Article-Discoveries

Dissecting the Pre-Columbian genomic ancestry of Native Americans along the Andes-Amazonia divide

Guido Alberto Gnecchi-Ruscone[*,1,¤], Stefania Sarno[*,†,1], Sara De Fanti[1], Laura Gianvincenzo[1], Cristina Giuliani[1], Alessio Boattini[1], Eugenio Bortolini[2], Tullia Di Corcia[3], Cesar Sanchez Mellado[4], Taylor Jesus Dàvila Francia[4], Davide Gentilini[5], Anna Maria Di Blasio[5], Patrizia Di Cosimo[6], Elisabetta Cilli[2], Antonio Gonzalez-Martin[7], Claudio Franceschi[8], Zelda Alice Franceschi[9], Olga Rickards[3], Marco Sazzini[†,1], Donata Luiselli[2], Davide Pettener[1]

[1]University of Bologna, Laboratory of Molecular Anthropology & Centre for Genome Biology, Dept. of Biological, Geological and Environmental Sciences, Bologna, Italy
[2]University of Bologna, Dept. of Cultural Heritage, Ravenna, Italy
[3]University of Rome Tor Vergata, Dept. of Biology, Rome, Italy
[4]National Intercultural University of Amazon, Faculty of Intercultural Education and Humanity, Ucayali, Peru
[5]Center for Biomedical Research & Technologies, Italian Auxologic Institute IRCCS, Milan, Italy
[6]Takesi Project, University of Bologna, Italy
[7]Complutense University of Madrid, Dept. of Zoology and Physical Anthropology, Madrid, Spain
[8]University of Bologna, Dept. of Experimental, Diagnostic and Specialty Medicine, Bologna, Italy
[9]University of Bologna, Dept. of History and Cultures, Bologna, Italy

[¤]Current address: Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany

[*] These authors contributed equally to this work
[†] Corresponding authors: E-mails: stefania.sarno2@unibo.it, marco.sazzini2@unibo.it

Abstract

Extensive European and African admixture coupled with loss of Amerindian lineages makes the reconstruction of pre-Columbian history of Native Americans based on present-day genomes extremely challenging. Still open questions remain about the dispersals that occurred throughout the continent after the initial peopling from the Beringia, especially concerning the number and dynamics of diffusions into South America. Indeed, if environmental and historical factors contributed to shape distinct gene pools in the Andes and Amazonia, the origins of this East-West genetic structure and the extension of further interactions between populations residing along this divide are still not well understood.

To this end, we generated new high-resolution genome-wide data for 229 individuals representative of one Central and 10 South Amerindian ethnic groups from Mexico, Peru, Bolivia and Argentina. Low levels of European and African admixture in the sampled individuals allowed the application of fine-scale haplotype-based methods and demographic modeling approaches. These analyses revealed highly specific Native American genetic ancestries and great intra-group homogeneity, along with limited traces of gene flow mainly from the Andes into Peruvian Amazonians. Substantial amount of genetic drift differentially experienced by the considered populations underlined distinct patterns of recent inbreeding or prolonged isolation. Overall, our results support the hypothesis that all non-Andean South Americans are compatible with descending from a common lineage, while we found low support for common Mesoamerican ancestors of both Andeans and other South American groups. These findings suggest extensive back-migrations into Central America from non-Andean sources or conceals distinct peopling events into the Southern Continent.

Introduction

The history of Native American populations is one of the most debated topics in the study of ancient human migrations, which puzzles academics from many different fields (Dillehay et al. 2009). Recently, new sources of evidence coming from genomic data of both modern populations and ancient human specimens have been contributing to unveil novel aspects on the genetic ancestry and population history of first Americans. Overall, it has been confirmed that present-day Native American groups descend from human expansions entering North America from East Asia through the Beringia land corridor, although subsequent timings, number of founder events and especially diffusion processes within the Americas are still a matter of intense debate (Skoglund and Reich 2016).

Ancient genomes from North America and Siberia revealed that present-day Northern Native American populations harbor an intricate mixture of four main streams of ancestry, which were brought into the continent during at least three different diffusion processes (Raghavan et al. 2014; Rasmussen et al. 2014; Rasmussen et al. 2015; Raghavan et al. 2015; Lindo et al. 2017; Moreno-Mayar et al. 2018a). While documenting a complex pattern of secondary migrations into North America, the first and oldest of these waves (i.e. First Americans, FA) was until recently supposed to be the one contributing to the ancestry of all present-day Central and South American groups (Schurr and Sherry 2004; Tamm et al. 2007; Wang et al. 2007; Kitchen et al. 2008; Fagundes et al. 2008; Perego et al. 2010; Reich et al. 2012; Battaglia et al. 2013). Consistently with this scenario, a 12.6

ka human sample recovered in western Montana (Anzick-1) was found to derive all of his ancestry from the same FA source and, in fact, resulted to be genetically closer to Native Central and South Americans than to any other Northern American group (Rasmussen et al. 2014). In addition, coalescent analyses of ancient mitochondrial genomes from South America further suggested a small population entering the Americas around 16 ka after a few millennia of Beringia standstill and rapidly expanding southward (Llamas et al. 2016). In agreement with archaeological records, and particularly with the presence in Southern Chile of one of the oldest American archeological sites (Monte Verde, 14-15 ka), these results pointed towards a strong founder effect and an early and rapid peopling from North to South America, plausibly along a costal Pacific route (Dillehay and Collins 1988; Dillehay et al. 2008).

However, other studies questioned the model of a single wave of genetically homogeneous migrants as being responsible for the entire ancestry of Central and South American populations (Skoglund et al. 2015; Brandini et al. 2017; Scheib et al. 2018). Accordingly, recent evidence based on genomic data generated from ancient human remains retrieved in Central and South America confirmed the occurrence of multiple waves of diffusion into the south of the continent and suggested a complex scenario involving the spread of ancient populations that were already genetically structured (Moreno-Mayar et al. 2018b; Posth et al. 2018). Among these samples, the oldest ones (dating to ~11-10 ka), either from North America (Nevada) or South America (West and East of the Andes), were those showing the highest genetic affinity with Anzick-1. This corroborates the hypothesis that the first diffusion from North into South America was extremely rapid (~1-2 ka) and was not limited to the West coast since it is supported by samples from the entire continent. However, the genetic footprints of this first peopling event were found to be subtler in more recent samples, suggesting an extensive population replacement beginning from around 9 ka, by a different ancestral lineage with respect to that represented in the Clovis-associated Anzick-1 (Posth et al. 2018). Subsequent migrations after the initial diffusion were associated with expansion from Mesoamerica occurred sometime after ~8.7 ka, which spread first southward (contributing to the ancestry of all present-day South Americans) and then northward, as suggested by ~2 ka ancient samples from Nevada (Moreno-Mayar et al. 2018b). Despite the common view that these processes contributed significantly to the formation of the modern South American genomic landscape, the two above mentioned studies did not clarify in detail the relative proportions of these ancestries in the genomes of contemporary populations, being instead focused mostly on the relationship between ancient samples. Finally, minor contributions to the South American gene pool (i.e. < 5% of ancestry) were ascribable to the affinity with an Austro-Melanesian-related ancestry source already attested for some present-day Amazonian groups (Skoglund et al. 2015). This pattern was recognized also in a ~10 ka sample from Brazil (Moreno-Mayar et al. 2018b), coupled with a newly described connection between ancient samples from the California Channel Islands and the Late Central Andes since around 4.2 ka (Posth et al. 2018).

Overall, these studies revealed that the dynamics of demographic events occurred between Central and South America subsequently to the initial peopling of these regions, as well as within the southern continent itself, have been more complicated than previously thought. Within South America, mitochondrial DNA, Y-chromosome and autosomal data showed a clear structure East/West of the Andes, in agreement with a long-

standing geographic barrier between Andeans and Amazonians (Luiselli et al. 2000; Tarazona-Santos et al. 2001; Fuselli et al. 2003; Reich et al. 2012; Homburger et al. 2015). In addition, populations from the Andean cordillera experienced an additional history of adaptation to high-altitude environments with respect to the other South Americans (Bigham et al. 2010; Crawford et al. 2017; Lindo et al. 2018b). More recently, the Andes were the cradle of the major South American Pre-Columbian civilizations, last of which the Inca Empire (D'Altroy 2014). In the same way, complex demographic histories and different patterns of gene flow among and within Central and South America may have further affected the genetic structuring of Southern Native Americans during and after the initial peopling process.

In conclusion, the dynamics that characterized the entering and diffusion of Native American ancestors in South America are still unclear and many questions remain open. More specifically, there is no clear evidence about 1) how the different diffusions into South America described by Moreno-Mayar et al. (2018b) and Posth et al. (2018) reconcile with the genetic structure observable in present-day South Americans East and West of the Andes, and 2) the extent to which subsequent contacts and gene flow between Central and South Americans, as well as between Andeans and Amazonians, occurred. Indeed, while local patterns of back migrations and gene flow between the Caribbean region and northern South America has been detected (Reich et al. 2012; Moreno-Estrada et al. 2013; Schroeder et al. 2017), it is still not clear if and to what extent the Andean and non-Andean gene pools have admixed after their initial split, with inevitable implications for the identification of a correct divergence time between the two groups. Hints from uniparental markers suggested that some gene-flow between the Andes and Amazon could have occurred (Barbieri et al. 2014; Di Corcia et al. 2017; Gómez-Carballa et al. 2018). Furthermore, they also show different patterns of genetic drift and gene flow, with larger effective population sizes and higher migration rates within the Andes, compared to lower gene flow and higher genetic drift in the eastern populations settled in Amazonian and Chaco regions (Tarazona-Santos et al. 2001; Lewis et al. 2005; Sevini et al. 2013).

From a genome-wide perspective, a strong limitation in the study of Native American population history is due to the dramatic demographic changes that they experienced after the European colonization of the 15th century (Llamas et al. 2016; Lindo et al. 2016; Lindo et al. 2018a). In fact, it is well known that, because of these processes, present-day American populations appear as a mixture of ancestral sources from different continents, mostly Europe and Africa, with varying proportions from one country to another. This makes extremely limited the possibility of making historical inferences and testing demographic models based only on the fractioned Native American genomic portions (Gravel et al. 2013; Moreno-Estrada et al. 2014; Homburger et al. 2015; Kehdy et al. 2015; Montinaro et al. 2015).

