage markers, including widely employed uniparental markers such as mtDNA, are widely distributed across different geographical regions within continents. This means that the genetic characterization of any such marker can only be used to assign ancestry across a broad geographic region. Despite this limitation, lineage-based tests of ancestry continue to be widely used for ancestry assignment.

The combination of historical evidence, based on written records, with genetic marker based data should be able to provide increased resolution for geographical ancestry assignment. Indeed, companies that provide genealogical services have begun to recognize this fact and provide the ability for users to integrate historical records with genetic data. However, to date there has not been any formal analytical framework for the integration of historical records with information gleaned from the characterization of genetic markers. Here, we provide such a framework in the form a straightforward Bayesian calculation for the combination of information taken from historical records with data on the geographical distribution of mtDNA haplotypes. Bayesian approaches have been shown to be quite useful for genetic marker based ancestry inference (Corander et al., 2004; Raj et al., 2014), but they have not been used to combine historical and genetic information as we have done here. We show that the incorporation of historical information using Bayes' rule, i.e., the calculation of posterior probabilities, can substantially increase the accuracy of ancestry assignment based on mtNDA haplotypes.

Our combined evidence Bayesian approach to ancestry assignment was applied to the question of the African geographic origins of Afro-Colombians. Despite the fact that there is a large population of Afro-descendants in Colombia, this particular community has been relatively understudied, in terms of their genetic ancestry, compared to nearby mestizo populations that have primarily European and Native American ancestry. Thus, the application of the combined evidence approach to ancestry assignment in this community has the potential to be impactful. In addition, descendants of Africans throughout the Americas have been particularly interested taking advantage of genetic approaches to genealogy to gain insight into their ancestral origins, which were often obscured by the harsh conditions of the slave trade. Our approach should prove to be relevant and applicable to other communities of Afro-descendants in the Americas. In fact, there is a well-established effort underway to record and analyze records of more than 35,000 trans-Atlantic voyages that carried more than 12 million Africans destined for the new world (http://www.slavevoyages.org/). Data from this detailed resource could be used as historical priors for ancestry assignment of African–Americans, Afro-Caribbeans and Afro-Brazilians.

Our results also underscore the utility of taking a personalized approach to ancestry inference. The individualized assignment of historical prior probabilities to different mtDNA haplotypes can often yield major changes in ancestry inference and leads to an overall improvement in ancestry assignment. The interrogation of more family-specific historical records could provide a way to formulate historical prior values that are even more individually tuned to personal family histories.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2015.08.015.

Acknowledgments

This work was funded by BIOS Centro de Bioinformática y Biología Computacional (062-2014) and the Georgia Institute of Technology Denning Global Engagement Seed Fund (320000118).

