

1 Fine-scale genomic analyses of admixed individuals reveal 2 unrecognized genetic ancestry components in Argentina

3 Native American, African and European genetic ancestries in Argentina

4 Pierre Luisi^{1,*}, Angelina García^{1,2,3}, Juan Manuel Berros⁴, Josefina M. B. Motti⁵, Darío
5 Demarchi^{1,2,3}, Emma Alfaro^{6,7}, Eliana Aquilano⁸, Carina Argüelles^{9,10}, Sergio
6 Avena^{11,12}, Graciela Bailliet⁸, Julieta Beltramo^{8,13}, Claudio M. Bravi⁸, Mariela Cuello⁸,
7 Cristina Dejean^{11,12}, José Edgardo Dipierri⁷, Laura S. Jurado Medina⁸, José Luis
8 Lanata¹⁴, Marina Muzzio⁸, María Laura Parolin¹⁵, Maia Pauro^{1,2,3}, Paula B. Paz
9 Sepúlveda⁸, Daniela Rodríguez Golpe⁸, María Rita Santos⁸, Marisol Schwab⁸, Natalia
10 Silvero⁸, Jeremias Zubrzycki¹⁶, Virginia Ramallo¹⁷, Hernán Dopazo^{18,19,*}

11

12

13 **1.** Universidad Nacional de Córdoba, Facultad de Filosofía y Humanidades,
14 Departamento de Antropología, Argentina

15

16 **2.** Instituto de Antropología de Córdoba (IDACOR), Consejo Nacional de
17 Investigaciones Científicas y Técnicas – Universidad Nacional de Córdoba, Córdoba,
18 Argentina

19

20 **3.** Universidad Nacional de Córdoba, Facultad de Filosofía y Humanidades, Museo de
21 Antropología, Córdoba, Argentina.

22

23 **4.** Consejo Nacional de Investigaciones Científicas y Técnicas – Laboratorio de
24 Análisis de Datos, Biocódices S.A, Buenos Aires, Argentina

25

26 **5.** Núcleo de Estudios Interdisciplinarios de Poblaciones Humanas de Patagonia
27 Austral (NEIPHA), Consejo Nacional de Investigaciones Científicas y Técnicas –
28 Universidad Nacional del Centro de la Provincia de Buenos Aires, Quequén,
29 Argentina.

30

31 **6.** Instituto de Ecorregiones Andinas (INECOA), Consejo Nacional de Investigaciones
32 Científicas y Técnicas – Universidad Nacional de Jujuy, Argentina

33

34 **7.** Instituto de Biología de la Altura, Universidad Nacional de Jujuy, Argentina

35

36 **8.** Instituto Multidisciplinario de Biología Celular (IMBICE), Consejo Nacional de
37 Investigaciones Científicas y Técnicas – Comisión de Investigaciones Científicas –
38 Universidad Nacional de La Plata, Argentina

39

40 **9.** Grupo de Investigación en Genética Aplicada (GIGA), Facultad de Ciencias Exactas,
41 Químicas y Naturales, Departamento de Genética, Instituto de Biología Subtropical

42 (IBS) - Nodo Posadas, Universidad Nacional de Misiones (UNaM) – Consejo Nacional
43 de Investigaciones Científicas y Técnicas, Posadas, Argentina.

44

45 **10.** Cátedra de Biología Molecular. Carrera de Medicina. Facultad de Ciencias de la
46 Salud. Universidad Católica de las Misiones (UCAMI), Posadas, Argentina.

47

48 **11.** Instituto de Ciencias Antropológicas (ICA), Facultad de Filosofía y Letras,
49 Universidad de Buenos Aires, Argentina

50

51 **12.** Centro de Estudios Biomédicos, Biotecnológicos, Ambientales y Diagnóstico
52 (CEBBAD), Universidad Maimónides, Buenos Aires, Argentina,

53

54 **13.** Laboratorio de Análisis Comparativo de ADN, Corte Suprema de Justicia de la
55 Provincia de Buenos Aires, La Plata, Argentina

56

57 **14.** Instituto de Investigaciones en Diversidad Cultural y Procesos de Cambio
58 (IIDyPCa), Consejo Nacional de Investigaciones Científicas y Técnicas – Universidad
59 Nacional de Río Negro, San Carlos de Bariloche, Argentina

60

61 **15.** Instituto de Diversidad y Evolución Austral (IDEAus), Consejo Nacional de
62 Investigaciones Científicas y Técnicas – Centro Nacional Patagónico, Puerto Madryn,
63 Argentina

64

65 **16.** Laboratorio de Genómica. Biocódices S.A., Buenos Aires, Argentina

66

67 **17.** Instituto Patagónico de Ciencias Sociales y Humanas (IPCSH) – Consejo Nacional
68 de Investigaciones Científicas y Técnicas - Centro Nacional Patagónico, Puerto
69 Madryn, Argentina

70

71 **18.** Departamento de Ecología, Genética y Evolución, Facultad de Ciencias Exactas y
72 Naturales, Universidad de Buenos Aires, Argentina

73

74 **19.** Consejo Nacional de Investigaciones Científicas y Técnicas – Biocódices S.A.,
75 Buenos Aires, Argentina

76

77 * Corresponding Authors:

78 Emails: pierresp@gmail.com (PL), dopazoh@gmail.com (HD)

79

80 **Abstract**

81 Similarly to other populations across the Americas, Argentinean populations trace
82 back their genetic ancestry into African, European and Native American ancestors,
83 reflecting a complex demographic history with multiple migration and admixture
84 events in pre- and post-colonial times. However, little is known about the sub-
85 continental origins of these three main ancestries. We present new high-throughput
86 genotyping data for 87 admixed individuals across Argentina. This data was
87 combined to previously published data for admixed individuals in the region and
88 then compared to different reference panels specifically built to perform population
89 structure analyses at a sub-continental level. Concerning the Native American
90 ancestry, we could identify four Native American components segregating in
91 modern Argentinean populations. Three of them are also found in modern South
92 American populations and are specifically represented in Central Andes, Central
93 Chile/Patagonia, and Subtropical and Tropical Forests geographic areas. The fourth
94 component might be specific to the Central Western region of Argentina, and it is
95 not well represented in any genomic data from the literature. As for the European
96 and African ancestries, we confirmed previous results about origins from Southern
97 Europe, Western and Central Western Africa, and we provide evidences for the
98 presence of Northern European and Eastern African ancestries.

99 **Author Summary**

100 The human genetic diversity in Argentina reflects demographic processes during
 101 which the European colonists invaded a territory where Native American
 102 populations were settled. During the colonial period, the slave trade also prompted
 103 many African people to move to Argentina. Little is known about the origins of these
 104 three continental ancestry components in Argentinean populations nowadays.
 105 Genotyping data for 87 admixed individuals throughout Argentina was generated
 106 and data from the literature was re-analyzed to shed light on this question. We
 107 confirmed that most of the European genetic ancestry comes from the South,
 108 although several individuals are related to Northern Europeans. We confirmed that
 109 the African origins in Argentina mainly trace back from Western and
 110 Central/Western regions, and we document some proportion of Eastern African
 111 origins poorly described before. As for the Native American ancestry, we identified
 112 that it can be divided into four main components that correspond to Central
 113 Chile/Patagonia, Subtropical and Tropical Forests, Central Andes and Central
 114 Western region of Argentina. In order to understand the specificity of the genetic
 115 diversity in Argentina, more effort is required to generate specific massive genomic
 116 knowledge at the local level.

117 **Introduction**

118 The first systematic investigation of human genetic variation in Argentina focused
 119 on a limited number of markers either uniparental (mtDNA, Y-STRs, Y-SNP; [1–10]

or autosomal (Short Tandem Repeats, Ancestry Informative Markers, Alu sequences, indels, and blood groups [11–16]). Studies based on autosomal markers identified an important inter-individual heterogeneity for the African, Native American and European genetic ancestry proportions [11–16]. Accordingly, most of the studies based on uniparental markers showed large differences in the genetic composition of Argentinean populations, accounting for the different demographic histories within the country [1,2,17–20,3–9,16]. Although the idea of a 'white' country with most of inhabitant's descendants from European immigrants has now been rejected by these studies, the Argentine founding myth of a white and European nation remains perceptible today [21].

There is a wealth of information about the European ancestors, both from familial stories and from historical records. On the contrary, little is known about the African and Native American populations that contributed to the admixture events between the continental components. The great wave of European immigration occurred between the mid-19th to mid-20th centuries. Although immigrants came from all over the continent, historical records attest that immigration waves from Southern Europe (mainly Italy and Spain) were predominant [22].

As for the African genetic origins in Argentina, historical records about the arrival of African slaves to Río de la Plata show that Luanda, in current Angola territory in Central Western region of Africa, was the main departure harbor, followed by ports located in the Gulf of Guinean and on the coast of the present-day Senegalese, Gambian and Sierra Leone territories. The Indian Ocean coasts, in the current territory of Mozambique, were also an important departure location [23–25].

Although the departure harbor is a poor proxy to infer the actual slaves' origins [25,26], the African uniparental lineages observed in Argentina are consistent with the historical records [27,28].

As for the Native American component, it is difficult to study its origin focusing on present-day communities since their organization has changed drastically after the arrival of the first conquerors in the 16th century [21]. During the period of conquest and colonization, wars, diseases and forced labor decimated the Native populations. The system of colonial exploitation also often meant the relocation of individuals, families, and communities [29]. Then, the expansion of the nation-state by the late 19th Century can be described as a territorial annexation process and subjugation of the indigenous peoples perpetrated by the Argentinean national armed forces between 1876 and 1917 [30].

Due to the specificity of the Argentinean demographic history, a remaining challenge is to unravel which populations from each continent contributed to the genetic pool in nowadays Argentinean populations leveraging genotype data for hundreds of thousands of autosomal markers from the whole genome. Recently, two articles presented high-throughput genotyping data for modern Argentine individuals, and provided the first insights to decipher which populations from each continent contributed to the genetic pool in nowadays Argentinean populations [31,32]. In both studies, it was found that the European ancestry in Argentina is mainly explained by Italian and Iberian ancestry components. Homburger *et al.* observed a strong gradient in Native American ancestry of South American Latinos between Andean and other South American Native American populations [31]. In addition,

Muzzio *et al.* found that African ancestry is explained by a Central Western (Bantu-influenced) component and a Western component.