In the present study, we aimed at investigating some aspects of the peopling processes of South America both 1) at a continent-wide scale in relation to Central American populations and 2) at a more local scale as concerns the interactions between the high-altitude Andeans and the neighboring populations from Peruvian Amazon and Argentinian Gran Chaco regions. To this end, we analyzed 229 individuals representative of 11 Central and South Native American ethnic groups whose DNA was genotyped for ~720,000 genome-wide single nucleotide polymorphisms (SNPs). Taking advantage from samples previously typed for uniparental markers,

we generated new genomic data for five ethnic groups from Peruvian Amazon (Barbieri et al. 2014; Di Corcia et al. 2017), one group from the Gran Chaco region (Sevini et al. 2013) and four high-altitude Andean ethno-linguistic groups from the Titicaca lake area in Peru (Barbieri et al. 2011), as well as for newly collected samples from the Bolivian Andes. In addition, we included one Mexican ethnic group (Tzotzil) as representative of the "Mayan Cluster" identified by Moreno-Estrada et al. (2014), which was missing in previous Native American reference datasets that we included in our study (Li et al. 2008; Reich et al. 2012). By applying fine-scale haplotype-based analyses and demographic modelling inferences, we provided new insights into the origins of ancestral gene pools East-West of the Andes-Amazonia divide, as well as on local patterns of isolation and admixture that differently shaped the genetic and cultural complexity of present-day South American populations.

Results

After the quality control (QC) steps detailed in Materials and Methods, we obtained an "extended" dataset consisting of 207,165 genome-wide SNPs typed in 178 newly analyzed samples from Mesoamerican (Meso) and South American (SA) populations, 431 individuals from 50 additional Amerindian groups already included in previous reference studies, and 92 non-Native American populations retrieved from the literature (**supplementary table S1, Supplementary Material online**; Li et al. 2008, Reich et al. 2012, The 1000 Genomes Project Consortium). We used this "extended" dataset to frame the genetic variation of analyzed populations into the context of worldwide genomic landscape (**supplementary results, supplementary fig. S1, Supplementary Material online**) and to assess the extent of non-Native American admixture in the studied Amerindian groups (**supplementary results, supplementary fig. S2, supplementary table S2, Supplementary Material online**). Overall, our newly generated data revealed very limited non-Native American admixture, with only one Wichi and two Yanesha samples showing appreciable levels of African ancestry and a low number of individuals per group presenting proportions of European admixture higher than the considered threshold (**supplementary results, supplementary fig. S2, Supplementary Material online**).

Native American genetic structure

Principal Component Analysis (PCA) performed only on the Native American populations retained in the pruned "un-admixed" dataset showed a good resemblance with both the geographic distribution and the linguistic affiliation of analyzed populations (**fig 1a**). Accordingly, it generally confirmed a pattern of North-to-South variability, with the exceptions of Costa Ricans and western Brazilians (i.e. Surui and Karitiana), which instead occupied an outlier position along PC1 and PC2, respectively. In this context, our newly analyzed SA groups formed two well-distinguishable clusters, encompassing all the Andeans from one hand and the Amazonians with Gran Chaco populations on the other.

In order to investigate more deeply patterns of Native American sub-structure and to infer proportions of different ancestral genetic components, we run the unsupervised ADMIXTURE analysis on this Native American "un-admixed" pruned dataset (**supplementary fig. S3, Supplementary Material online**), including

a European population (i.e. CEU) as a further check for non-Amerindian gene flow. At the best-fit model of K = 8, all Native Americans clustered according to seven highly-specific genetic components, also corroborating the absence of any detectable European admixture since the remaining last component was restricted exclusively to CEU (**fig 1b, supplementary fig. S3 and S4, Supplementary Material online**). Overall, the detected Native American genetic ancestries revealed a clear geographic distribution. One component is highly represented in Mesoamerican populations and gradually decreases southward. Another component is mostly observed in Costa Rican groups, such as Maleku, Teribe, Bribri and Cabecar, being also present at lower proportions in Colombian populations (i.e. Waunana, Embera, Wayuu and Kogi). Importantly, two other components were highly enriched in all Andeans and in Peruvian Amazonian populations respectively, thus suggesting an East-West structuring pattern between different Andean- and Amazonian-specific ancestries. Instead, the remaining three components resulted to be private respectively of Karitiana, Surui and Wichi, although this last one was observed also in Chane, Guarani and Jamanadi. Interestingly, at K = 9 Cashibo acquired a private genetic component as well (**supplementary fig. S3, Supplementary Material online**).

Outgroup-*f3* statistics were used to formally infer the sharing of genetic drift between couples of populations (i.e. genetic relatedness between groups). In agreement with ADMIXTURE and PCA results, all non-Andean SA populations were found to be more closely related to each other than to all Andeans and finally to all Meso populations, and symmetrically all the Andean groups appeared to be more closely related to each other, than to all the other SA and then to Mesoamericans (**supplementary fig. S5, Supplementary Material online**). The sole exception to this trend was represented by Costa Rican groups, particularly Cabecar, to which almost all Amazonians (except Shipibo and Yanesha) and Grand Chaco populations are genetically closer with respect to the Andeans.

Consistently with these results and previous studies, the topologies of phylogenetic trees reconstructed with TreeMix generally confirmed a North-to-South progressive pattern of population splitting, with Meso branching out from the tree before the split of Costa Rican and SA groups (**supplementary fig. S6 and S7, Supplementary Material online**; Reich et al. 2012). However, allowing for migration events among populations revealed more complex patterns of genetic relationships between groups, involving changes in the order of splits between Costa Ricans and SA or between the Andean and the non-Andean SA major clades, as well as some connections between single populations (**supplementary results, supplementary fig. S6 and S7, Supplementary Material online**).

Intra-population patterns of genomic diversity

To better understand how the different histories of Meso, Andean and non-Andean SA groups have shaped their genomic diversity, we explored patterns of within-population genetic variation. In particular, to test how the demographic and evolutionary history of each population may have affected the observed ancestry patterns, we calculated the extension of regions with continuous homozygous SNPs (i.e. runs of homozygosity, ROH) and we classified them according to length into three different classes (see Materials and Methods). By investigating the distribution of ROH length over all individuals in each population, we found that Peruvian

and Brazilian Amazonians particularly represented by Surui, Karitiana and Cashibo, showed enrichment of longer ROH classes (**fig. 2a**), especially if compared to Andean groups and to the Wichi from Gran Chaco, who instead harbor shorter ROH segments.

These patterns were further explored with fastIBD by comparing values of identity by descent (IBD) sharing within the analyzed "un-admixed" Native American populations and by visualizing the distribution of the total length of shared IBD segments at different bin thresholds (see Materials and Methods). Consistently with ROH results, values of within-population average IBD-sharing (W_{AB}) appeared significantly higher for Cashibo, Surui and Karitiana. Furthermore, these groups showed tract lengths distributions that – if compared to the rest of Native American groups – are particularly shifted towards the highest classes of IBD binning, which indicates more recent genetic relatedness (**fig. 2b, supplementary fig. S8, Supplementary Material online**). Compared to the rest of Native Americans, also Wichi and Cabecar showed relatively high values of W_{AB}, but their tract lengths distributions are within the ranges observed for all the other Amerindian populations analyzed (**fig. 2b**).

Fine-scale inter-population haplotype sharing

To evaluate at a finer scale the genomic structure of un-admixed Native American groups, we applied the fineSTRUCTURE clustering algorithm to the CHROMOPAINTER "chunk-counts" matrix of individual haplotype sharing. We first included also European (CEU), East Asian (CHB) and African (YRI) groups to definitely verify the absence of post-Colombian admixture (s**upplementary results and supplementary fig. S9, Supplementary Material online**), and then we considered only the Native American groups (excluding Chipewyan) to specifically focus on intra-Amerindian haplotype sharing patterns.

Overall, clusters of genetically homogeneous individuals identified by fineSTRUCTURE largely matched with population labels (**fig. 3**, **supplementary fig. S10, Supplementary Material online**). Few exceptions involved individuals belonging to closely related groups. For example, one Guahibo individual clustered with all the Piapoco, as does one Guaranì with neighboring Chane and one Shipibo within the Cashibo cluster. Analogously, Aymara individuals previously sampled in Bolivia (Reich et al. 2012) and Bolivian Aymara from our study appeared highly intermingled, as well as some Cabecar individuals with the other Costa Rican groups (i.e. Maleku, Teribe, Bribri). Albeit we caution that fineSTRUCTURE hierarchical clustering does not imply any evolutionary relationship between distinct clusters, and thus should not be interpreted as a phylogenetic tree (Lawson et al. 2012; Leslie et al. 2015), if considering the clusters independently (**fig. 3)**, they perfectly matched with the clades identified by TreeMix and were consistent with the broad pattern described by outgroup-*f3* statistics (**supplementary fig. S5, S6 and S7, Supplementary Material online**). In fact, all Andeans formed a clade that departs from all the other SA. Similarly, and in agreement with the genetic difference between Peruvians and other Amazonian groups appreciable with TreeMix, all Peruvian Amazons formed a separate clade among each other, the sole exception being Huambisa that instead clustered with all the other Amazonian groups, as well as with Chane and Guarani (**fig. 3**). As concerns Mesoamericans, they all clustered together presenting internal relationships again in agreement with TreeMix results, i.e. Pima split

first with respect to the Central Mexican groups of Tepehuano, Zapotec and Mixe on one hand and the Southern Mexican Tzotzil and Guatemala populations on the other. Finally, all the Wichi (i.e. the 17 new individuals from our study and four previously published by Reich et al. 2012) formed an outlier cluster, and so did a separate clade encompassing all the Costa Rican groups (i.e. Cabecar, Maleku, Teribe, Bribri).

Comparison between the clustering pattern and the "chunk-lengths" matrix pinpointed additional interesting features (**supplementary fig. S11, Supplementary Material online**). First, Karitiana, Surui, Cashibo, Wichi and Cabecar, who showed higher proportions of homozygous segments and of total length of shared IBD (**fig. 2b**), were also the ones presenting the lowest (~ 0) proportion of haplotype "copying" with other groups. However, the difference between these groups is that while Surui, Karitiana and Cashibo clustered within their corresponding clade (i.e. West and Peruvian Amazonian, respectively), the Wichi and Cabecar were outliers with respect to all the other groups. Furthermore, only few sets of populations revealed evident traces of high haplotype sharing outside from their own cluster, namely the Chane and Guarani with the Wichi, and the Colombian Wayuu, Waunana, Embera and Kogi with the Costa Rican clade, thus signaling possible events of gene flow between these populations. However, attempts to date these admixture events using the GLOBETROTTER pipeline, which is based on the "chunk-lengths" matrix produced by CHROMOPAINTER, were unsuccessful. In fact, the co-ancestry curves were too noisy to successfully fit an exponential function describing the admixture parameters (**supplementary fig. S12, Supplementary Material online**). This may be due in part to the low haplotype resolution and ascertainment bias of SNP-chip data and in part to the fact that the populations involved in this study are related to each other to the point that the method is unable to produce clear patterns of haplotype chunks belonging to one or another ancestral source.