References

Aulicino, E.D., 2013. Genetic Genalogy: The Basics and Beyond. AuthorHouse, Bloomington.
Bedoya, G., Montoya, P., Garcia, J., Soto, I., Bourgeois, S., Carvajal, L., Labuda, D., Alvarez, V., Ospina, J., Hedrick, P.W., Ruiz-Linares, A., 2006. Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate. Proc. Natl. Acad. Sci. U. S. A. 103, 7234–7239.
Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., Ostrer, H., 2010. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc. Natl. Acad. Sci. U. S. A. 107 (Suppl. 2), 8954–8961.
Cann, R.L., Stoneking, M., Wilson, A.C., 1987. Mitochondrial DNA and human evolution. Nature 325, 31–36.
Carvajal-Carmona, L.G., Soto, I.D., Pineda, N., Ortiz-Barrientos, D., Duque, C., Ospina-Duque, J., McCarthy, M., Montoya, P., Alvarez, V.M., Bedoya, G., Ruiz-Linares, A., 2000. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. Am. J. Hum. Genet. 67, 1287–1295.
Carvajal-Carmona, L.G., Ophoff, R., Service, S., Hartiala, J., Molina, J., Leon, P., Ospina, J., Bedoya, G., Freimer, N., Ruiz-Linares, A., 2003. Genetic demography of Antioquia (Colombia) and the Central Valley of Costa Rica. Hum. Genet. 112, 534–541.
CIA, 2014. The World Factbook. US Government, Langley.
Congiu, A., Anagnostou, P., Milia, N., Capocasa, M., Montinaro, F., Destro Bisol, G., 2012. Online databases for mtDNA and Y chromosome polymorphisms in human populations. J. Anthropol. Sci. 90, 201–215.
Corander, J., Waldmann, P., Marttinen, P., Sillanpaa, M.J., 2004. BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics 20, 2363–2369.
Cordoba, L., Jair Garcia, J., Hoyos, L.S., Duque, C., Rojas, W., Caravajal, S., Escobar, L.F., Reyes, I., Cajas, N., Sanchez, A., Garcia, F., Bedoya, G., Ruiz-Linares, A., 2012. Composicion genetica de una poblacion del suroccidente de Colombia. Revista Colombiana de Antropologia 48, 21–48.
Emery, L.S., Magnaye, K.M., Bigham, A.W., Akey, J.M., Bamshad, M.J., 2015. Estimates of continental ancestry vary widely among individuals with the same mtDNA haplogroup. Am. J. Hum. Genet. 96, 183–193.
Fitzpatrick, C., Yeiser, A., 2005. DNA & Genealogy. Rice Book Press, Houston.
Gates Jr., H.L., 2010. Exploring Our Roots. PBS, Boston.
Jobling, M.A., Tyler-Smith, C., 2003. The human Y chromosome: an evolutionary marker comes of age. Nat. Rev. Genet. 4, 598–612.
Maya Restrepo, L.A., 2005. Brujería y reconstrucción de identidades entre los africanos y sus descendientes en la Nueva Granada Siglo XVII. Ministerio de Cultura, Bogotá.
Pakendorf, B., Stoneking, M., 2005. Mitochondrial DNA and human evolution. Annu. Rev. Genomics Hum. Genet. 6, 165–183.
Potter-Phillips, D., 1999. History of genealogy. Family Chronicle (http://www.familychronicle.com/HistoryOfGenealogy.html).
Raj, A., Stephens, M., Pritchard, J.K., 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197, 573–589.
Rodriguez, J.A., 2008. Bantúes y otros Africanos en Colombia, Velorios Y Santos Vivos. Comunidades Negras, Afrocolombianas, Raizales Y Palenqueras. Museo Nacional De Colombia, Bogotá.
Salas, A., Richards, M., Lareu, M.V., Scozzari, R., Coppa, A., Torroni, A., Macaulay, V., Carracedo, A., 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am. J. Hum. Genet. 74, 454–465.
Salas, A., Carracedo, A., Richards, M., Macaulay, V., 2005. Charting the ancestry of African Americans. Am. J. Hum. Genet. 77, 676–680.
Shriver, M.D., Kittles, R.A., 2004. Genetic ancestry and the search for personalized genetic histories. Nat. Rev. Genet. 5, 611–618.
Stefflova, K., Dulik, M.C., Barnholtz-Sloan, J.S., Pai, A.A., Walker, A.H., Rebbeck, T.R., 2011. Dissecting the within-Africa ancestry of populations of African descent in the Americas. PLoS One 6, e14495.
Stumpf, M.P., Goldstein, D.B., 2001. Genealogical and evolutionary inference with the human Y chromosome. Science 291, 1738–1742.
Wang, S., Ray, N., Rojas, W., Parra, M.V., Bedoya, G., Gallo, C., Poletti, G., Mazzotti, G., Hill, K., Hurtado, A.M., Camrena, B., Nicolini, H., Klitz, W., Barrantes, R., Molina, J.A., Freimer, N.B., Bortolini, M.C., Salzano, F.M., Petzl-Erler, M.L., Tsuneto, L.T., Dipierri, J.E., Alfaro, E.L., Bailliet, G., Bianchi, N.O., Llop, E., Rothhammer, F., Excoffier, L., Ruiz-Linares, A., 2008. Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet. 4, e1000037.
Zakharia, F., Basu, A., Absher, D., Assimes, T.L., Go, A.S., Hlatky, M.A., Iribarren, C., Knowles, J.W., Li, J., Narasimhan, B., Sidney, S., Southwick, A., Myers, R.M., Quertermous, T., Risch, N., Tang, H., 2009. Characterizing the admixed African ancestry of African Americans. Genome Biol. 10, R141.