Despite these progresses, the efforts to understand the demographic history that shaped genetic diversity in Argentina have inevitably been scarce, particularly for African and Native American components. Studies of a wider region, namely South America can provide important insights. A previous study showed that Western and Central Western African ancestries are common across the Americas, particularly in Northern latitudes, while the influence of South/Eastern African ancestry is greater in South America [33]. In another study in Brazil, two African ancestries have been observed: a Western African one and another associated with Central East African and Bantu populations, the latter being more present in the Southeastern and Southern regions [34]. Ancient DNA analyses of the peopling of the Americas suggest that Native American populations of South America descend from two streams from Northern America: one mainly present in the Andes and another one present elsewhere [35]. These streams have replaced the first people settled in the region, whose ancestry was related to the Clovis culture [35]. Genetic continuity for the Native American component appears to have prevailed in the region ever since these replacements [35–37]. However, little is known about the legacy in modern populations of the different ancestries associated to the several waves of population arrivals in South America. Modern Native American populations in the Southern Cone of South America seems to be divided between a component that includes populations in Tropical and Subtropical Forests and another component that includes Andean, Central Chile and Patagonian populations [38]. A recent genomic

study of ancient and modern populations from Central Chile and Western Patagonia further identified that they are differentiated from the Andean and Subtropical and Tropical Forests populations [37]. In another genomic study of modern samples (in which the Southern Cone is only represented by the Gran Chaco region), it has been found that all non-Andean South American populations are likely to share a common lineage, while they are unlikely to share with the Andeans any common ancestor from Central America [39], supporting the hypothesis of many back migrations to Central America from non-Andean South American populations [39]. It has also been described that the genetic interactions between the peopling routes on both sides of the Andes were limited [39,40].

Here, we carry out a fine-scale population genomic study to get insight into the genetic structure and the complex origins of the African, European and Native American ancestries in Argentina populations. Using Affymetrix, Axiom-Lat1 array technology, we genotyped 87 admixed individuals nationwide for more than 800,000 SNPs covering coding and non-coding regions of the genome. Additionally, a dataset with ~500 individuals from modern and ancient populations throughout South America was constructed. We also pulled together genotype data for Sub-Saharan African and European populations from the literature. The new data generated in the present study, compared to those data sets, allows to broaden the knowledge of the sub-continental origins of the three main genetic ancestries of Argentinean populations.

Results and Discussion

Studied Populations

High-throughput genotyping data was generated for 87 admixed individuals throughout Argentina (**Fig 1** and **S1 Table**). For clarity in the visualization of the results, we used provinces and regions to classify the admixed individuals analyzed as shown in **Fig 1**. This classification has not been used for any statistical analyses. The generated data was compared to different data sets to understand the origins of the genetic diversity in the country. These data sets are called **DS<n>** and are described in **S2 Table**.

Fig 1. Sample locations from the present study. The samples are divided according to regions for visualization only. N: The number of samples that passed genotyping Quality Controls.

Ancestry in a worldwide context

In order to characterize the genetic diversity observed in Argentina within a worldwide genetic context, Principal Component Analysis (PCA [41]; **S1 Fig**) was applied to the dataset **DS1**, which contains genotype data for the samples from Argentina generated in the present study, as well as other admixed and Native American individuals from South America [31,37,38,42] and individuals from Europe and Africa from the 1000 Genomes Project (1KGP; [42]). The PCA shows that Argentinean individuals have different proportions of European, Native American and, much less represented, African ancestry (**S1A Fig**). This pattern, which has already been documented in other admixed populations from South America

[31,42], including in Argentina [13–16,32] supports the heterogeneous genetic origins throughout the country. In addition, PC4 discriminates between Southern and Northern European individuals (**S1B Fig**) while PC6 separates Luhya population, a Bantu-speaking population in Kenya, from Western African populations (**S1C Fig**). The Argentinean samples do not exhibit any gradient along these two PCs. But, PC3 and PC5 discriminate three different main Native American ancestries that are also represented in the Argentinean samples.

We ran unsupervised clustering models with Admixture software [43] on **DS1**. We used from 3 to 12 putative ancestral populations. Comparing the cross-validation scores obtained for each run, we estimated that the genotype data analyzed was best explained with a model with $K=8$ ancestral populations (**S2A Fig**). At $K=3$ (**S2B Fig** and **S3 Table**), the algorithm allows estimating the proportions of European, Native American and African ancestry. For $K=4$ to $K=7$, (**S2C-F Figs**), the model detects sub-continental ancestries. At $K=8$ (**Fig 2; S3 Table**), the European ancestry is divided into Northern and Southern components (dark and light blue, respectively), while African ancestry is composed of Westernmost African (green), Gulf of Guinea (light green) and Bantu-influenced (dark green) components. Moreover, three components are observed for Native American ancestry: one in Central Chile/Patagonia populations (in orange; hereafter referred to as **CCP**), another in Central Andes populations (in yellow; hereafter referred to as **CAN**) and a third one in populations from Tropical and Subtropical Forests (in pink; here after referred to as **STF**).

Fig 2. Admixture Analyses in a Worldwide Context.

Admixture for $K=8$. **1st row:** 1000 Genomes Samples from Europe, Africa and South America [42]. **2nd row:** Modern Native American samples grouped following [38]. **3rd row:** Chilean, Peruvian and Argentinean admixed samples from [31]. **4th row:** Argentinean admixed samples from the present study. Samples are grouped according to regions (**Fig 1**); CYA: Cuyo Region in Argentina; NEA: Northeastern Region, NWA: Northwestern Region; PPA: Pampean Region; PTA: Patagonia Region. CLM: Colombians from Medellin; PEL: Peruvians from Lima.

We confirmed that the estimates of the continental ancestry proportion obtained from a model with $K=3$ and $K=8$ are highly consistent (**S3 Fig**). Moreover, the eigenvectors from PCA and the ancestry proportion estimates with Admixture are well correlated (**S4 Fig**).

From the Admixture results for $K=8$ (**Fig 2**), we observed that the European ancestry for Argentinean samples, is divided in Southern and Northern components, the former being the most abundant. The low proportion of African ancestry in Argentinean samples makes difficult to interpret its sub-continental origins from analyses within a global context. Surprisingly, all three components of Native American ancestry are present in most Argentinean samples (**Fig 2**). They exhibit mid proportions of different Native ancestries suggesting either the result of a mixture between these three ancestry components or an underrepresentation of Native American in the reference panels currently available. Such mixture pattern is not observed in other South American countries. Indeed, the Native American ancestry for Peruvian, Chilean and Colombian admixed samples is mainly represented by CAN, CCP and STF, respectively. This is consistent with the geographical area where the admixed individuals have been sampled, and the genetic ancestry of the indigenous communities from each country.

Sub-continental ancestry components in Argentina

To decipher the sub-continental origins of the European, African and Native American ancestries, we first estimated local ancestry patterns in phased **DS2** and **DS3 (DS2p and DS3p)** separately. Across phased chromosomes, we assigned whether a genomic region was of Native American, European or African ancestry using RFMix v2 [44]. We applied principal component and unsupervised clustering analyses on the masked data (**S5 Fig**) compiled with reference data sets describing the genetic diversity within each continent (**S2 Table**).

The genetic legacy of European migration in Argentina

We used **DS4**, a combination of the masked genotype data for admixed individuals with a set of European individuals carefully selected to be representative of the genetic diversity in their sampling area [45,46].

From the PCA, we observed that most Argentinean individuals cluster with Iberians and Italians (**Fig 3**), as previously described [31,32]. However, some individuals cluster with Central and Northern Europeans.

Fig 3. European Ancestry Specific Principal Component Analysis.

(A) Multidimensional scaling scatterplot (MDS) from Euclidian distances calculated from weighted 5-first European-specific Principal Components performed using the European reference samples, and admixed samples masked for African ancestry. Individuals from the European reference panel are colored according to main European geographic regions as shown in **(B)** while South American individuals are represented as shown in the legend. Elbow method to choose the best number of PCs to compute MDS is shown in **S6 Fig**.

304 We estimated that the genotype data analyzed was best explained with an
305 admixture model with $K=2$ (**S7A Fig**). Although it is rather difficult to assign a clear
306 label to each of these two ancestral populations, it seems that the algorithm
307 discriminates between Central/Northern (dark blue) and Southern Europe (light
308 blue) components (**S7B Fig**). For $K=3$, we still observe a Central/Northern European
309 component (dark blue), while the South can be divided into a component
310 particularly represented in Iberian individuals (turquoise) and another component
311 more frequent in Southeastern Europeans and Italians (light blue). The Iberian
312 component is highly represented in Southern American samples (**S7C Fig**), most
313 likely reflecting the legacy of first Iberian migrations in South America during
314 colonial period [11]. Samples from Argentina and Chile exhibit higher proportions of
315 Southeastern/Italian and Northern European ancestries than Colombians, as well as
316 lower Iberian ancestry proportions (**S8 Fig**). We observed no significant difference
317 in the proportions of any European ancestry between Argentinean and Chilean
318 samples (**S8 Fig**). However, both PCA and Admixture shows that the individuals
319 with most Southeastern/Italian ancestry are from Argentina, This is consistent with
320 a previous study [31], and it can be explained by the many arrivals from Italy during
321 the great wave of European immigration in the 19th and 20th centuries [22].

322 Moreover, Argentinean individuals with higher proportion of Central/Northern
323 European ancestry are from Misiones province (NEA), consistent with the historical
324 record of settlement of Polish, German, Danish and Swedish colonies promoted by
325 governmental or private enterprises in the province [47].

Different African ancestry components in Argentina

To investigate the sub-continental components that explains African ancestry in Argentina, we used **DS5** which combines masked genotype data from admixed individuals with a published data set of African individuals [42,48–50] (**S9A Fig**). The African reference populations used here can be divided into five main groups: Bantu-influenced, Hunter-Gatherers, Western African, San and Eastern African populations [48].

From the PCA applied to this data, we observed that African ancestry in Argentina has small genetic affinity with San and Hunter-Gatherers populations. Indeed, according to PC1 and PC2, Argentinean individuals are located within a group of African individuals belonging to Bantu-influenced, Western African and Eastern African populations, and clearly distinct from San and Hunter-Gatherer populations (**S9B Fig**). Moreover, on PC3, most Argentinean are closer to Western African and Bantu-influenced populations, while several Argentinean individuals, particularly from Northern Argentina, are located within a gradient towards Eastern African populations (**S9D Fig**). According to PC4, most of the Argentinean individuals are closer to the Bantu-influenced populations while others individuals are closer to Western Africa populations, and several other Argentinean individuals are found in a cline between these two groups (**S9C Fig**). These patterns were confirmed by applying Admixture algorithm. The Cross-Validation procedure points to a best fit of the data with $K=5$ (**S10A Fig**), a model for which we observe that the Bantu-influenced and Western Africa components are the most represented in Argentinean individuals. Moreover, this analysis also showed that some individuals exhibit

smaller, yet important, proportions of Eastern African ancestry, particularly in Northern Argentina (**Fig 4**). Although, the important missing genotype rate in masked data for admixed individuals could bias PCA and Admixture results, the results obtained by both methods are highly consistent for admixed individuals (**S11 Fig**).