Demographic modelling

We attempted to formally assess the genealogical relationships between the Andean and non-Andean SA with respect to the Meso populations with simplified four-population tree-like models by applying *f4* and D-statistics for all possible combinations of the studied groups (**supplementary table S3 and S4, Supplementary Material online**). Overall, tests in the form of (CHB, Meso; SA, SA) and (CHB, Andean; Meso, non-Andean) confirmed that SA groups are consistent with forming a clade with respect to the Meso groups. The only populations breaking this trend were the Cabecar when considered in the Meso position and the Tzotzil when SA were specified as combinations of Andean and Amazonian groups, respectively (**supplementary results, supplementary table S3 and S4, Supplementary Material online**). In fact, Andeans and non-Andeans resulted differently related to Mesoamericans when we tested the topology in the form of (CHB, Meso; Meso, Andeans or non-Andeans) when the two Meso populations were the Zapotec and the Tzotzil, respectively (**supplementary results, supplementary table S5, Supplementary Material online**).

That being so, to identify demographic models explaining the intricate relationships between Mesoamericans, Andean and non-Andean SA, we finally used the admixture graph (AG) approach as described in Materials and Methods. The simplest AG test (**supplementary fig. S13a, table S6, Supplementary Material online**)

modeled Andean and non-Andean SA as descending from a common ancestral population that is a sister group of the Zapotec and provided good fits except for some non-Andean populations (i.e. Guahibo and Huambisa). However, poor fits to the data extended to all non-Andean groups when we included the Tzotzil in the demography as the last Meso group before the divergence within SA **(supplementary fig. S13b, Supplementary Material online)**. In these cases, none of the combinations between Andeans and non-Andeans can be successfully modeled as forming a clade with respect to the Tzotzil (**supplementary table S7, Supplementary Material online**). Since AGs without admixture represent poor fits in the history of these populations, we tried to model alternative topologies allowing for mixture events. In particular, following the results of *f4* and D analyses (**supplementary results, Supplementary Material online**), we tested AG configurations connecting the Andeans to the Zapotec or the non-Andeans to the Tzotzil through one admixture event between a lineage ancestral to these Mesoamerican groups and the other SA ancestral pool, respectively (**supplementary fig. S14a and 14b, table S8 and S9, respectively, Supplementary Material online**). Interestingly, while both such cases provided no good fit, a demography where the Andeans are instead admixed between a deeper Mesoamerican node (i.e. ancestral to the Tepehuano) and the non-Andean SA lineage showed several good fits and thus cannot be definitively ruled out **(supplementary fig. S14c, table S10, Supplementary Material online)**. However, among all the tested demographic models, the ones that maximized the fits to the data are those where the Tzotzil were modeled as a mixture of ancestry strands related to a lineage leading to all non-Andean SA and to a node ancestral to the Zapotec (**fig. 4a, supplementary table S11, Supplementary Material online**). Furthermore, Guahibo and Huambisa groups can be successfully modeled only if considering a further admixture event between a node ancestral to the Andeans and a node ancestral to the Zapotec, before the above-mentioned admixture involving non-Andean groups with the Tzotzil (**supplementary fig. S14d, table S12, Supplementary Material online**).

We finally attempted to test the demography within SA and especially between the different non-Andean clades taking into account the results from Reich et al. (2012) and modeling our newly generated data (**fig. 4b, supplementary fig. S15, Supplementary Material online**). We found good fits for the Gran Chaco (represented by the Wichi) as the first non-Andean clade branching out, and the Guarani could be successfully modeled only as admixed between a node ancestral to this Gran Chaco lineage and a node leading to other Amazonians, thus confirming results from Reich et al. (2012).

Importantly, the Peruvian Amazonian groups (i.e. Cashibo, Shipibo, Yanesha) best fit when modeled as admixed between the SA Amazonian lineage and the Andean clade (**fig. 4b, supplementary table S13, Supplementary Material online**). On the contrary, trying to fit them in a demography without admixture generally resulted in *f-statistics* that are more than $|Z|>3$ standard errors from expectation, thus supporting a model with admixture as a better choice (**supplementary fig. S15, Supplementary Material online**). It is worth nothing that in such a model the Yanesha presented an extra affinity with the YRI outgroup, i.e. an outlier *f4* statistics ($Z < -3$) in the form (YRI, Zapo/Wichi; Surui/Karitiana, Yanesha). This result complies with some outlier *f4* and D statistics in the form (CHB, Meso; non-Andean, Yanesha) (**supplementary table S3, Supplementary Material online**) and may suggest a possible remnant of cryptic post-Columbian African

admixture undetected by previous analyses. The Ashaninka revealed good fits for both models – i.e. either accounting for additional mixture or not - but again with a slight increase in fit for the admixture case (**supplementary fig. S15 and table S13, Supplementary Material online**). Overall, the "Andean" admixture component in Peruvian Amazonians was very low, ranging from ~5% in Ashaninka to ~15% in Yanesha, which is consistent with them harboring mostly a non-Andean and specifically an Amazonian genetic ancestry (**fig. 4b**).

Discussion

To shed light into the genetic history of SA populations with fine-scale genomic analyses and to overcome the inferential limitations imposed by recent post-Columbian admixture, we genome-wide genotyped individuals representative of 10 South and one Central American ethnic groups. In particular, to address the investigation of both broader and local scale patterns of peopling processes and of genetic relationships East and West of the Andes/Amazonian divide, we integrated previous Native American reference panels (Reich et al. 2012) with new data from high-altitude Andean groups from Peru and Bolivia, Peruvian Amazonians, Wichí from the Gran Chaco and Mexican Mayan Tzotzil (see Materials and Methods). The reduced non-Native American ancestry detected especially in the newly typed samples (**supplementary results, supplementary fig. S2, supplementary fig. S9, Supplementary Material online**) allowed us to exclude recently admixed individuals, still relying on a good sample size per group (**supplementary table S1, Supplementary Material online**). Global population structure analyses on the Native American "un-admixed" dataset revealed a clear-cut pattern of structuring between groups, coupled with substantial intra-population homogeneity (**fig.1**). Overall, individuals belonging to the same population formed tight clusters on the PCA space (**fig.1a**) and presented similar admixture proportions (**fig. 1b**). Even at the finer-scale structuring level explored by haplotype-based fineSTRUCTURE analyses, individuals were consistently found to cluster according to their respective population, with only few exceptions of single samples assigned to neighboring groups (**fig. 3, supplementary fig. S10, Supplementary Material online**).

Genetic relationships between populations were broadly concordant with their language family affiliation and corresponded to geographic locations at a local scale (**fig.1a**). The Meso groups showed a general North to South clustering pattern (with the exception of the outlier position of Costa Ricans), while among the Peruvian samples emerged a sharp distinction between the tight cluster of high-altitude Andeans and the Amazonians, the latter grouping with the bulk of other non-Andean SA from Brazil, Colombia and Gran Chaco (**fig. 1a**). Inferences of ancestry proportions showed the presence of distinct Native American genetic components largely corresponding to one Central American (i.e. highest in all Mexican groups), one Costa Rican, one Andean and different non-Andean SA components maximized in Peruvian Amazonians, Brazilian Surui and Karitiana, and Wichi from the Gran-Chaco (**fig. 1b, supplementary fig. S3, Supplementary Material online**).

Proportions of Meso components (**fig. 1b**) were observed at different levels among some non-Andean populations, suggesting shared ancestry or recent contacts between Meso and SA groups. In particular, this

latter case could explain the proportions of the Costa Rican-like component observed in Northern SA from Colombia (i.e. Kogi, Embera, Waunana, Wayuu), who in fact occupied an intermediate position in the PCA with respect to the neighboring Amazonians (**fig. 1a**). In agreement with previous studies (Reich et al. 2012, Homburger et al. 2015), admixture between Colombian groups and Costa Ricans emerged also from several TreeMix runs (**supplementary results, supplementary fig. S6 and S7, Supplementary Material online**) and was supported by the high sharing of haplotypes between these two clusters revealed by CHROMOPAINTER analyses (**supplementary fig. S9 and S11, Supplementary Material online**). Patterns of haplotype sharing from outside their own-specific cluster were observed also for the Chane and Guarani groups, which revealed significant proportions of the Wichi-like component. In fact, they clustered with the Wichi in TreeMix phylogenies, although showing migration edges with Amazonians (**supplementary results, supplementary fig. S6 and S7, Supplementary Material online**).

Analyses of intra-population diversity, measuring both the length of genotype homozygous tracts (**fig. 2a**) and the genome-wide haplotype IBD sharing between individuals belonging to the same group (**fig. 2b**), concurrently confirmed a general pattern of higher drift experienced by non-Andean SA and Costa Rican groups with respect to the Meso and Andean populations. This likely reflects known differences in the past population histories and effective population sizes between these groups (Wang et al. 2007). In fact, during pre-Columbian times the area of present-day Mexico in Mesoamerica and the Andes witnessed the rise of complex urban societies, while in other regions the populations remained mainly organized in smaller groups thus probably incrementing inbreeding within populations and experiencing variable degrees of isolation (D'Altroy 2014; Arias et al. 2018). Nevertheless, detected differences in intra-group genetic patterns allowed the distinction between the effects of substantial inbreeding and/or small effective population sizes (N_e) from the ones of prolonged isolation. For instance, the Brazilian Amazonian groups of Surui and Karitiana and the Peruvian Cashibo, besides exhibiting private genetic components according to ADMIXTURE analysis (**fig. 1b**), also presented higher long-tract ROH and IBD values (**fig. 2**), as well as longer tip branches in the inferred TreeMix trees (**supplementary fig. S6 and S7, Supplementary Material online**) and admixture graphs (**supplementary tables S4-S12, Supplementary Material online**), thus signaling evidence of recent population-level relatedness and high genetic drift. This confirmed previous results obtained for Surui and Karitiana (Wang et al. 2007, Li et al. 2008, Verdu et al. 2014), and is in line with the reduced uniparental lineage composition already observed for Cashibo (Di Corcia et al. 2017). On the contrary, Wichi and Cabecar, despite showing overall high levels of homozygosity and intra-group haplotype sharing, presented an average shorter length of homozygous tracts and chunks of shared haplotypes, more compatible with a prolonged isolation rather than high inbreeding (**fig. 2**). This is reflected also in their outlier position with respect to other Central and South American clusters identified by fineSTRUCTURE (**fig. 3, supplementary fig. S10, Supplementary Material online**).

For what concerns the Wichi, these patterns were in agreement with the strong founder effect and the subsequent high diversification of mitochondrial lineages observed in a previous study (Sevini et al. 2013), thus confirming that this region has been long populated and that Wichi remained genetically isolated from

both the neighboring Andes on the West and Amazonia on the North. Effects of such an isolation were evident in the population-specific clustering of Wichi both in genotype-based ADMIXTURE and haplotype-based fineSTRUCTURE analyses (**fig. 1b** and **fig 3**). As for their relationships with the other main branches of SA lineages (i.e. Andean and Amazonian), the instable position of Wichi in TreeMix phylogenies was paralleled by an ancestral connection between Meso and Andean clades each time the Gran Chaco group split out before the Andeans instead of being a sister clade of all the other non-Andean SA (**supplementary results, supplementary fig. S6 and S7, Supplementary Material online)**. Consistently with the closer relationship with Amazonians outlined by both PCA and outgroup-*f3* analyses, formal tests of treeness through four population statistics (**supplementary table S3, Supplementary Material online)** showed that the Wichi belong to the same non-Andean lineage of all Amazonians. Importantly, admixture graphs further suggested that they likely descend from one of the first splits within this lineage (**fig. 4b, supplementary fig. S15, Supplementary Material online**).

In fact, when formally tested with *f4* and D statistics, the non-Andean SA together with Costa Ricans were consistent with forming a clade with respect to Meso and Andeans (**supplementary results, supplementary table S3, Supplementary Material online**). This is also in agreement with them being more closely related to each other than to all Andeans according to the outgroup-*f3* analyses (**supplementary fig. S5, Supplementary Material online**). Instead, *f4* and D statistics revealed that Andeans and non-Andeans are differently related to the Meso groups. In fact, all non-Andean SA and Cabecar if tested as forming a clade together with the Andeans showed a significant extra genetic affinity with the Tzotzil (**supplementary results, supplementary table S3, Supplementary Material online**). Furthermore, when we directly assessed to which Meso lineage the SA are more closely related, the Andeans revealed a closer relationship to the Zapotec with respect to the Tzotzil, whereas the opposite applied for all non-Andean and Cabecar groups (**supplementary results, supplementary table S3, Supplementary Material online**). Admixture graphs corroborated these patterns and formally confirmed that for Meso and SA populations a demographic model that follows the simple tree-like topology of TreeMix does not fit with the data unless accounting for at least one admixture event between the main branches (**supplementary fig. S13a, table S4, Supplementary Material online**).

More specifically, the best fitting model obtained by testing all different combinations of possible admixture events suggested by outlier *f4* and D statistics, was the one where the Tzotzil descend from an admixture between a Meso branch and the South American lineage contributing to all present-day non-Andean populations, after the divergence between Andean and non-Andean SA (**fig. 4a**). This demographic model is in accordance with two opposite scenarios (and with all possible events in between). According to the first scenario, if the split between the Andean and non-Andean lineages happened within SA, back-migrations and/or bi-directional gene flow between Meso and SA extended far beyond the attested contacts between North SA and the Caribbean region (see also Reich et al. 2012; Moreno-Estrada et al. 2013). In particular, these processes would have involved northward diffusions of the ancestors of all non-Andean SA that can be genetically detected up to present-day highlands of Southern Mexico. However, this hypothesis does not seem to find support from the ancient genomes recently studied by Moreno-Mayar et al. (2018b) and Posth et al.

(2018). In fact, both these studies did not detect any evidence of back migrations from South America into Central America, even though the only ancient data from the entire Meso and North SA regions (i.e. northern Peru) are represented by two samples from Belize dated > 7.4 ka (Posth et al. 2018). Therefore, it is not possible to exclude that such population movements occurred after this period. According to the second scenario, if the split between the Andean and non-Andean lineages occurred before the entrance into SA (i.e. somewhere in North America or most likely in Mesoamerica), the admixture involving the Tzotzil ancestors would have happened outside SA between two already diverged Meso lineages, one ancestral to the Zapotec and the other one instead related to all present-day non-Andean SA. This latter lineage could therefore represent an additional gene pool, distinct from the one leading to the Andeans, which spread into South America on one side and admixed in Mesoamerica on the other. However, it is not possible to attest which of these events happened earlier.

Interestingly, we cannot exclude an alternative model where the Andeans are admixed between the non-Andean SA lineage and a more ancient Meso branch of the Northern Mexican Tepehuano lineage (**supplementary fig. S14c, table S8, Supplementary Material online**). This alternative demography is more concordant with the second scenario underlying the previous model and reconciles the pattern observed in several TreeMix replicates, which showed a migration edge between the Andean cluster and a node ancestral to all the Mesoamericans (**supplementary results, supplementary fig. S6 and S7, Supplementary Material online**). By considering recent findings emerged from ancient DNA, the connection between the Andeans and a northern Meso lineage (i.e. ancestral to Tepehuano) could be associated with the small amount of gene-flow (2-4 %) identified by Posth et al. (2018) between ancient individuals from the California Channel Islands (California) and the ancient Late Central Andeans, which dates back to sometime after ~4.2 ka.

Finally, both models described so far, together with TreeMix results (**fig. 4a, supplementary fig. S6, S7 and S14c, Supplementary Material online**), could imply the same event of slow population replacement that started around 9 ka and went on for several thousands of years (Moreno-Mayar et al. 2018b; Posth et al. 2018). Indeed, a connection between the high-altitude Andeans and a more rooted (i.e. ancient) Meso lineage with respect to non-Andeans is in agreement with the longer standing genetic continuity attested so far in the Andes (started 8-9 ka; Lindo et al. 2018b; Posth et al. 2018) as compared to other regions of SA. For instance, in Patagonia this population replacement occurred not earlier than ~5 ka (Moreno-Mayar et al. 2018b). Therefore, such a long-term population replacement could be at the origin of the East and West of the Andes genetic structure that we observe in present-day Native South Americans.

Within South America, we finally confirmed the role of sharp genetic barrier represented by the Andes. This pattern holds even considering the geographically close groups from both the Gran Chaco region in Argentina and the neighboring Peruvian Amazon included in our newly typed samples. Nevertheless, even if these low-altitude Peruvian groups are indeed of Amazonian ancestry (i.e. more closely related to the rest of non-Andean populations; **fig. 1a**, **supplementary fig. S5, Supplementary Material online**), a small proportion of gene-flow from the Andeans into the Peruvian Amazonians can be detected (**fig. 4b**), in agreement with previous results from uniparental data (Barbieri et al. 2014; Di Corcia et al. 2017). Conversely, we did not detect

evidence of recent gene flow in the opposite direction, from non-Andean sources into the Andean groups, as suggested in a recent study (Harris et al. 2018).

While being aware that the over-simplified nature of demographic models cannot explain the actual complexity of thousands of years of population history, and that isolation-by-distance may account for a great deal of the genetic structure among present-day populations (thus masking most of the demographic events occurred through time), our results support the view that a simple divergence from common Mesoamerican ancestors along with an unidirectional latitudinal expansion is not sufficient to explain the genetic diversity of Native South Americans.

Overall, the present study reconciles the genetic structure of modern South American populations with the recent findings emerged from ancient DNA analyses (Moreno-Mayar et al. 2018b; Posth et al. 2018) and provide intriguing hypotheses that could be tested with new data from additional ancient samples. Furthermore, while future studies will further benefit from the analyses of new complete genomes, our results stress the importance of implementing accurate sampling strategies and of selecting representative populations based on historical/linguistic and anthropological information to add new insights into the pre-Columbian history of Native Americans.

Materials and Methods

Samples collection and genotyping

In this study, we analyzed a total of 229 individuals belonging to 11 ethnic groups from Meso and South America, namely: Tzotzil from Chiapas (Mexico); Cashibo, Shipibo, Huambisa, Ashaninka and Yanesha from Peruvian Amazon (Peru); Quechua, Aymara and Uros from the Peruvian Andes (Peru); Aymara from the Bolivian Andes (Bolivia); Wichi from the Gran-Chaco (Argentina).

Subjects were surveyed for being native of their respective ethnic group by at least three generations. Saliva samples from Bolivian Aymara were collected with the Oragene-DNA Self Collection Kit OG500 (DNA Genotek Inc., Ottawa, Ontario, Canada). Genomic DNA was purified with the *prepIT-L2P* protocol (DNA Genotek, Ottawa, Ontario, Canada) and quantified by fluorometric methods (Qubit® dsDNA BR Assay Kit, Life Technologies, Carlsbad, CA, USA). The other samples were collected and processed for DNA extraction as described elsewhere (Barbieri et al. 2011, Sevini et al. 2013, Barbieri et al. 2014, Moreno-Estrada et al. 2014, Di Corcia et al. 2017).

The participants provided a written informed consent to data treatment and project objectives. Approvals from local Institutional Review Boards were obtained as well. In particular, authorization by the Unidad de Identificación Genética (UNIGEN) de la Universidad Mayor de San Andrés (UMSA) was obtained for the collection of new samples from Bolivia. Approvals by the representative of the regional organization of Ucayali (AIDESEP) and by the president of COSHIKOX (Consejo Shipibo Conibo Xetebo), by the representative of the Yanesha political association and the FECONAYA (Federacion de Comunidades Nativas Yanesha), by the University Hospital of Maternity and Neonatology of the Universidad Nacional de Cordoba and the Ministry of Health of the province of Chaco, as well as by the University of Guadalajara, the National

Institute of Medical Sciences and Nutrition Salvador Zubirán (INNSZ) and the National Institute of Genomic Medicine (INMEGEN), were previously obtained for already collected samples from Peru (Barbieri et al. 2011; Barbieri et al. 2014; Di Corcia et al. 2017), Argentina (Sevini et al. 2013) and Mexico (Moreno-Estrada et al. 2014).