Fig 4. African Ancestry-Specific Admixture Analysis.

Admixture for $K=5$. **1st row**: 1,685 reference samples with >99% of African ancestry. Populations are grouped following [48]. **2nd and 3rd rows**: Peruvian and Argentinean admixed samples from [31] and from the present study. Samples are grouped according to region. CYA: Cuyo Region; NEA: Northeastern Region, NWA: Northwestern Region; PPA: Pampean Region; PTA: Patagonia Region. Genotype data for admixed samples were masked for African ancestry.

The legacy of Western Africa on the African genetic diversity in the Americas has been preeminent [51–53], along with an impact of Bantu-influenced populations from Central Western Africa, particularly in Brazil and the Caribbean [51,52]. These two African ancestries have also been previously documented in Argentina from studies of autosomal [32] and maternal markers [27,28]. Maternal lineages specific to populations from Central Western Africa –particularly from Angola- are the most common African lineages in Argentina according to studies in the Central region [27] and in four urban centers (Puerto Madryn, Rosario, Resistencia and Salta) [28]. These results are concordant with the predominance of the Bantu-influenced origin identified in the present study. In addition, the presence of Southeastern Africa maternal lineages in Argentina [27,28] is consistent with African ancestry of this origin identified in previous studies in other South America countries [33,34], and with the Eastern African ancestry identified here.

An unrecognized genetic Native American component in Argentina

We compared masked genotype data for admixed individuals with masked genotype data of modern Native American reference populations from South America [37,38] and a set of ancient DNA data from the region [35–37,54,55] (**DS6**; **Fig 5A**). The PCA (**Fig 5B**) and Admixture analyses (best model with $K=3$) confirmed the three main South Native American ancestry components described by de la Fuente *et al.* (2018), and identified in our analyses at the global level.

Fig 5. Native American ancestry-specific principal component analysis. **(A)** Localization map of the studied samples. Color and point coding matches sample groups: ancient populations are grouped following [35], modern Native American populations are grouped following [38] while Argentinean admixed sample locations are grouped according to Regions (**Fig 1**). For clarity, some geographic coordinates have been slightly changed. **(B)** Multidimensional scaling scatterplot from Euclidian distance calculated from weighted Native American ancestry specific Principal Components. Elbow method to choose the best number of PCs to compute MDS is shown in **S12 Fig**.

Similar results were also observed when applying Admixture algorithm (best model with $K=3$; **S13A Fig**). The proportion estimates for the three main Native American ancestries observed for the South America Native American individuals and for the ancient samples (**Fig 6**) highly correlate with estimates from unmasked data (**S14 Fig**), and are consistent with CAN, CCP and STF labels that we attributed. In Argentina, the three Native American ancestries are observed in almost all the admixed individuals studied. In Northeastern Argentina (NEA), STF ancestry is the most frequent. In Northwestern Argentina (NWA), CAN ancestry is not clearly dominating since STF is also observed in important proportions. In the South, CCP is observed in greater proportions.

Fig 6. Native American Ancestry-Specific Admixture Analysis.
Admixture for $K=3$. **1st row:** Ancient samples grouped following [35]. **2nd row:** Modern Native American samples grouped following [38]. **3rd row:** Chilean, Peruvian and Argentinean admixed samples from [31]. **4th row:** Argentinean admixed samples from the present study; sample locations are grouped according to Regions (**Fig 1**); CYA: Cuyo Region; NEA: Northeastern Region, NWA: Northwestern Region; PPA: Pampean Region; PTA: Patagonia Region. Genotype data for all modern samples were masked for Native American ancestry..

Although migration events tend to reduce the genetic distance among these components, we still identified correlations between geographic coordinates and the proportions of each Native American ancestry component. Indeed, CAN is found in higher frequencies further North (**S15A Fig**) and West (**S15A Fig**), STF is higher further North (**S15B Fig**) and East (**S15E Fig**), while the proportion of CCP is higher further South (**S15C Fig**).

Many individuals from the Cuyo and Pampean regions of Argentina (San Juan and Córdoba provinces as well as South of Buenos Aires province) exhibit intermediate position in PCA (**Fig 5**) and mid proportion estimates with Admixture (**Fig 6**). This pattern can be interpreted as the result of a mixture between different ancestries (scenario 1) or relative limited shared history with any of them (scenario 2). In the intent to discriminate between both scenarios, we increased the number of ancestral populations to 4 and 5. Model with $K=4$ (**S13C Fig**) does not help to address this question, and model with $K=5$ informs that scenario 2 is the most likely for Middle Holocene samples from the Southern Cone (**S13D Fig**). Models including more ancestral populations ($K>5$) did not allow robust ancestry proportion estimates most likely due to the masked data leveraged here. Many regions of Argentina are underrepresented in the reference panel because of the scarcity of

Native American communities in most of the Argentinean territory. Altogether, PCA and admixture analyses are not sufficient to unequivocally contrast whether some Native American genetic diversity specific to Argentina is not represented in the reference panel.

In order to circumvent these limitations, we performed an objective quantitative approach based on *K*-means clustering (**S16 Fig**). We assigned the 452 modern South American individuals (the 53 ancient samples included in **DS6** were not analyzed here) to a Native American ancestry cluster. We assigned 163, 161 and 70 individuals to the clusters representing CAN, STF and CCP, respectively, and 32 individuals were assigned to a fourth cluster (**S4 Table**). The remaining 26 individuals were removed for further analyses because their group assignment was not consistent across the three clustering approaches. We acknowledge that these groups are culturally, ethnically and linguistically heterogeneous. However, we argue that analyzing such groups built from genetic similarities may provide interesting insights into evolutionary mechanisms that shaped the Native American ancestry in South America.

Most individuals from Calingasta, (located in the Northwest Monte and Thistle of the Prepuna ecoregion; San Juan Province) and from Santiago de Chile were assigned to the fourth group. The genealogical record for the Calingasta individuals attests to a local origin of their direct ancestors up to two generations ago, and they have mtDNA sub-haplogroups predominant in the Cuyo region (**S1 Table**; [56,57]). On the other hand, all the Huilliche and Pehuenche individuals from Central Chile [37]

have been consistently assigned to CCP. Altogether, we decided to refer to this fourth group hereafter as Central Western Argentina hereafter (CWA).

Relationship among the four identified Native American components

Significant positive f_3 scores to test for Treeness were obtained for the six possible pairwise comparisons among the four Native American clusters identified (**Fig 7A** and **S5A Table**). In addition to confirming the differentiation among the three components described in our reference panel (CCP, CAN and STF), this analysis showed that the fourth group (CWA) most likely represents a Native American component never described in any previous study based on autosomal genetic markers. In fact, the genetic differentiation (measured with F_{ST} index) between CWA and any other cluster is not lower than for other comparisons (**Fig 7B**). The lowest F_{ST} was observed between STF and CAN, probably due to the fact that STF encompasses the Northern Andes region (**Fig 7B**). Moreover, the distribution of $1 - f_3(YRI; Ind1, Ind2)$ between pairs of individuals from different groups (**S17 Fig**) is an additional argument to discard a scenario of mixture (mentioned before as scenario 1).

Furthermore, f_4 analyses showed that (i) CAN has no particular genetic affinity with any component relative to the others; (ii) STF is closer to CAN as compared to CCP and CWA; and (iii) CWA and CCP exhibit higher genetic affinity between them than with CAN or STF (**Fig 7C; S5B Table**). However, a neighbor-joining analysis [58] from distances of the form $1/f_3(YRI; X, Y)$ suggests that CAN is more closely related to CCP and CWA than to STF (**Fig 7D; S5C Table**).

Fig 7. Relationship among the four Native American groups identified.
(A) $f_3(\text{Target}; S1, S2)$ to test for treeness. **(B)** F_{ST} between pairs of Native American groups. **(C)** $f_4(\text{YRI}, \text{Target}; S1, S2)$ to test whether Target shares more ancestry with $S1$ or $S2$. Since f_4 is symmetrical when switching $S1$ and $S2$, only positive comparisons are shown. **(D)** Neighbor-joining tree estimated from distance matrix of the form $1/f_3(\text{YRI}; X, Y)$. The tree was estimated using ancient sample from Upward Sun River site in Beringia (USR1) as outgroup. CAN: Central Andes; STF: Subtropical and Tropical Forests; CCP: Central Chile / Patagonia; CWA: Central Western Argentina; YRI: Yoruba from 1KGP.
 Vertical segments are the ± 3 standard errors intervals in A and B.

Based on these evidences, we argue that CWA may represent a Native American ancestry that diverged from CCP and established on the Eastern side of the Andes in the Cuyo region.

The existence of a specific differentiated component in the Central Western and Central regions of Argentina has been previously suggested from maternal lineages analyses accounting for the genetic relationship between these two regions [10] and the presence of specific clades underrepresented elsewhere [10,27,59]. These studies support the hypothesis of a common origin and/or important gene flow [57], and the authors referred to a meta-population with great temporal depth and differentiated from other regions in Argentina. Moreover, the ethnographic description of the populations that were settled at the moment of contact with the Spanish colonies accounts to a potential relationship between Huarpes in the Cuyo region and Comechingones in present-day Córdoba province ([60], cited by [61]). Altogether, these facts support the Central Western Argentina (**CWA**) label that we attributed to this fourth component.

Complementary historical facts may explain the representation of this component in Santiago de Chile and its absence in Huilliche and Pehuenche populations. First, in

pre-Columbian times, it seems to have existed an ethnic boundary around the 36° S Latitude, since Central Andean civilizations could not expand further South. [62,63]. Later, the Spanish could neither settle those territories. Second, the Spanish colonists that established in Santiago de Chile organized massive deportation of Huarpes individuals from Cuyo region to palliate the lack of indigenous workforce [64]. For example, in Santiago de Chile, in 1614, 37% of the indigenous people that lived in the suburbs were Huarpes according to the chronicler Vázquez de Espinosa [65].

The present study, in which we analyzed individuals that do not belong to any indigenous community, made possible the identification of a Native American component not previously reported from autosomal markers.

Genetic affinity of the four Native American ancestry components with ancient populations

We evaluated the genetic affinity of the four Native American ancestry components to ancient samples from the literature [35–37,54,55,66,67]. Graphical summaries of pairwise f_3 -based distances are presented in **S18 Fig**.

When comparing the genetic affinity of a given component with the different ancient groups using either the f_3 -outgroup or the f_4 statistics (**S19F Fig** and **S5D,E Tables**), we identified that CAN tends to exhibit greater genetic affinity with ancient Andean populations than with other ancient groups (**S19A,E Figs**). Strikingly, the genetic affinity of this component with both Late Andes and Early Andean ancient groups would point to a genetic continuity across the whole temporal transect that the archaeological samples provide. This would suppose that the replacement of an

early population arrival by a later stream of gene flow in Central Andes identified previously [35] does not fully explain the current-day genetic diversity for CAN. We observed higher genetic affinity with ancient Southern Cone groups for both CCP and CWA (**S19C,D,G,H Figs**). As for STF, we observed intermediate genetic affinity with ancient samples from both the Andes and the Southern Cone (**S19B,F Figs**). The fact that there is no ancient sample group exhibiting outstanding genetic affinity with STF points to the underrepresentation of this component in ancient samples. First, the geographical range covered by ancient samples that could represent this component is restricted to Brazil, while STF is a heterogeneous group that includes relatively isolated populations [39] from the Gran Chaco, the Amazonas and Northern Andes. Moreover, the most recent samples that could represent STF are aged ~6700BP, and gene flow with other components since then may have contributed to dissolve the genetic affinity of STF with ancient samples in Brazil analyzed here.

Then, we evaluated the relationship of the time depth of the ancient samples from either the Andes or the Southern Cone, with their genetic affinity to the modern components of Native American ancestry (**Fig 8**). We observed a statistically significant relationship between the age of the ancient Southern Cone samples and their genetic affinity with CCP and CWA. This means that the older the ancient sample from the Southern Cone, the lower the shared drift with CCP and CWA. On the other hand, no statistically significant relationship was identified for STF and CAN ($P = 0.523$ and $P=0.596$, respectively; **Fig 8A**). These patterns could be due to a relationship between geography and the age of the ancient samples because the

most recent samples are concentrated in the Southern tip of the subcontinent (**Fig 5A**). Moreover, the number of SNPs with genotype data tends to decrease with the age of the ancient samples due to DNA damage, and thus inducing a potential bias towards significant positive correlations. To simultaneously correct both these two putative confounding effects, we repeated the analyses but using a correction for the ancient sample age (the residuals of the linear regression between the age of the ancient samples and their geographic coordinates) and a correction for genetic affinity estimates (the residuals of the linear regression between f_3 and the number of SNPs to estimate it). This correction intensified the relationship described for CCP and CWA (**Fig 8C**). It also allowed to actually identifying significant relationships for STF and CAN. On the other hand, CAN is the only modern Native American component that exhibits a significant relationship between its genetic affinity with ancient Andean samples and their age (**Fig 8B**). This pattern holds after correction for geography (**Fig 8D**). Repeating the same analyses using f_4 statistics, we reached the same conclusions (**S20 Fig**).

Fig 8. Changes across time of genetic affinity of the four Native American with ancient samples.

Each point represents a f_3 -Outgroup score of the form $f_3(YRI; X, Ancient)$ vs the age of ancient samples, where X is one of the four identified modern Native American components, and *Ancient* is an ancient group. X is represented by the color of the square while the symbol inside the square represents *Ancient*. The legend for plotted symbols is shown in Fig 5. **(A)** f_3 vs age of ancient samples from Southern Cone. **(B)** f_3 vs age of ancient samples from the Andes. **(C)** f_3 vs age of ancient samples from Southern Cone considering correction for both f_3 and age. **(D)** f_3 vs age of ancient samples from the Andes considering correction for both f_3 and age. Linear regression slopes and the associated P -values are shown. CAN: Central Andes; STF: Subtropical and Tropical Forests; CCP: Central Chile / Patagonia; CWA: Central Western Argentina; YRI: Yoruba from 1KGP

Early divergence among the four Native American components in Argentina

Using another setting of the f_4 statistics (**S5F Table; S21 Fig**), we observed that all Southern Cone ancient samples – except Los Rieles ~10900 BP – are more similar to CCP than to CWA (**S21F Fig**), pointing to divergence time between those two components older than ~7700 BP. CWA is not closer to any Brazilian and Andean ancient sample as compared to CCP, reinforcing the idea that CWA is not a mix of CCP, CAN and STF related ancestries. Moreover, we observed that all the ancient groups of the Southern Cone have greater genetic affinity with CCP than with CAN and STF (**S21B,D Figs**). Ancient groups of the Southern Cone (with the exception of Los Rieles ~10900 BP) tend to exhibit higher genetic affinity with CWA than with STF (**S21C Fig**). In addition, the comparisons including STF and CWA also point to a closer genetic relation between ancient groups from the Southern Cone with CWA, although statistical significance is reached only for Late Holocene Patagonia samples.

All the ancient samples from the Andes, independently of their previous assignment to the Early or Late population stream [35], are genetically more similar to CAN than to any of the three other components (**S21A,D,E Figs**). These results support further the hypothesis that the ancestry associated with an early Andean population arrival remains present to some extent in modern Central Andean populations and has not been totally replaced by a later stream of gene flow. Moreover, the archaeological samples in Brazil have higher genetic affinity with CAN than with CWA and CCP (not statistically significant for the latter comparisons). This suggests that divergence of

596 STF with CCP and CWA could have occurred before the divergence of STF with CAN
597 (also observed in **S18A Fig**).

598 Altogether, the trends depicted by these analyzes are expected under a model of
599 early divergence among the four components. Our results also support the
600 hypothesis of genetic continuity over long period of time for the different
601 components after their split.

602 In order to get insights into the past genetic influence among the four components
603 since their divergence, we applied a last f_4 -statistics analysis (**S5G Table; S22 Fig**).
604 We observed that the influence of CAN appears to have been more pronounced into
605 STF component than into CWA and CCP components (**S22B,G,I Figs**). The genetic
606 affinity of CAN with STF, CCP and CWA is reduced as compared to both ancient Early
607 and Late Andes samples (**S22A,H,J Figs**), implying that that the genetic legacy of
608 Andean ancient populations into CAN remains important nowadays. The mutual
609 influence of CCP and CWA components seems to have been important in remote
610 times (**S22K,L Figs**). However, the available ancient samples provide much more
611 precise information on the influence of CCP into CWA component than the other
612 way around. Indeed, CWA does not exhibit higher genetic affinity to any ancient
613 group relative to CCP (**S22K Fig**). The archaeological record for which genetic data
614 has been generated misrepresents CWA since its early divergence with CCP, as well
615 as the common ancestors specific to these two components.

Conclusions

European ancestry in modern Argentinean populations is mostly explained by Southern European origins, although we identified several individuals with Northern European ancestry. African ancestry in Argentina traces its origins from three main components. Although the Bantu-influenced and Western components are clearly the most represented, we also found that an Eastern origin explains some of the African ancestry in Argentina and represents up to ~30% of the African ancestry for some Argentinean individuals. Our work shows that studying more admixed individuals, with a particular focus on extending the geographical coverage of the Argentinean territory, would help to identify the genetic legacy from secondary migration streams, and thus to get a better representation of the complex origins of African and European ancestries in the country.

Concerning Native American ancestry, we concluded that Argentinian populations share, in varying proportions, from three distinct components previously described: Central Andes, Central Chile/Patagonia and Subtropical and Tropical Forests. Moreover, we present evidence supporting the hypothesis that the Native American ancestry in the Central Western region of Argentina may derive from a fourth component that diverged early from the other Native American components, and maintained a tight link with the Central Chile/Patagonia component. This relation is not explained by a putative contribution of admixed individuals from Santiago de Chile (**S23 Fig**). Having identified this component from admixed individuals demonstrates that focusing only on indigenous communities is insufficient, at least

in Argentina, to fully characterize the Native American genetic diversity and decipher the pre-Columbian history of Native Americans. Indeed, most indigenous communities have been culturally annihilated and invisibilized [68,69] to the point that several Argentinean regions were considered “Indian free” in the mid-20st century [70]. However, the cultural incorporation did not necessarily imply a biological extinction. Although studies based on samples from indigenous communities [37–39] provide decisive information to understand the evolutionary history of Native American ancestry, alternative strategies must be considered to fill this gap in the effort to more fully describe the Native American ancestry (e.g. see [71]). Studying admixed individuals can be complex, and leveraging a pure statistical approach, we grouped individuals from rather culturally, ethnically, linguistically, and genetically heterogeneous groups to represent the four Native American components discussed here. Yet, the present study provides useful insights into the routes followed by the main population arrivals in the Southern Cone (**S24 Fig**).

Further efforts are needed to better characterize the Native American ancestry component identified in the Central Western region of Argentina. Particularly, we encourage future studies to confirm the tentative geographical label that we suggest here, and to estimate its influence in the region. Besides these specific questions, many other general questions remain to be answered to better understand the pre-Columbian population dynamics in the Southern Cone such as the time and place of the splits among the components described here, and the extent of genetic exchanges among them. More genotype data for ancient samples, modern

indigenous communities and admixed individuals, particularly in Central, Northeastern and Patagonia regions of Argentina, would help to decipher these issues.

The genomic characterization of populations is an unavoidable practice for many issues ranging from the understanding of our biological heritage, the rational use of biobanks, the definition of an adequate reference genome, the estimation of polygenic risk scores, the study and treatment of simple and complex diseases, and the design of a national program of genomic medicine in our country. This study is a joint effort of Argentinean institutions funded by the national scientific system, and represents the first milestone of the *Consortio PoblAR*, a national consortium for creating a public reference biobank to support biomedical genomic research in Argentina [72]. Genomic knowledge of local populations should be a priority for developing countries to achieve an unbiased representation of diversity in public databases and the scientific development in periphery countries.

Material and Methods

Studied Samples

We genotyped 94 individuals with the Axiom LAT1 array (Affymetrix) from 24 localities and 17 provinces across Argentina (**Fig 1**). These samples were selected among 240 collected by different population genetics groups (Consortio PoblAR) during past sampling campaigns with a biological anthropology focus. According to the available information (e.g. interviews, genealogical information, etc.), each

682 PoblAR research group selected for this study some samples, maximizing the odds
683 that they come from individuals with greater Native American ancestry. For
684 example, surnames were used as a proxy to achieve this objective and the
685 permanence of ancestors in national territory has been another variable that was
686 taken into account. The analyzed 94 samples were also selected to ensure an
687 extended geographical range and were included when they presented sufficient
688 DNA concentration and Native American maternal lineage. Moreover, among the
689 males, we prioritized those with Native American paternal lineage.