On April 8th, 2013 the Bioethics Committee of the University of Bologna also approved all the procedures concerning this study (within the framework of the ERC-2011-AdG 295733 project). Moreover, this study was designed and conducted according to the relevant guidelines, regulations and ethical principles for medical research involving human subjects stated by the WMA Declaration of Helsinki.

All DNA samples (n = 229) were genotyped for ~720,000 SNPs distributed along the whole genome at an average spacing of 4 Kb, with the HumanOmniExpress 1.1 BeadChip (Illumina, San Diego, CA, USA). Genotyping experiments were performed at the facilities of the Center for Biomedical Research & Technologies of the Italian Auxologic Institute.

Data curation

Obtained genotype data were filtered using the PLINK software package v.1.07 (Purcell et al. 2007) by applying a series of QC steps to remove individuals and variants with low call rates, SNPs with ambiguous alleles and inbred individuals. More precisely, we excluded variants with missing call rates exceeding 5%, SNPs showing significant deviations from the Hardy-Weinberg equilibrium ($p < 0.01$) and those with ambiguous A/T or G/C strand polymorphisms. As for per-individual QC, we removed samples showing more than 2% of missing genotypes (n = 28) and/or a high degree of IBD sharing (n = 24). In particular, we estimated inbreeding for each pair of individuals on an LD pruned dataset ($r^2 > 0.1$), by calculating the genome-wide proportion of shared alleles and we randomly excluded one individual from each pair showing an IBD coefficient higher than 0.25. Moreover, because populations that experienced long-term isolation and small N_e are generally characterized by higher mean values of IBD, we assessed with the Grubb test (package *outlier* of the R software; Komsta, 2011) the presence of outlier values in the IBD distribution of all possible pairs of individuals belonging to the same ethnic group. We then removed one individual from every outlier pair showing *p-values* lower than 0.05.

A final high-density "clean" dataset of 178 samples typed for 660,772 SNPs was used for merging with a reference population panel of publicly available genome-wide data from the HGDP project (Li et al. 2008), the 1000 Genomes Project (The 1000 Genomes Project Consortium) and Native American populations from Reich et al. (2012) (**supplementary table S1, Supplementary Material online**). Before merging procedure, we performed the same QC described above on each reference dataset separately. After merging, we obtained an "extended" dataset including 431 additional individuals from 50 Native American ethnic groups typed for a common set of 207,165 SNPs. This dataset was used to perform the haplotype-based analyses described below, and thinned for genotype-based analyses by removing SNPs in high LD ($r^2 > 0.2$) within a sliding window of 50 SNPs advanced by 5 SNPs at the time, as well as variants with a minor allele frequency (MAF) < 0.01, thus obtaining a pruned "extended" dataset consisting of 96,991 SNPs.

Population structure and admixture analyses

PCA were carried out on the pruned "extended" dataset by using the *smartpca* method implemented in the EIGENSOFT package v6.0.1 (Patterson et al. 2006). PCA was first applied on all the worldwide populations to check for the presence of genotyping errors or inconsistency between the data (**supplementary fig. S1, Supplementary Material online)**. Then, we performed PCA only on the Native American groups included in the pruned "un-admixed" dataset, i.e. after having checked for limited non-Native American admixture (**fig 1a**). Indeed, since recent events of European and African admixture may complicate the study of pre-Columbian history, we assessed the presence of non-Native American genetic components in the analyzed American populations of the pruned "extended" dataset by running unsupervised clustering analyses implemented in ADMIXTURE v.1.22. (Alexander et al. 2009). First, we tested K = 2 to K = 15 clusters including 32 European, African, and Asian groups in addition to 66 American groups and we excluded all American individuals (or entire groups) showing proportions of European and African ancestry higher than 2% and 1% respectively, by considering K = 6 because at this number of clusters all the main non-Native American ancestral components were resolved (**supplementary fig. S2, Supplementary Material online)**. Then, we replicated ADMIXTURE testing from K = 2 through K = 10 on the remaining 43 Amerindian populations contained in this pruned "un-admixed" dataset, including also European ancestry CEU to further check for the absence of external admixture (**fig. 1b**, **supplementary fig. S3, Supplementary Material online)**. For each K tested, we performed 50 independent ADMIXTURE runs with a different random seed to monitor convergence and only those with the highest log-likelihood were considered for the plots. Concurrently, we calculated cross-validation (CV) errors for each K in order to identify the most reliable number of genetic clusters concordant with the data (**supplementary fig. S4, Supplementary Material online)**. The reliability of obtained ADMIXTURE results was further assessed applying the *pong* algorithm (Behr et al. 2016). This method identifies the number of modes (i.e. number of different Q matrices) present across the 50 independent runs performed for each given K and evaluates the average pairwise similarity within and between the eventual different modes. The maximum-likelihood clustering approach implemented in ADMIXTURE indeed ignores possible cases of multimodality, i.e. multiple sets of membership coefficients inferred from a set of runs on the same data that may differ non-trivially as belonging to different modes (Behr et al. 2016). The ADMIXTURE performed on the "extended" dataset produced highly consistent results across the different runs, and up to K = 9 *pong* identified no more than two modes per K, with the average similarity being always > 99.9% within mode and > 88% between modes. In particular, the K = 6 run that we considered for inferring European and African admixture belongs to the major mode (i.e. the one replicated in most runs) of the two identified, with a between-mode similarity of 88%. The ADMIXTURE performed on the "un-admixed" dataset produced highly consistent results. No more than four modes per K were identified and the average similarity within mode and between modes were always > 99.9% and > 87% respectively. More specifically, the K = 8 run reported in **fig. 1b** belongs to the major out of four modes with a between-mode similarity of 93%.

To formally test for the presence of African and European admixture on the pruned "un-admixed" dataset, we

also calculated *f3*-statistics using the *qp3Pop* program implemented in the ADMIXTOOLS v3.0 package (Reich et al. 2009). We considered negative statistics with Z-score values below -2 as significant signals of admixture (**supplementary table S2, Supplementary Material online**). The same software was used to apply an outgroup-*f3* approach in order to measure the sharing of genetic drift (i.e. genetic similarity) between each pair of Native American populations (Raghavan et al. 2014; **supplementary fig. S5, Supplementary Material online**). In particular, we tested each possible pair of Native American groups with $N \geq 5$ as sources of admixture and an outgroup population (YRI) as "target" of such an admixture.

Finally, we used TreeMix v1.12 (Pickrell & Pritchard 2012) to build phylogenetic trees on the Amerindian populations present in the pruned "un-admixed" dataset including CEU as root population for an additional check of further signals of European gene flow. TreeMix was first used to construct a tree without allowing any migration and then we tested sequentially 1 to 4 migration events. These analyses were performed both including only populations with $N \geq 5$ individuals, as well as on the whole dataset implementing the TreeMix sample size correction flag (**supplementary fig. S6 and S7, Supplementary Material online**).

Intra-population genetic structure

To explore patterns of within population genetic variation, we calculated the extension of ROH segments and the average length of genome shared IBD. ROH were calculate on the "un-admixed" dataset considering only groups with $N \geq 5$ individuals (to reduce possible biases due to small sample sizes) by using the command "*--homozyg*" implemented in the PLINK software package v.1.07 (Purcell et al. 2007) under default parameter settings. SNPs were considered to be part of a homozygous segment, when the proportion of overlapping homozygous windows was above 5% (Anagnostou et al. 2017). Then, we considered a default Gaussian fitting of the ROH length distribution, using the *Mclust* function from the R package mclust V3 (Fraley and Raftery, 2002), which identified three different ROH length classes (**fig. 2a**).

Patterns of IBD sharing within populations were estimated on the phased data by using the *fastIBD* method implemented in the BEAGLE 3.3 software (Browning and Browning, 2011). The phase of haplotypes for the "un-admixed" dataset was statistically reconstructed using SHAPEIT2 v2.r790 (Delaneau et al. 2013) by applying default parameters and HapMap phase 3 recombination maps. FastIBD was run ten times for each chromosome using different random seeds. To call IBD blocks we post-processed results with the 'plus-process-fibd.py' pipeline modified by Ralph et al. 2013. We set the fastIBD threshold to 1e-10 and considered only blocks longer than 1 cM. As summary IBD-statistics, we computed the total length of genome shared IBD averaged over the number of possible pairs of individuals within each population (W_{AB} metric; Atzmon et al. 2010). The average IBD-sharing was calculated for nine different bin categories (**supplementary fig. S8, Supplementary Material online)** (Moreno-Estrada et al. 2014).

CHROMOPAINTER and fineSTRUCTURE analyses

To explore fine-grained population structure and define clusters of genetically homogeneous individuals, we exploited the haplotype-based approach implemented in CHROMOPAINTERv2/ fineSTRUCTURE. Samples

were phased with the SHAPEIT software as specified above. We applied the CHROMOPAINTERv2/fineSTRUCTURE pipeline (Lawson et al. 2012) separately, but following the same steps detailed below to 1) the "un-admixed" dataset including CEU, CHB, YRI to further control for allele sharing pattern with non-Native American populations and 2) the "un-admixed" dataset including only the Meso and South American groups (**fig. 3, supplementary fig. S9, S10 and S11, Supplementary Material online)**.

We first estimated the mutation/emission and the switch rate parameters with 10 steps of the Expectation-Maximisation (E-M) algorithm on a subset of chromosomes {4, 10, 15, 22} using every individual both as "donor" and "recipient". Then, we averaged the obtained values across chromosomes (weighting by the number of markers) and individuals, and we used the estimated mutation/emission and switch rate parameters to run CHROMOPAINTER again on all chromosomes, considering a parameter k = 50 to specify the number of expected chunks to define a region. This value was suggested to be preferable compared to the default value of 100 (Leslie et al. 2015) when painting closely related populations. The obtained matrix of haplotype sharing "chunk" counts was summed up across all the 22 autosomes and submitted to the fineSTRUCTURE clustering algorithm version fs2.1 (Lawson et al. 2012). We ran fineSTRUCTURE pipeline by setting 1,000,000 "burn-in" MCMC iterations, followed by additional 2,000,000 iterations and sampling the inferred clustering patterns every 10,000 runs. Finally, we set 1,000,000 additional hill-climbing steps to improve posterior probability and merge clusters in a step-wise fashion. Individuals were hierarchically assembled into clusters until reaching the final configuration tree. We then applied the GLOBETROTTER algorithm (Hellenthal et al. 2014) in the attempt to infer dates for the admixture events between Native American groups suggested by different analyses. We run GLOBETROTTER on CHROMOPAINTER runs performed only on the Meso and South American groups of the "un-admixed" dataset, grouping the samples according to the clusters identified by fineSTRUCTURE and excluding each time from the donors the cluster tested as target of admixture in GLOBETROTTER. In particular, we tried to fit the Waunana_Embera and the Wayuu clusters as target of admixture by using all the other Native American clusters of the "extended un-admixed" dataset as parental proxies (**supplementary fig. S12, Supplementary Material online)**. Moreover, we performed an additional CHROMOPAINTER/GLOBETROTTER workflow on the phased high-density "clean" dataset (i.e. consisting of only our newly-typed populations, but relying on a greater number of markers), in the attempt to date the admixture events observed for the Peruvian Amazonian groups between all Andean and non-Andean possible sources present in the original dataset (**supplementary fig. S12, Supplementary Material online**). All GLOBETROTTER runs were conducted according to guidelines reported in Hellenthal et al. (2014) and performing a first run standardizing over a null individual.