690 All saliva and blood samples were collected under written informed consent to
691 participate to this study. The informed consents for research use were approved by
692 several Ethics Committees from (i) Hospital Zonal de Puerto Madryn (Resolución
693 009/2015), (ii) Hospital Zonal de San Carlos de Bariloche (Resolución 1510/2015),
694 (iii) Hospital Provincial de Pediatría Dr. Fernando Barreyro, (iv) Investigaciones
695 Biomédicas IMBICE (RENIS CE000023), (v) Provincia de Jujuy, (vi) Hospital Italiano
696 of Buenos Aires (protocolo 1356/09); and (vii). Centro de Educación Médica e
697 Investigaciones Clínicas “Norberto Quirno” (Resolución 612/2018). The biological
698 samples were coded and anonymized, as per the Argentina National Law 25.326 of
699 Protection of Personal Data.

700 87 samples and 791,543 autosomal variants passed the standard Affymetrix
701 genotyping Quality Controls (**S1 Table**).

702 Most of the genotype data processing was performed using in-house scripts in R
703 [73] and perl [74], leveraging *plink2* [75], *vcftools v1.13* [76], and *bedTools v2* [77].

We compiled the genotype data for the 87 Argentinean samples with different genotype data available in the literature. We focused our study on biallelic SNPs (removing indels and SNPs with more than 2 alleles). Any putative inconsistent strand had been fixed processing to the relevant flip, filtering out any SNP with ambiguous genotype (A/T, G/C).

Cryptic relatedness among samples were assessed using King software [78]. To avoid any 1st degree relationship, we filtered out individuals, minimizing the total number of removals. No admixed individuals had been removed at this step. We thus build different dataset arrangements (named **DS<n>**) that we analyzed through this work (**S2 Table**).

Argentinean genetic diversity within a worldwide context

To analyze genetic diversity in Argentina within a worldwide context, we built the Dataset1 (**DS1**). This dataset contains 87 Argentinean samples, 654 African, 503 European and 179 South American samples from 1KGP [42], 54 modern unrelated Chilean samples [37], and 161 Native-American individuals from South America [38]. Moreover, **DS1** included genotype data from [31] which consists in 82 individuals from Lima (Peru), 27 from Santiago de Chile, and 161 from Argentinean urban centers.

We filtered out any variant and individual exhibiting in the compiled data set more than 2% and 5% of missing genotypes (--geno 0.02 and --mind 0.05 flags in plink 1.9), respectively, as well as Minor Allele Frequency below 1% (--maf 0.01 flag in plink 1.9). The filtered data was then pruned for Linkage-Disequilibrium (--indep-

pairwise 50 5 0.5 flag in plink 1.9). The combined data set has a total intersection of 59,237 SNPs and 1,908 individuals. With this curated data we performed Principal Component Analyses [41] (**S1 Fig**) and Admixture [43] (**Fig 2, S2 Fig and S3 Table**).

Local Ancestry

Local ancestry analyses rely on haplotype reconstruction (phasing) and require high SNP density. Since, admixed individuals of interest were genotyped on different microarray platforms, we decided to perform two separate local ancestry analyses on two different data sets (**DS2** and **DS3**). **DS2** consists of 87 Argentinean individuals from the present study, and 54 Chilean individuals [37], all of them genotyped with the Axiom LAT1 microarray. **DS3** includes 175 Argentinean, 27 Chilean, and 119 Peruvian genotyped with the Illumina OMNI1 microarray [31]. Both **DS2**, **DS3** were merged with 1KGP data consisting in 503 phased reference samples for each of the African and European genetic ancestry, and 347 Latin American individuals.

The 87 Argentinean, and the 54 Chilean samples were phased with shapeIT2 [79,80]. The genetic map, and 5,008 haplotypes panel provided by 1KGP were downloaded from http://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3/. Algorithm, and model parameters were used by default, filtering out monomorphic SNPs, and those with more than 2% of missing genotype. We obtained phased genotype data for 608,501 autosomal SNPs. This data was then merged with phased 1KGP genotype data for African, American and European samples described before. Since this data set derived from DS2, we call it **DS2p** (phased DS2).

The 175 Argentinean, 27 Chilean and 119 Peruvian samples from [31] were phased separately using the same procedure with shapeIT2. After filtering for missing genotypes and merging with phased 1KGP genotype data for African, American and European samples we obtained phased data for 694,626 autosomal SNPs. Since this data set derived from DS3, we call it **DS3p** (phased DS3).

In **DS2** and **DS3** we ran Admixture using K parameter minimizing the Cross-Validation score. We used individuals with more than 99% of Native American ancestry as references for local ancestry estimation. For **DS2** we used $K = 7$, and Native American ancestry was defined as the sum of the two American specific components observed (**S25 Fig**). According to this criterion, 48 individuals were assigned as Native American reference. All the other American samples were defined as Admixed.

For **DS3**, we used $K = 5$, and Native American ancestry was defined as the single American specific component observed (**S26 Fig**). According to this criterion, 19 individuals were assigned as Native American reference. All the other American samples were defined as Admixed.

RFMix v2 ([44] downloaded at <https://github.com/slowkoni/rfmix> on 15th August 2018) was run on **DS2P** and **DS3P** separately using parameter settings similar to [81]. The reference panels consist of the African and European samples from 1KGP, as well as Native American individuals identified through Admixture procedure described before. We used 1 Expectation-Maximization iteration (-e 1), actualizing the reference panel in this process (--reanalyze-reference). We used CRF spacing

size and random forest window size of 0.2 cM (-c 0.2 and -r 0.2). We use a node size of 5 (-n 5). We set the number of generations since admixture to 11 (-G 11) considering the estimates from [31]. The forward-backward output was then interpreted to assign allele ancestry to the one exhibiting major posterior probability, conditioning that it was greater than 0.9. Otherwise, the allele ancestry was assigned to Unknown (UKN). As a sanity check, the global ancestry proportions estimated through this RFMix analysis were compared with those obtained with Admixture software. The global ancestry proportion estimates obtained by both procedures matched very well: spearman's correlation greater than 0.9 in American samples for any of the 3 continental ancestries (**S27 Fig**). Moreover, the ancestry ditypes assigned in admixed individuals from 1KGP (included in both DS2p and DS3p) were highly consistent between both independent masking procedures (**S28 Fig**).

Ancestry Specific Population Structure

In order to analyze the ancestry-specific population structure we masked the data, i.e. for each individual, we assigned missing genotype for any position for which at least one of the two alleles was not assigned to the relevant ancestry. In other words, to study ancestry A, we kept for each individual, regions exhibiting ancestry A on both haplotypes (ditypes) as illustrated in **S5 Fig**

European ancestry specific population structure

To study European ancestry specific population structure, we analyzed together masked data for this ancestry for Colombian individuals from 1KG and individuals

from **DS2P** and **DS3P** excluding individuals from Chilean Native American communities [37]. This data was merged with a set of reference individuals with European ancestry [46], which is a subset of the POPRES dataset [45]. We call this data set as **DS4**. We removed individuals with less than 30% SNPs with the ancestry ditypes (--mind 0.7 with *plink* 1.9). We also removed SNPs with more than 50% of missing genotypes (--geno 0.5 with *plink* 1.9). Thus, **DS4** contains 132 modern Argentinean individuals (29 from the present study and 103 from [31]), 17 individuals from Santiago de Chile, 4 from Lima and 74 from Colombia [42]. **DS4** encompasses 29,347 SNPs of which 27,634 remained after LD-pruning (--indep-pairwise 50 5 0.5 flag in *plink2*).

Smartpca from Eigensoft v7.2.0 was run on **DS4** [41] with the *lsqproject* option ON. We report the PCA results summarized into a 2-dimensional space by applying Multidimensional Scaling on weighted Euclidian distance matrix for the first N PCs. We weighted each PC by the proportion of variance it explains. We selected the N most informative PCs according to the Elbow method on the proportion of explained variance. Admixture [43] was run with *K* ranging from 2 to 10 with cross-validation procedure.

African ancestry specific population structure

To study African ancestry specific population structure, we analyzed together masked data for this ancestry for individuals from **DS2P** and **DS3P**. This data was merged with a compilation of reference individuals with African ancestry from [42,48–50]. We removed African individuals with less than 99% of African ancestry

when comparing them to the 2504 individuals from 1KGP (Admixture with $K=7$ minimizing cross-validation score). We thus reduced the African reference to 1685 individuals. We call as **DS5** the data set containing both the masked data for admixed South American individuals and African reference individuals.

We removed SNPs with more than 10% of missing genotypes (--geno 0.1 with *plink* 1.9), and individuals with less than 5% of the ancestry ditypes (--mind 0.95 with *plink* 1.9). Thus, **DS5** contains, 26 modern Argentinean individuals (all from the present study), and 12 individuals from Lima (9). **DS5** consisted in 137,136 SNPs, of which 128,086 remained after LD-pruning (--indep-pairwise 50 5 0.5 flag in *plink2*).

PCA and Admixture were performed as for European ancestry specific population structure analyses (described before).

Native American ancestry specific population structure

To study Native American ancestry specific population structure, we analyzed together masked data for this ancestry for individuals from **DS2P** and **DS3P**. This data was merged with pseudo-haploid data for ancient samples within South and Central America [35–37,54,55], as well as with masked data for Native American individuals from [38]. The pseudo-haploid data for ancient samples was downloaded from the Reich Laboratory webpage (<https://reich.hms.harvard.edu/downloadable-genotypes-present-day-and-ancient-dna-data-compiled-published-papers>) the 15th of May 2019. For each sample, we used annotations (geographic coordinates and approximate date) from the metafile provided at the same url.

We call this data set **DS6**. We removed individuals with less than 30% SNPs with the ancestry ditypes (--mind 0.7 with *plink* 1.9). We also removed SNPs with less than 50% individuals with the ancestry ditypes (--geno 0.5 with *plink* 1.9). **DS6** contains 146 modern Argentinean individuals (74 from the present study and 72 from [31]), 22 individuals from Santiago and 77 from Lima, along with 207 Native South American individuals from [37,38] and 53 ancient samples. **DS6** encompasses 47,003 SNPs, of which 39,423 remained after LD-pruning (--indep-pairwise 50 5 0.5 flag in *plink2*).

PCA and Admixture were performed as for European ancestry specific population structure analyses (described before), with the difference that the *poplistname* option was set for *smartpca* in order to estimate the PC using only modern samples and project the ancient samples.

Statistical assignation of modern South American individuals to Native American components

Given a distance matrix among modern individuals from **DS6**, we performed *K-means* clustering with the number of clusters ranging from 2 to 20 and selected the *K-means* output minimizing the Bayesian Information Criterion (BIC). This procedure was applied to three different distance matrix: (i) Weighted Euclidean distance based on the two first PCs from Principal Component Analysis (ii) Euclidean distance based on the ancestry proportions estimated from Admixture ($K=3$) and (iii) $1 - f_3(Ind1, Ind2, Yoruba)$, where *Ind1* and *Ind2* are two individuals. In the three cases, the *K-means* procedure minimized the BIC when considering 4 clusters. These clusters correspond to the three Native American Component

discussed in this paper (CAN, CCP and STF), along with one laying in-between, which we finally attributed to Central Western Argentina. However, individual assignment to one of these four clusters was not totally consistent according to the distance matrix we used. To obtain a robust assignation, an individual was assigned to a given cluster when it consistently belonged to it across the three *K*-means procedures, otherwise it had been removed for following analyses (**S16 Fig**). The cluster assignation for each individual is given in **S4 Table**.

We computed pairwise F_{ST} among the groups using smartpca [41] with fstonly and lsqproject set to YES, and all the other settings left at default. Standard errors were estimated with the block jackknife procedure [41].

***F*-statistics to infer relationship between the four Native American components in Argentina and their genetic affinity with ancient populations**

Starting from genotype data for the individuals in **DS6**, we also included genotype data from Yoruba (YRI) population from [42], Mixe population from [38] and, pseudo-haploid data for Anzick individual from the Clovis culture [67] and USR1 individual from Upward Sun River in Beringia [66]. The resulting data set is called **DS7**. Within **DS7**, we grouped modern individuals according to their assigned Native American group and we removed those with inconsistent assignation (**S16 Fig** and **S4 Table**). Thus, **DS7** contains 426 modern individuals, 55 ancient samples, 108 Yoruba and 17 Mixe individuals. **DS7** encompasses 88,564 SNPs. Note that we did not apply any SNP filtering overall **DS7** in order to maximize the number of SNPs included in each group comparison considered.

Using modern individuals from **DS7**, and considering any possible combination of the four groups identified (STF, CCP, CAN and CWA), we computed the f_3 statistics [82] in the form of $f_3(\textit{Target}; S1, S2)$. This allowed to contrast whether *Target*, *S1* and *S2* could be organized in the form of a phylogenetic tree (positive f_3) or whether the *Target* group is the result of mixture between *S1* and *S2* groups (negative f_3). We also computed $f_4(\textit{YRI}, \textit{Target}; S1, S2)$ to test whether group *Target* shares more evolutionary history with group *S1* (negative f_4) or group *S2* (positive f_4). Moreover, we reconstructed the Neighbour-joining tree from the matrix of distances with Phylip v3.2 [58] and *USR1* as outgroup. The distances were estimated as $1/f_3\text{-outgroup}(\textit{YRI}; X, Y)$.

We computed $f_3\text{-Outgroup}$ and f_4 statistics in order to estimate the genetic affinity of the four Native American components with ancient populations from South and Central America. We computed the $f_3\text{-outgroup}$ and f_4 statistics in the form of $f_3(\textit{YRI}; X, \textit{Ancient})$ and $f_4(\textit{YRI}, X; \textit{USR1}, \textit{Ancient})$, where *X* the represent the cluster containing all individuals assigned to a given Native American ancestry component. We also computed f_4 of the form $f_4(\textit{YRI}, \textit{Ancient}; X, Y)$. This f_4 setting allowed testing which of *X* or *Y*, each referring to one of the four Native American components, shares more ancestry with a given *Ancient* group. *X* and *Y*.

We finally computed f_4 of the form $f_4(\textit{Ancient}, X; Y, \textit{YRI})$ to test whether a given modern Native American component *Y* shares exhibit closer genetic affinity with a given *Ancient* group (negative f_4) or with another modern Native American component *X* (positive f_4).

All the results based on *F-statistics* are listed in **S5 Table**. We assessed significance of a comparison considering 3 standard errors ($|Z| > 3$) from block jackknife procedure [82].

Geographical maps

All maps have been generated in R using, *maps* package , (<https://cran.r-project.org/web/packages/maps/index.html>). **S24 Fig** includes a raster downloaded at <https://www.naturalearthdata.com/> and integrated in R with *raster* (<https://CRAN.R-project.org/package=raster>) and *RStoolbox* (<http://bleutner.github.io/RStoolbox>) packages.

Abbreviations

1KGP: 1000 Genomes Project
AMBA: Metropolitan Area of Buenos Aires
BIC: Bayesian Information Criterion
CAN: Central Andes
CCP: Central Chile / Patagonia
CLM: Colombians from Medellin
CWA: Central Western Region in Argentina
CYA: Cuyo Region in Argentina
DS: Data set

- 923 **LD:** Linkage Disequilibrium
- 924 **NEA:** Northeastern Region in Argentina
- 925 **NWA:** Northwestern Region in Argentina
- 926 **PCA:** Principal Component Analysis
- 927 **PEL:** Peruvians from Lima
- 928 **PPA:** Pampean Region in Argentina
- 929 **PTA:** Patagonia Region in Argentina
- 930 **SNP:** Single Nucleotide Polymorphism
- 931 **STF:** Subtropical and Tropical Forests
- 932 **YRI:** Yoruba individuals in Ibadan, Nigeria

933 **Acknowledgments**

- 934 In memoriam of Raúl Carnese.
- 935 We thank Rolando González-José, Andrea Llera and Mariana Berenstein for the
- 936 management and promotion of the Consorcio Poblar.
- 937 We thank all the participants of the different sampling campaigns across Argentina
- 938 from which the newly genotyped samples are derived from.
- 939 We thank Ricardo A. Verdugo, Etienne Patin, Luca Pagani, Marta E. Alarcón
- 940 Riquelme and María Teruel Artacho who kindly shared genotype data included in
- 941 this study.

942 We thank Laura Fejerman for her proofreading of the manuscript.

943 This project has been founded by FONCyT-ANPCyT (PICT2014-1597 attributed to

944 HD and PICT3655-2016 attributed to PL) and CONICET (PIP 0208/14 attributed to

945 HD and VR)

946 **Disclosure Declaration**

947 PL provides consulting services to myDNAmap S.L.

948 JMB and JZ are employed by Biocódices S.A; HD is the scientific director of

949 Biocódices S.A.

950 Biocodices provided support in the form of part-time employment salaries for

951 authors HD, JMB and JZ.

952 Neither Biocódices nor myDNAmap have made any decision in the design of the

953 study, sampling biological material, the collection, and the analysis of the data.

954 These companies have also not made any decision on the preparation and

955 publication of the work. Biocódices provided inputs for DNA's sample extraction and

956 access to computing cluster resources.

957 **Data Availability Statement**

958 The data analyzed here comprises both newly generated and previously reported

959 data sets. Access to publicly available data sets should be requested through the

960 distribution channels indicated in each published study. For newly genotyped

961 samples, individual genotype data will be made available at
962 [http://biocodices.hopto.org/Affymetrix Pop Arg/](http://biocodices.hopto.org/Affymetrix_Pop_Arg/) at time of publication.

963 Bibliography

- 964 1. Bravi CM, Sans M, Bailliet G, Martínez-Marignac V, Portas M, Barreto I, et al.
965 Characterization of mitochondrial DNA and Y-chromosome haplotypes in a
966 Uruguayan population of African ancestry. Hum Biol. 1997;69: 641–652.
- 967 2. Sala A, Penacino G, Corach D. Comparison of allele frequencies of eight STR
968 loci from Argentinean Amerindian and European populations. Hum Biol.
969 1998;70: 937–947.
- 970 3. Goicoechea A, Carnese FR, Dejean C, Avena SA, Weimer TA, Franco MHL, et
971 al. Genetic Relationships Between Amerindian Populations of Argentina. Am J
972 Phys Anthropol. 2001;115: 133–143.
- 973 4. Dejean CB, Crouau-Roy B, Goicoechea AS, Avena SA, Carnese FR. Genetic
974 variability in Amerindian populations of northern Argentina. Genet Mol Biol.
975 2004;27: 489–495.
- 976 5. Fejerman L, Carnese F, Goicoechea A, Avena S, Dejean C, Ward R. African
977 ancestry of the population of Buenos Aires. Am J Phys Anthropol. 2005;128:
978 164–70.
- 979 6. Wang S, Ray N, Rojas W, Parra M V., Bedoya G, Gallo C, et al. Geographic
980 patterns of genome admixture in latin American mestizos. PLoS Genet.

981 2008;4: e1000037.

982 7. Bobillo MC, Zimmermann B, Sala A, Huber G, Röck A, Bandelt H-J, et al.

983 Amerindian mitochondrial DNA haplogroups predominate in the population

984 of Argentina: towards a first nationwide forensic mitochondrial DNA

985 sequence database. Int J Legal Med. 2010;124: 263–268.

986 8. Pauro M, Garcia A, Bravi CM, Demarchi DA. Distribucion de haplogrupos

987 mitocondriales aloctonos en poblaciones rurales de Cordoba y San Luis. Rev

988 Argentina Antropol Biol. 2010;12: 47–55.

989 9. Pauro M, García A, Nores R, Demarchi D a. Analysis of uniparental lineages in

990 two villages of Santiago Del Estero, Argentina, seat of Pueblos de Indios in

991 colonial times. Hum Biol. 2013;85: 699–720.

992 10. García A, Pauro M, Bailliet G, Bravi CM, Demarchi DA. Genetic variation in

993 populations from central Argentina based on mitochondrial and Y

994 chromosome DNA evidence. J Hum Genet. 2018.

995 11. Martínez Marignac VL, Bertoni B, Parra EJ, Bianchi NO. Characterization of

996 admixture in an urban sample from Buenos Aires, Argentina, using

997 uniparentally and biparentally inherited genetic markers. Hum Biol. 2004;76:

998 543–57.

999 12. Seldin MF, Tian C, Shigeta R, Scherbarth HR, Silva G, Belmont JW, et al.

1000 Argentine Population Genetic Structure: Large Variance in Amerindian

1001 Contribution. Am J Phys Anthropol. 2007;132: 455–462.

1002 13. Corach D, Lao O, Bobillo C, van Der Gaag K, Zuniga S, Vermeulen M, et al.

1003 Inferring continental ancestry of argentineans from autosomal, Y-
1004 chromosomal and mitochondrial DNA. *Ann Hum Genet.* 2010;74: 65–76.

1005 14. Parolin ML, Toscanini UF, Velázquez CL, Berardi GL, Holley A, Tamburrini C, et
1006 al. Genetic admixture patterns in Argentinian Patagonia. *PLoS One.* 2019;14:
1007 e0214830.