Tests for treeness

We assessed consistency with a four-population tree topology using the functions implemented in ADMIXTOOLS v3.0 to calculate *f4* (Reich et al. 2009) and *D*-statistics (Green et al. 2010). In particular, we tested if all SA form a consistent clade with respect to all Meso and an outgroup (Han Chinese, CHB),

considering all possible combinations of populations with *f4* and *D* in the form (CHB, Meso; SA, SA) (**supplementary table S3, Supplementary Material online)**. Using again CHB as outgroup, we then tested whether Meso populations were consistent with forming a clade with non-Andean SA (i.e. Amazonians and Gran Chacos) testing all possible combinations of *f4* and *D* in the form (CHB, Andeans; Meso, non-Andean SA) (**supplementary table S4, Supplementary Material online**). Finally, we tested if Andean and non-Andean SA were differently related to present-day Meso groups by testing all possible combinations of *f4* and *D* in the form (CHB, Meso; Meso, Andeans) and (CHB, Meso; Meso, non-Andean SA) (**supplementary table S5, Supplementary Material online)**.

Admixture Graphs

To test more refined demographic hypotheses explaining relationships between Meso and SA populations, as well as within SA itself, we applied the *f*-statistic based modelling approach implemented in the *qpGraph software* of the ADMIXTOOLS v3.0 package (Reich et al. 2012). We set Yoruba from Nigeria (YRI) as the outgroup and CHB as the last non-Native American population in root position to all Meso and SA considered groups. To keep the models simple, in order to avoid overfitting and neglect gene-flow between closely related populations that will add unnecessary complexity (Patterson et al. 2012), we considered one population as representative of the main clusters that were identified according to TreeMix and fineSTRUCTURE. In particular, we used Northern Tepehuano, Southern Zapotec and Mayan Tzotzil as representative of the three Mesoamerican clades and then we iteratively tested all possible combinations of one non-Andean and one Andean group as representatives of the two main South American gene pools (**fig. 4a, supplementary fig. S13 and S14, table S6-S12, Supplementary Material online**). As for the intra-SA models, we also added an additional Andean population (but not all three together to simplify the otherwise complex high intra-Andean gene flow) and one representative for each non-Andean cluster (i.e. Grand Chaco, Peruvian Amazonians and Brazilian Amazonians) (**fig. 4b, supplementary fig. S15 and table S13, Supplementary Material online)**. Unless otherwise specified, we considered as significant evidence of rejection the models presenting one or more outlier f-statistics (with Z-score > |3|) and significant p-values (< 0.05) for the nominal χ^2 statistic, indicator of no evidence for a poor fit (Patterson et al. 2012). Inversely, models with no outlier f-statistics and presenting slightly non-significant p-values (> 0.01 and < 0.05) are still discussed as reasonable fits in the absence of better fitting alternative models.

Supplementary Material

Supplementary results, Supplementary figures S1-S15 and tables S1-S13 are available at Molecular Biology and Evolution on line (http://www.mbe.oxfordjournals.org/).

Data Availability

The genotype data generated during the current study is available at https://figshare.com/articles/South_American_dataset_Gnecchi-Ruscone_et_al_2019_/7667174

Acknowledgments

We would like to thank all the local communities who participated to the study without whom this work would have not been possible. We are also grateful to Hector Rangel-Villalobos (Instituto de Investigación en Genética Molecular, Centro Universitario De La Ciénega, Universidad de Guadalajara, Jalisco, Mexico) and to all participants in sampling expeditions for their valuable help in organizing sample collections. Finally, we would like to thank Cosimo Posth (Max Planck Institute for the Science of Human History, Jena, Germany) for his helpful advises in the interpretation of the obtained results. G.A.G.R., S.D.F. and S.S. were supported by the ERC-2011-AdG295733 to D.P.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. Nature 526:68–74.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Anagnostou P, Dominici V, Battaggia C, Pagani L, Vilar M, Wells RS, Pettener D, Sarno S, Boattini A, Francalacci P, et al. 2017. Overcoming the dichotomy between open and isolated populations using genomic data from a large European dataset. Scientific Reports 7:41614.

Arias L, Schröder R, Hübner A, Barreto G, Stoneking M, Pakendorf B. 2018. Cultural Innovations Influence Patterns of Genetic Diversity in Northwestern Amazonia. Mol. Biol. Evol.

Atzmon G, Hao L, Pe'er I, Velez C, Pearlman A, Palamara PF, Morrow B, Friedman E, Oddoux C, Burns E, et al. 2010. Abraham's Children in the Genome Era: Major Jewish Diaspora Populations Comprise Distinct Genetic Clusters with Shared Middle Eastern Ancestry. Am J Hum Genet 86:850–859.

Barbieri C, Heggarty P, Castrì L, Luiselli D, Pettener D. 2011. Mitochondrial DNA variability in the Titicaca basin: Matches and mismatches with linguistics and ethnohistory. Am. J. Hum. Biol. 23:89–99.

Barbieri C, Heggarty P, Yang Yao D, Ferri G, De Fanti S, Sarno S, Ciani G, Boattini A, Luiselli D, Pettener D. 2014. Between Andes and Amazon: the genetic profile of the Arawak-speaking Yanesha. Am. J. Phys. Anthropol. 155:600–609.

Battaglia V, Grugni V, Perego UA, Angerhofer N, Gomez-Palmieri JE, Woodward SR, Achilli A, Myres N, Torroni A, Semino O. 2013. The first peopling of South America: new evidence from Y-chromosome haplogroup Q. PLoS ONE 8:e71390.

Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics 32:2817-2823.

Bigham A, Bauchet M, Pinto D, Mao X, Akey JM, Mei R, Scherer SW, Julian CG, Wilson MJ, López Herráez D, et al. 2010. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. 6:e1001116.

Brandini S, Bergamaschi P, Cerna MF, Gandini F, Bastaroli F, Bertolini E, Cereda C, Ferretti L, Gómez-Carballa A, Battaglia V, et al. 2018. The Paleo-Indian Entry into South America According to Mitogenomes. Mol Biol Evol 35:299–311.

Browning BL, Browning SR. 2011. A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88:173–182.

Crawford JE, Amaru R, Song J, Julian CG, Racimo F, Cheng JY, Guo X, Yao J, Ambale-Venkatesh B, Lima JA, et al. 2017. Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. Am. J. Hum. Genet. 101:752–767.

D'Altroy TN. 2014. The Incas. New York: Wiley-Blackwell.

Delaneau O, Zagury J-F, Marchini J. 2013. Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods 10:5–6.

Di Corcia T, Sanchez Mellado C, Davila Francia TJ, Ferri G, Sarno S, Luiselli D, Rickards O. 2017. East of the Andes: The genetic profile of the Peruvian Amazon populations. Am. J. Phys. Anthropol. 163:328–

338.

Dillehay TD. 2009. Probing deeper into first American studies. Proc. Natl. Acad. Sci. U.S.A. 106:971–978.

Dillehay TD, Collins MB. 1988. Early cultural evidence from Monte Verde in Chile. Nature 332:150–152.

Dillehay TD, Ramírez C, Pino M, Collins MB, Rossen J, Pino-Navarro JD. 2008. Monte Verde: seaweed, food, medicine, and the peopling of South America. Science 320:784–786.

Fagundes NJR, Kanitz R, Eckert R, Valls ACS, Bogo MR, Salzano FM, Smith DG, Silva WA, Zago MA, Ribeiro-dos-Santos AK, et al. 2008. Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am. J. Hum. Genet. 82:583–592.

Fraley C, Raftery AE. 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association 97:611–631.

Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, Pettener D. 2003. Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. Mol. Biol. Evol. 20:1682–1691.

Gómez-Carballa A, Pardo-Seco J, Brandini S, Achilli A, Perego UA, Coble MD, Diegoli TM, Álvarez-Iglesias V, Martinón-Torres F, Olivieri A, et al. 2018. The peopling of South America and the trans-Andean gene flow of the first settlers. Genome Res. 28:767–779.

Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, Kenny EE, Gignoux CR, Maples BK, Guiblet W, et al. 2013. Reconstructing Native American migrations from whole-genome and whole-exome data. PLoS Genet. 9:e1004023.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A Draft Sequence of the Neandertal Genome. Science 328:710–722.

Greenberg JH. 1987. Language in the Americas. Stanford: Stanford University Press.

Harris DN, Song W, Shetty AC, Levano KS, Cáceres O, Padilla C, Borda V, Tarazona D, Trujillo O, Sanchez C, et al. 2018. Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. Proc. Natl. Acad. Sci. U.S.A. 115:E6526–E6535.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. Science 343:747–751.

Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD, et al. 2015. Genomic Insights into the Ancestry and Demographic History of South America. PLoS Genet. 11:e1005602.

Kehdy FSG, Gouveia MH, Machado M, Magalhães WCS, Horimoto AR, Horta BL, Moreira RG, Leal TP, Scliar MO, Soares-Souza GB, et al. 2015. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proc. Natl. Acad. Sci. U.S.A. 112:8696–8701.

Kitchen A, Miyamoto MM, Mulligan CJ. 2008. A three-stage colonization model for the peopling of the Americas. PLoS ONE 3:e1596.

Komsta L. 2011. outliers: Tests for outliers. Available from: https://CRAN.R-project.org/package=outliers

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. PLoS Genet. 8:e1002453.

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Wellcome Trust Case Control Consortium 2, et al. 2015. The fine-scale genetic structure of the British

population. Nature 519:309–314.