1008 15. Avena S, Via M, Ziv E, P??rez-Stable EJ, Gignoux CR, Dejean C, et al.
1009 Heterogeneity in genetic admixture across different regions of argentina. *PLoS*
1010 *One.* 2012;7: e34695.

1011 16. García A, Dermarchi DA, Tovo-Rodrigues L, Pauro M, Callegari-Jacques SM,
1012 Salzano FM, et al. High interpopulation homogeneity in Central Argentina as
1013 assessed by ancestry informative markers (AIMs). *Genet Mol Biol.* 2015;38:
1014 324–331.

1015 17. Salas A, Richards M, Lareu M-V, Scozzari R, Coppa A, Torroni A, et al. The
1016 African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am J Hum*
1017 *Genet.* 2004;74: 454–465.

1018 18. García A, Pauro M, Nores R, Bravi CM, Demarchi DA. Phylogeography of
1019 mitochondrial haplogroup D1: An early spread of subhaplogroup D1j from
1020 Central Argentina. *Am J Phys Anthropol.* 2012;149: 583–590.

1021 19. Corach D, Figueira Risso L, Marino M, Penacino G, Sala A. Routine Y-STR
1022 typing in forensic casework. *Forensic Sci Int.* 2001;118: 131–135.

1023 20. Salas A, Jaime JC, Álvarez-Iglesias V, Carracedo Á. Gender bias in the
1024 multiethnic genetic composition of central Argentina. *J Hum Genet.* 2008;53:

- 1025 662–674.
- 1026 21. Martínez Sarasola C. Nuestros paisanos los indios. 9th ed. Buenos Aires: Del
1027 Nuevo Extremo; 2011.
- 1028 22. Devoto F. Historia de la inmigración en la Argentina. 1a ed. Buenos Aires:
1029 Editorial Sudamericana; 2003.
- 1030 23. Studer E. La trata de negros en el Río de la Plata durante el siglo XVIII. Buenos
1031 Aires: Instituto de Historia de Argentina “Doctor Emilio Ravignani”; 1958.
- 1032 24. Rodríguez Molas R. Los sometidos de la conquista. Centro Edi. Buenos Aires;
1033 1985.
- 1034 25. Borucki A. The Slave Trade to the Río de la Plata, 1777–1812: Trans-Imperial
1035 Networks and Atlantic Warfare. *Colon latin Am Rev.* 2011;20: 81–107.
- 1036 26. Guzmán F. Africanos en la Argentina. Una reflexión desprevenida. Andes.
1037 2006;17.
- 1038 27. Pauro M. Análisis molecular de linajes uniparentales en poblaciones humanas
1039 del centro de Argentina. Universidad de Córdoba. 2015.
- 1040 28. Di Fabio Rocca F. La presencia subsahariana en el acervo génico de
1041 poblaciones cosmopolitas de la Argentina. Universidad de Buenos Aires. 2016.
- 1042 29. Mandrini R. La Argentina aborigen. De los primeros pobladores a 1910. Siglo
1043 XXI. Buenos Aires; 2008.
- 1044 30. Lenton DI, Delrio WM, Pérez PMV, Papazian AER, Nagy MA, Musante M.
1045 Argentina’s Constituent Genocide: Challenging the Hegemonic National

1046 Narrative and Laying the Foundation for Reparations to Indigenous Peoples.
1047 Armen Rev. 2012;53: 63–84.

1048 31. Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-
1049 Tello P, et al. Genomic Insights into the Ancestry and Demographic History of
1050 South America. PLOS Genet. 2015;11: e1005602.

1051 32. Muzzio M, Motti JMB, Sepulveda PBP, Yee M, Santos R, Ramallo V, et al.
1052 Population structure in Argentina. PLoS One. 2018;13: e0196325.

1053 33. Gouveia MH, Borda V, Leal TP, Moreira RG, Bergen AW, Kehdy FSG, et al.
1054 Origins, admixture dynamics and homogenization of the African gene pool in
1055 the Americas. Mol Biol Evol. 2020;Epub ahead: pii: msaa033.

1056 34. Kehdy FSG, Gouveia MH, Machado M, Magalhães WCS, Horimoto AR. Origin
1057 and dynamics of admixture in Brazilians and its effect on the pattern of
1058 deleterious mutations. Proc Natl Acad Sci U S A. 2015;112: 8696–8701.

1059 35. Posth C, Nakatsuka N, Lazaridis I, Fehren-schmitz L, Krause J, Reich D, et al.
1060 Reconstructing the Deep Population History of Central and South America
1061 Article Reconstructing the Deep Population History of Central and South
1062 America. Cell. 2018;175: 1185–1197.

1063 36. Moreno-Mayar JV, Vinner L, Damgaard PDB, Fuente C De, Margaryan A,
1064 Orbegoza MI, et al. Early human dispersals within the Americas. Science.
1065 2018;362: pii: eaav2621.

1066 37. de la Fuente C, Ávila-arcos MC, Galimany J, Carpenter ML, Homburger JR.
1067 Genomic insights into the origin and diversification of late maritime hunter-

gatherers from the Chilean Patagonia. *Proc Natl Acad Sci U S A*. 2018;115:
E4006–E4012.

38. Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, et al.
Reconstructing Native American population history. *Nature*. 2012;488: 370–4.

39. Gneccchi-ruscone GA, Sarno S, Fanti S De, Gianvincenzo L, Giuliani C, Boattini A,
et al. Dissecting the Pre-Columbian Genomic Ancestry of Native Americans
along the Andes – Amazonia Divide. *Mol Biol Evol*. 2019;36: 1254–1269.

40. Gomez-Carballa A, Pardo-Seco, Brandini S, Achilli A, Perego UA, Coble MD, et
al. The peopling of South America and the trans-Andean gene flow of the first
settlers. *Genome Res*. 2018 Jun;28(6):767-779.

41. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS
Genet*. 2006;2: 2074–2093.

42. The 1000 Genomes Project Consortium. A global reference for human genetic
variation. *Nature*. 2015;526: 68–74.

43. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry
in unrelated individuals. *Genome Res*. 2009;19: 1655–64.

44. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A discriminative
modeling approach for rapid and robust local-ancestry inference. *Am J Hum
Genet*. 2013;93: 278–288.

45. Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, et al. The
Population Reference Sample, POPRES: A Resource for Population, Disease,

1089 and Pharmacological Genetics Research. Am J Hum Genet. 2008;83: 347–358.

1090 46. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes
1091 mirror geography within Europe. Nature. 2008;456: 98–101.

1092 47. Gallero MC, Krautstoft EM. Proceso de poblamiento y migraciones en la
1093 Provincia de Misiones , Argentina (1881-1970). Avá Rev Antropol. 2010;16:
1094 1881–1970.

1095 48. Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, et al.
1096 Dispersals and genetic adaptation of Bantu-speaking populations in Africa and
1097 North America. Science. 2017;546: 543–546.

1098 49. Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Romero IG, et al.
1099 Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex
1100 Influences on the Ethiopian Gene Pool. Am J Hum Genet. 2012;91: 83–96.

1101 50. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al.
1102 Genomic Variation in Seven Khoe-San Groups Reveals Adaptation and
1103 Complex African History. Science. 2012;338: 374–9.

1104 51. Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR,
1105 et al. Reconstructing the Population Genetic History of the Caribbean. PLoS
1106 Genet. 2013;9.

1107 52. Fortes-lima C, Gessain A, Ruiz-linares A, Restrepo BN, Rojas W, Migot-nabias
1108 F, et al. Genome-wide Ancestry and Demographic History of African-
1109 Descendant Maroon Communities from French Guiana and Suriname. Am J
1110 Hum Genet. 2017;101: 725–736.

- 1111 53. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al.
1112 Genome-wide patterns of population structure and admixture in West
1113 Africans and African Americans. *Proc Natl Acad Sci U S A*. 2010;107: 786–791.
- 1114 54. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M,
1115 et al. Genomic evidence for the Pleistocene and recent population history of
1116 Native Americans. *Science* (80-). 2015;349: aab3884.
- 1117 55. Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, et al. The genetic
1118 prehistory of the Andean highlands 7000 years BP though European contact.
1119 *Sci Adv*. 2018;4: eauu4921.
- 1120 56. Motti JMB, Muzzio M, Ramallo V, Kladniew BR, Alfaro EL, Dipierri JE, et al.
1121 Origen Y Distribución Espacial De Linajes Maternos Nativos En El Noroeste Y
1122 Centro Oeste Argentinos. *Rev Argentina Antropol Biol*. 2013;15: 3–14.
- 1123 57. Motti J. Caracterización de Linajes Maternos en la Población Actual del
1124 Noroeste y Centro-Oeste Argentinos. Universidad Nacional de La Plata. 2012.
- 1125 58. Felsenstein J. PHYLIP-Phylogeny Inference Package (Ver. 3.2). *Cladistics*.
1126 1989;5: 164–166.
- 1127 59. Motti JMB, Schwab ME, Beltramo J, Jurado-medina LS, Muzzio M, Ramallo V, et
1128 al. Regional differentiation of native American populations from the analysis
1129 of maternal lineages. *Intersecc en Antropol*. 2017;18: 271–282.
- 1130 60. Canals Frau S. Etnología de los Huarpes. Una síntesis. *Anales del Instituto de*
1131 *Etnología Americana*. Anales del Instituto de Etnología Americana, Tomo VII.
1132 Mendoza; 1946. pp. 9–147.