Lewis CM, Tito RY, Lizárraga B, Stone AC. 2005. Land, language, and loci: mtDNA in Native Americans and the genetic history of Peru. Am. J. Phys. Anthropol. 127:351–360.

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319:1100–1104.

Lindo J, Achilli A, Perego UA, Archer D, Valdiosera C, Petzelt B, Mitchell J, Worl R, Dixon EJ, Fifield TE, et al. 2017. Ancient individuals from the North American Northwest Coast reveal 10,000 years of regional genetic continuity. Proc. Natl. Acad. Sci. U.S.A. 114:4093–4098.

Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, Cybulski JS, Willerslev E, DeGiorgio M, Malhi RS. 2016. A time transect of exomes from a Native American population before and after European contact. Nat Commun 7:13175.

Lindo J, Rogers M, Mallott EK, Petzelt B, Mitchell J, Archer D, Cybulski JS, Malhi RS, DeGiorgio M. 2018a. Patterns of Genetic Coding Variation in a Native American Population before and after European Contact. Am. J. Hum. Genet. 102:806–815.

Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, Watson JT, Llave CV, Witonsky D, Beall C, et al. 2018b. The genetic prehistory of the Andean highlands 7000 years BP though European contact. Science Advances 4:eaau4921.

Llamas B, Fehren-Schmitz L, Valverde G, Soubrier J, Mallick S, Rohland N, Nordenfelt S, Valdiosera C, Richards SM, Rohrlach A, et al. 2016. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. Sci Adv 2:e1501385.

Luiselli D, Simoni L, Tarazona-Santos E, Pastor S, Pettener D. 2000. Genetic structure of Quechua-speakers of the Central Andes and geographic patterns of gene frequencies in South Amerindian populations. Am. J. Phys. Anthropol. 113:5–17.

Montinaro F, Busby GBJ, Pascali VL, Myers S, Hellenthal G, Capelli C. 2015. Unravelling the hidden ancestry of American admixed populations. Nat Commun 6:6596.

Moreno-Estrada A, Gignoux CR, Fernández-López JC, Zakharia F, Sikora M, Contreras AV, Acuña-Alonzo V, Sandoval K, Eng C, Romero-Hidalgo S, et al. 2014. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science 344:1280–1285.

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martínez RJ, Hedges DJ, Morris RW, et al. 2013. Reconstructing the population genetic history of the Caribbean. PLoS Genet. 9:e1003925.

Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas A-S, Sikora M, et al. 2018. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. Nature 553:203–207.

Moreno-Mayar JV, Vinner L, Damgaard P de B, Fuente C de la, Chan J, Spence JP, Allentoft ME, Vimala T, Racimo F, Pinotti T, et al. 2018. Early human dispersals within the Americas. Science:eaav2621.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient Admixture in Human History. Genetics 192:1065–1093.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2:e190.

Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012. Genomic patterns of

homozygosity in worldwide human populations. Am. J. Hum. Genet. 91:275–292.

Perego UA, Angerhofer N, Pala M, Olivieri A, Lancioni H, Hooshiar Kashani B, Carossa V, Ekins JE, Gómez-Carballa A, Huber G, et al. 2010. The initial peopling of the Americas: a growing number of founding mitochondrial genomes from Beringia. Genome Res. 20:1174–1179.

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

Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, Rohland N, Nägele K, Adamski N, Bertolini E, et al. 2018. Reconstructing the Deep Population History of Central and South America. Cell 175:1185-1197.e22.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Orlando L, Metspalu E, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 505:87–91.

Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A-S, et al. 2015. POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science 349:aab3884.

Ralph P, Coop G. 2013. The geography of recent genetic ancestry across Europe. PLoS Biol. 11:e1001555.

Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW, Rasmussen S, Moltke I, Albrechtsen A, Doyle SM, et al. 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature 506:225–229.

Rasmussen M, Sikora M, Albrechtsen A, Korneliussen TS, Moreno-Mayar JV, Poznik GD, Zollikofer CPE, de León MP, Allentoft ME, Moltke I, et al. 2015. The ancestry and affiliations of Kennewick Man. Nature 523:455–458.

Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, et al. 2012. Reconstructing Native American population history. Nature 488:370–374.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461:489–494.

Scheib CL, Li H, Desai T, Link V, Kendall C, Dewar G, Griffith PW, Mörseburg A, Johnson JR, Potter A, et al. 2018. Ancient human parallel lineages within North America contributed to a coastal expansion. Science 360:1024–1027.

Schroeder H, Sikora M, Gopalakrishnan S, Cassidy LM, Maisano Delser P, Sandoval Velasco M, Schraiber JG, Rasmussen S, Homburger JR, Ávila-Arcos MC, et al. 2018. Origins and genetic legacies of the Caribbean Taino. Proc. Natl. Acad. Sci. U.S.A. 115:2341–2346.

Schurr TG, Sherry ST. 2004. Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence. Am. J. Hum. Biol. 16:420–439.

Sevini F, Yao DY, Lomartire L, Barbieri A, Vianello D, Ferri G, Moretti E, Dasso MC, Garagnani P, Pettener D, et al. 2013. Analysis of population substructure in two sympatric populations of Gran Chaco, Argentina. PLoS ONE 8:e64054.

Silverman H, Isbell WH eds. 2008. The Handbook of South American Archaeology. New York, NY: Springer

New York Available from: http://link.springer.com/10.1007/978-0-387-74907-5

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D. 2015. Genetic evidence for two founding populations of the Americas. Nature 525:104–108.

Skoglund P, Reich D. 2016. A genomic view of the peopling of the Americas. Curr. Opin. Genet. Dev. 41:27–35.

Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, Mulligan CJ, Bravi CM, Rickards O, Martinez-Labarga C, Khusnutdinova EK, et al. 2007. Beringian standstill and spread of Native American founders. PLoS ONE 2:e829.

Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Labarga CM, Rickards O, Tyler-Smith C, Pena SD, Santos FR. 2001. Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. Am. J. Hum. Genet. 68:1485–1496.

Verdu P, Pemberton TJ, Laurent R, Kemp BM, Gonzalez-Oliver A, Gorodezky C, Hughes CE, Shattuck MR, Petzelt B, Mitchell J, et al. 2014. Patterns of admixture and population structure in native populations of Northwest North America. PLoS Genet. 10:e1004530.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, et al. 2007. Genetic variation and population structure in native Americans. PLoS Genet. 3:e185.

Figures

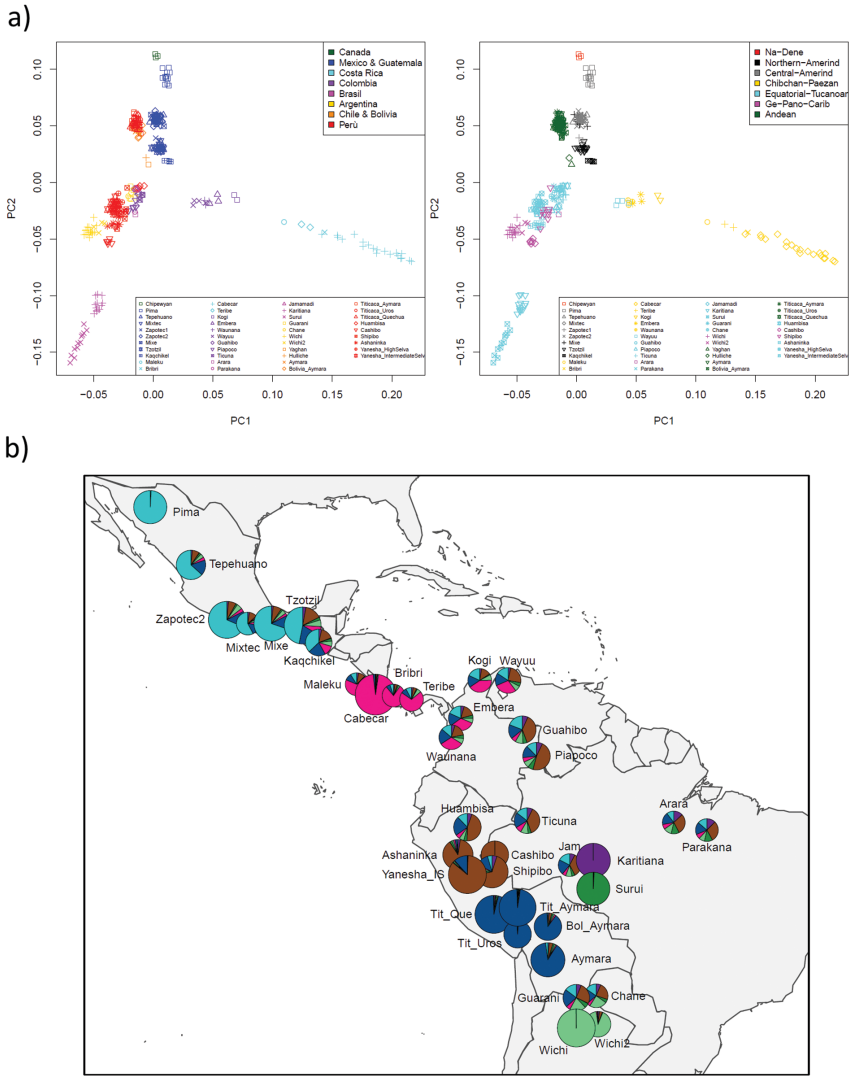

FIG. 1. Principal component and ADMIXTURE analyses performed on Native American populations included in the pruned "un-admixed" dataset. **a)** Plot of PC1 vs PC2 for the 43 un-admixed Native American groups reported in the bottom legends of left and right plots. Individuals are color-coded according to their country of origin (left) or language family affiliation (right). In order to allow continent-wide comparison, we used the same Greenberg's classification (1987) of languages as in Reich et al. (2012) **b)** Results of ADMIXTURE unsupervised cluster-based analysis at K = 8. Average proportions of inferred ancestral components are plotted at population-level. Pie charts diameters are proportional to the sample sizes of each considered group ranging from N = 1 to N = 20 (full set of populations, K tested and CV-errors are reported in **supplementary fig. S3 and S4, Supplementary Material online**). The geographical map has been plotted using the R software [v.3.2.4]

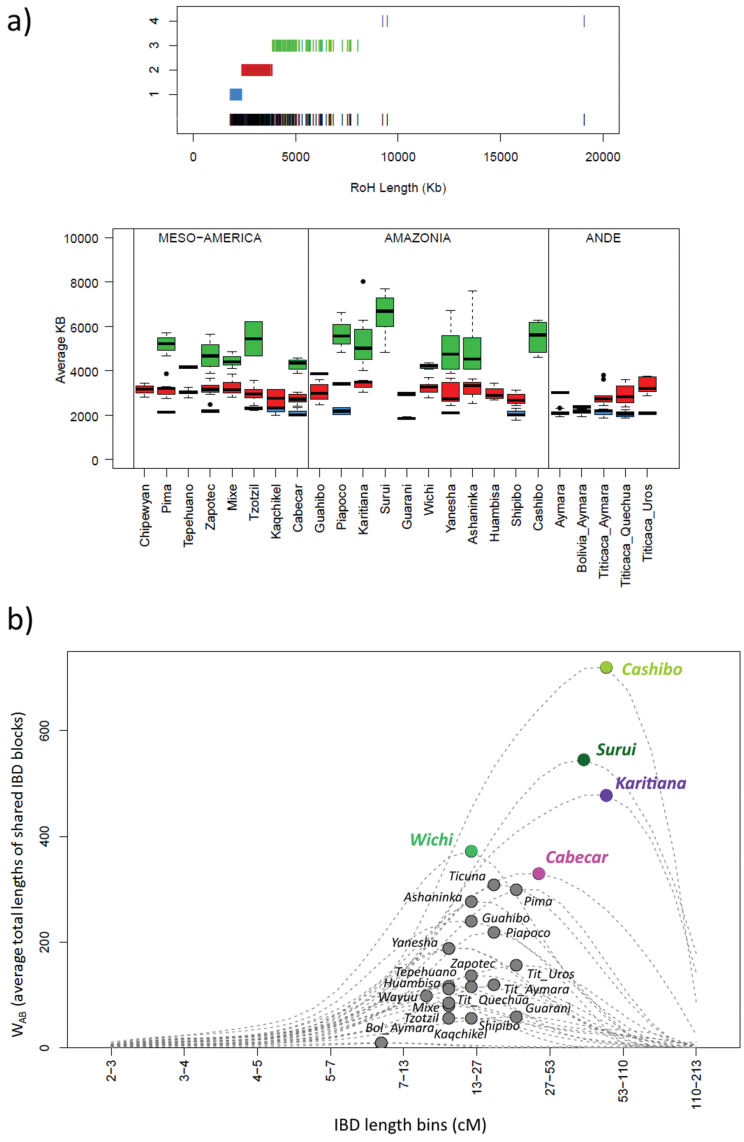

FIG. 2. Intra-population patterns of homozygosity and haplotype sharing. **a)** Runs of Homozygosity (ROH) calculated for the Native American groups with N ≥ 5 included in the "un-admixed" dataset. Top panel shows the distribution of all ROH lengths (black) and their inferred assignment into four bin classes identified by *Mclust* (blue, red, green, purple for class 1 to 4, respectively). Since, class 4 is represented by only three outlier individuals, they were removed from downstream analyses. Bottom panel shows the average length of ROHs over all individuals within each Native American population for each of three considered length classes (i.e. 1 to 3). **b)** Pattern of intra-population haplotype sharing measured as the average total length of genome shared IBD between every couple of samples within each population (W_{AB}). Within-population IBD-sharing was calculated for nine bins of IBD lengths, corresponding to different degrees of relatedness according to Moreno-Estrada et al. (2014). Dashes lines represent the distribution of inferred statistics over the considered length classes (see also **supplementary fig. S8**). The mode of the distribution is plotted as the corresponding labelled point for each population.

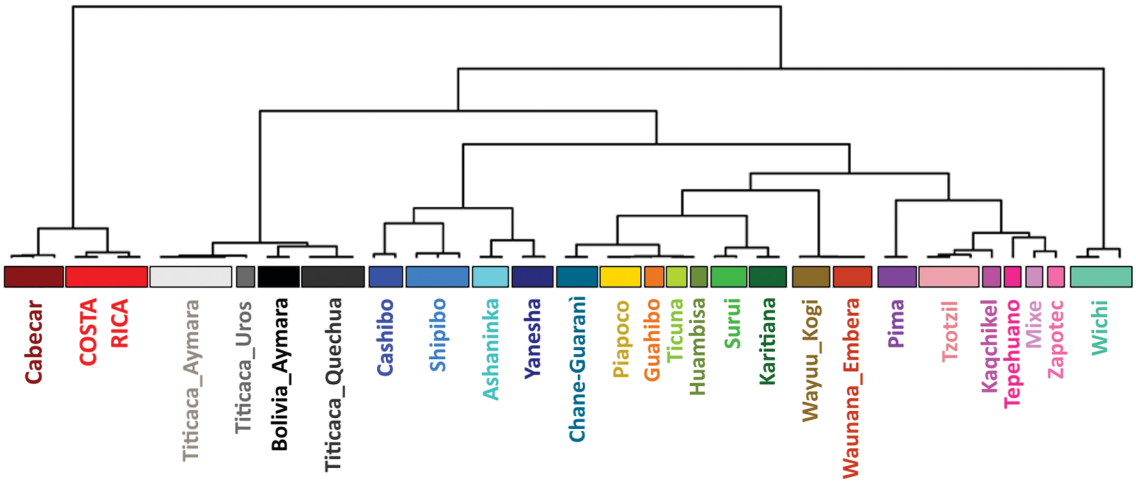

FIG. 3. fineSTRUCTURE hierarchical clustering dendrogram calculated between pairs of Native American individuals of the "un-admixed" dataset. The 26 clusters highlighted with different colors are highly concordant with the actual population labels, with the exclusions of partially overlapping geographically close groups of Costa Rica (i.e. Maleku, Bribri, Teribe and some Cabecar samples), Chane and Guarani, Wayuu and Kogi, Waunana and Embera, Aymara from Bolivia. In the figure, these samples were thus merged in the same cluster. For detailed annotation of individuals inside each cluster see **supplementary fig. S10, Supplementary Material online.**

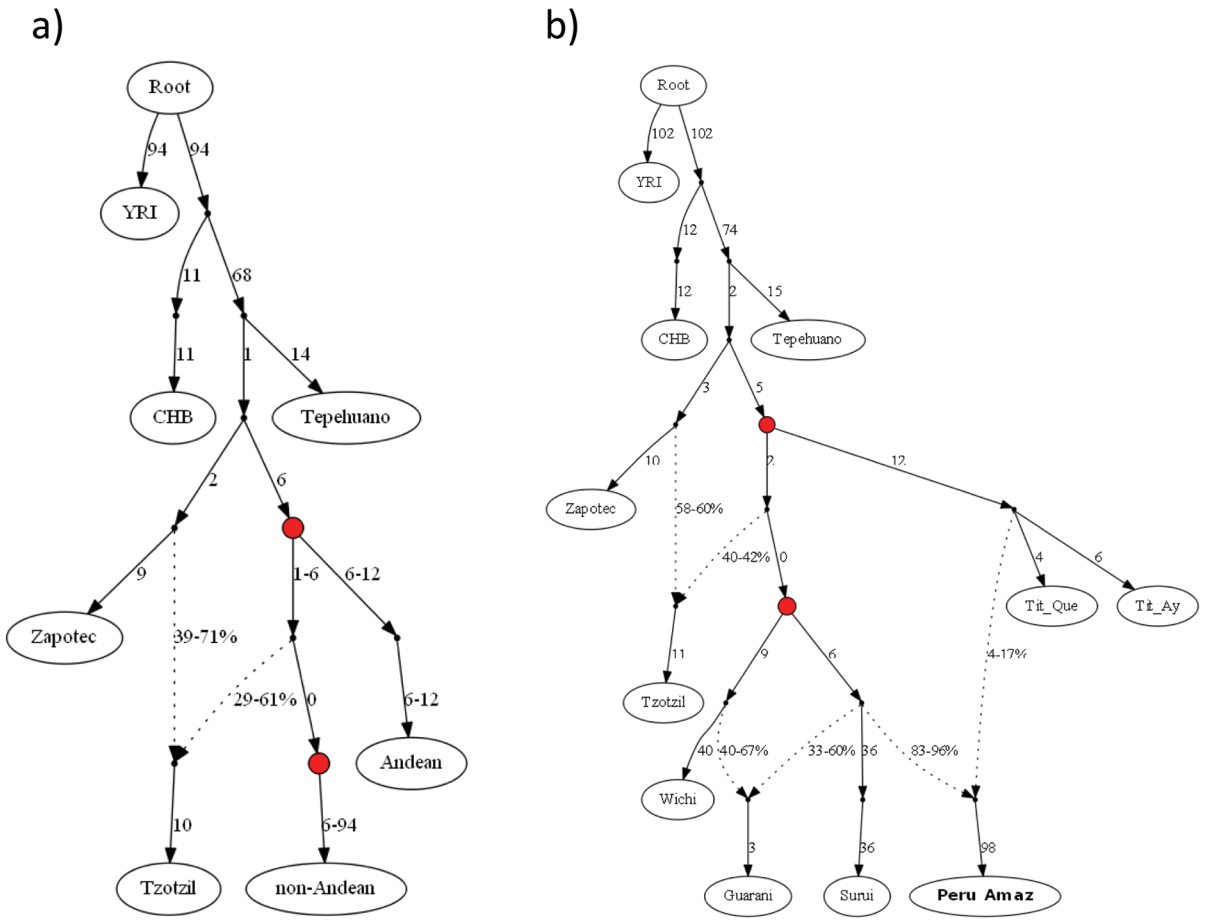

FIG. 4. Best-fitting admixture graphs (AG) obtained with *qpGraph*. **a)** Schematic summary of models testing a topology where the Tzotzils descend from an admixture between a node ancestral to the Zapotec and the non-Andean lineage, while using all possible combinations of Andean and non-Andean populations. **b)** Schematic summary of all AGs obtained testing in turn four Peruvian Amazonian groups (i.e. Cashibo, Shipibo, Yanesha and Ashaninka) as admixed between a non-Andean, specifically Amazonian, lineage and a node ancestral to the Andeans. Dotted lines represent the two-way admixture events tested and the percentages of ancestry on each line denote the proportions of admixture relative to the two admixing lineages. Units along solid lines indicate the measure of drift. Ranges of admixture proportions and drift lengths represent the min and max values reported in **supplementary table S11 and S13, Supplementary Material online** for all of the tests performed. Red nodes represent the two possible events of diffusion into South America hypothesized in the Discussion section.