- 1133 61. Escolar D. Los dones étnicos de la Nación. Identidades huarpe y modos de
1134 producción de soberanía en Argentina. Prometeo E. Buenos Aires; 2007.
- 1135 62. Bengoa J. Historia de los antiguos mapuches del sur. Desde antes de la llegada
1136 de los españoles hasta las paces de Quilín. Editorial. Santiago; 2003.
- 1137 63. Galdames S. ¿Detuvo la batalla del Maule la expansión inca hacia el sur de
1138 Chile? Cuadernos Hist. 2017;3: 7–25.
- 1139 64. Michieli CT. Antigua historia de Cuyo. Ansilta Ed. San Juan; 1994.
- 1140 65. Gascón M. Cuyo en el espacio imperial. La fase de configuración: 1580-1680.
1141 Rev Tefros. 2011;9: 1–20.
- 1142 66. Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst
1143 J, et al. Terminal Pleistocene Alaskan genome reveals first founding
1144 population of Native Americans. Nature. 2018;553: 203–207.
- 1145 67. Rasmussen M, Anzick SL, Waters MR, Skoglund P, Degiorgio M, Jr TWS, et al.
1146 The genome of a Late Pleistocene human from a Clovis burial site in western
1147 Montana. Nature. 2014;506: 225–229.
- 1148 68. Michieli CT. La disolución de la categoría Jurídico-Social de “indio” en el siglo
1149 XVIII: El caso de San Juan (Región de Cuyo). Publicacio. Universidad Nacional
1150 de San Juan; 2000.
- 1151 69. Katzer L. El mestizaje como dispositivo biopolítico. Biblos. Pueblos indígenas
1152 Interculturalidad, colonialidad, política. Biblos. Buenos Aires; 2009. pp. 59–7.
- 1153 70. Briones C. Construcciones de aboriginalidad en Argentina. Société suisse des

1154 Américanistes / Schweizerische Amerikanisten-Gesellschaft. 2004;Bulletin 6:
1155 73–90.

1156 71. Mas-sandoval A, Arauna LR, Gouveia MH, Barreto ML, Horta BL, Lima-costa
1157 MF, et al. Reconstructed Lost Native American Populations from Eastern
1158 Brazil Are Shaped by Differential Jê/Tupi Ancestry. *Genome Biol Evol.*
1159 2019;11: 2593–2604.

1160 72. Dopazo H, Llera AS, Berenstein M, Gonzáles-josé R. Genomas , enfermedades y
1161 medicina de precisión : un Proyecto Nacional. *Ciencia, Tecnol y Política.*
1162 2019;2: 1–10.

1163 73. Team RC. R: A language and environment for statistical computing. R
1164 Foundation for Statistical Computing, Vienna, Austria. 2019.

1165 74. Wall L, Christiansen T, Orwant J. Programming perl. O'Reilly M. 2000.

1166 75. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-
1167 generation PLINK: Rising to the challenge of larger and richer datasets.
1168 *Gigascience.* 2015;4: 1–16.

1169 76. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, Depristo MA, et al. The
1170 variant call format and VCFtools. *Bioinformatics.* 2011;27: 2156–2158.

1171 77. Quinlan AR, Hall IM. BEDTools : a flexible suite of utilities for comparing
1172 genomic features. *Bioinformatics.* 2010;26: 841–842.

1173 78. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust
1174 relationship inference in genome-wide association studies. *Bioinformatics.*

1175 2010;26: 2867–73.

1176 79. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for
1177 thousands of genomes. Nat Methods. 2011;9: 179–181.

1178 80. Delaneau O, Marchini J, Zagury J. A linear complexity phasing method for
1179 thousands of genomes. Nat Methods. 2012;9: 179–181.

1180 81. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al.
1181 Human Demographic History Impacts Genetic Risk Prediction across Diverse
1182 Populations. Am J Hum Genet. 2017;100: 635–649.

1183 82. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient
1184 admixture in human history. Genetics. 2012;192: 1065–93.

1185

Supplementary Figure Captions

S1 Fig. Principal component analysis in a worldwide Context.

A: PC2 vs PC1; **B:** PC4 vs PC3; **C:** PC6 vs PC5. The percentage of variance explained by each principal component (PC) is given. Each point represents an individual following the color and point codes given in legend.

S2 Fig. Admixture analysis in a worldwide context.

(A) Cross-Validation score for Admixture runs on the worldwide meta dataset (**DS1**) with K from 3 to 12. **(B-G)** Admixture results with $K=3$ to $K=8$. 1KGP: 1000 Genomes Project; CYA: Cuyo Region; NEA: Northeastern Region, NWA: Northwestern Region; PPA: Pampean Region; PTA: Patagonia Region.

S3 Fig. Comparison of different admixture continental ancestry proportion estimates in a worldwide context.

Comparison of the African, European and Native American ancestry proportion estimates obtained with Admixture models with $K=3$ and $K=8$ applied to **DS1**. **(A)** African ancestry proportions for $K=3$ are as observed in green in **S2B Fig**, while for $K=8$ they are estimated as the sum of the three greenish colors observed in **Main Fig 2**. **(B)** European ancestry proportions for $K=3$ are as observed in blue in **S2B Fig**, while for $K=8$ they are estimated as the sum of the two bluish colors observed in **Main Fig 2**. **(C)** Native American ancestry proportions for $K=3$ are as observed in orange in **S2B Fig**, while for $K=8$ they are estimated as the sum of the three reddish colors observed in **Main Fig 2**.

S4 Fig. Correlation between eigenvectors and ancestry proportion estimates from analyses in a worldwide context.

Comparison of ancestry proportion estimates from Admixture model with $K=8$ and the 6 first Principal Components (PCs) in **DS1**. **(A)** PC1 vs African ancestry proportions (estimated as the sum of the three greenish colors observed in **Main Fig 2**). **(B)** PC2 vs European ancestry proportions (estimated as the sum of the two bluish colors observed in **Main Fig 2**). **(C)** PC2 vs Native American ancestry proportions (estimated as the sum of the three reddish colors observed in **Main Fig 2**). **(D)** PC6 vs Bantu-influenced ancestry proportions (dark olive green in **Main Fig 2**). **(E)** PC6 vs Western African ancestry proportions (dark green in **Main Fig 2**). **(F)** PC4 vs Southern European ancestry proportions (light blue in **Main Fig 2**). **(G)** PC4 vs Northern European ancestry proportions (dark blue in **Main Fig 2**). **(H)** PC3 vs Central Chile / Patagonia ancestry proportions (orange in **Main Fig 2**). **(I)** PC3 vs Central Andes ancestry proportions (yellow in **Main Fig 2**). **(J)** PC5 vs Subtropical and Tropical Forests ancestry proportions (pink in **Main Fig 2**).

S5 Fig. Example of a local ancestry output.

(A) RFMIX output for a given admixed individual. **(B)** Masked genotype showing diatypes of Native American (red), European (blue) and African (green) ancestry.

- 1225 Gaps are represented in grey and regions with unassigned ancestry (Unknown) are
1226 in black.
- 1227 **S6 Fig. Choice of the number of principal components from European**
1228 **ancestry-specific principal component analysis.**
1229 Elbow method to determine which PC minimizes the angle of the curve from the
1230 chart “Percentage of variance explained *versus* Number of PCs”
- 1231 **S7 Fig. European ancestry specific admixture analysis.**
1232 **(A)** Cross-Validation scores for K from 2 to 10. **(B)** Admixture for $K=2$. **(C)**
1233 Admixture for $K=3$. CYA: Cuyo Region; NEA: Northeastern Region, NWA:
1234 Northwestern Region; PPA: Pampean Region; PTA: Patagonia Region.
- 1235 **S8 Fig. Comparision of different European ancestry proportions in South**
1236 **America.**
1237 Comparison of ancestry proportion estimates from European Ancestry Specific
1238 Admixture Analyses ($K=3$) among samples from the Argentina, Chile and Colombia
1239 **(A)** Southeaster/Italian ancestry (light blue in **S7C Fig**). **(B)** Iberian ancestry
1240 (turquoise in **S7C Fig**). **(C)** Northern Ancestry (dark blue in **S7C Fig**). P -value of the
1241 Wilcoxon test for each pairwise comparison is shown.
- 1242 **S9 Fig. African ancestry-specific principal component analysis.**
1243 **(A)** Localization map of the 1685 reference samples with >99% of African ancestry.
1244 **(B-C)** Principal Components performed using the African reference samples
1245 (represented as in panel **A**), and South American samples masked for African
1246 ancestry.
- 1247 **S10 Fig. African ancestry-specific admixture analysis.**
1248 **(A)** Cross-validation scores K from 2 to 10. **(B)** Admixture for $K=2$. **(C)** Admixture
1249 for $K=3$. **(D)** Admixture plots for $K=4$. CYA: Cuyo Region; NEA: Northeastern Region,
1250 NWA: Northwestern Region; PPA: Pampean Region; PTA: Patagonia Region.
- 1251 **S11 Fig. Correlation between eigenvectors and ancestry proportion estimates**
1252 **from African ancestry specific analyses**
1253 Comparison of ancestry proportion estimates from Admixture model with $K=5$ and
1254 some Principal Components (PCs) in admixed samples from **DS5**. **(A)** PC3 vs
1255 Western African ancestry proportions (yellow in **S10 Fig**). **(B)** PC3 vs Bantu-
1256 influenced ancestry proportions (blue in **S10 Fig**). **(C)** PC4 vs Eastern African
1257 ancestry proportions (orange in **S10 Fig**).
- 1258 **S12 Fig. Choice of the number of principal components from Native American**
1259 **ancestry-specific principal component analysis.**
1260 Elbow method to determine which PC minimizes the angle of the curve from the
1261 chart “Percentage of variance explained *versus* Number of PCs”
- 1262 **S13 Fig. Native American ancestry-specific admixture analysis.**

1263 **(A)** Cross-validation scores K from 2 to 10. **(B)** Admixture for $K=2$. **(C)** Admixture
1264 for $K=4$. **(D)** Admixture plots for $K=5$. CYA: Cuyo Region; NEA: Northeastern Region,
1265 NWA: Northwestern Region; PPA: Pampean Region; PTA: Patagonia Region.

1266 **S14 Fig. Comparison of Native American ancestry proportion estimates**
1267 **obtained with admixture on unmasked data ($K=8$) and masked data ($K=3$).**
1268 Unmasked and masked data refers to **DS1** and **DS6**, respectively. Spearman
1269 correlation coefficients and associated P -values are shown. CAN: Central Andes; STF:
1270 Subtropical and Tropical Forests; CCP: Central Chile / Patagonia.

1271 **S15 Fig. Correlation of Native American ancestry proportions and geographic**
1272 **coordinates in Argentina**
1273 **(A)** Central Andes ancestry proportions vs Latitude. **(B)** Central Andes ancestry
1274 proportions vs Longitude. **(C)** Subtropical and Tropical Forests ancestry
1275 proportions vs Latitude. **(D)** Subtropical and Tropical Forests ancestry proportions
1276 vs Longitude. **(E)** Central Chile/Patagonia ancestry proportions vs Latitude. **(F)**
1277 Central Chile/Patagonia ancestry proportions vs Longitude. Linear regression
1278 slopes and the associated P -values are shown.

1279 **S16 Fig. Individual assignment to a Native American ancestry cluster.**
1280 Consensus cluster assignment of South American individuals based on three K -
1281 *means* procedures run with different pairwise distances among individuals. **(Top)**
1282 K -*means* results using Ancestry-Specific PCA and Admixture (ASPCA and AS-
1283 Admixture), and f_3 results to compute pairwise distances. Individuals are
1284 represented as in **Main Fig 5**. Insets: BIC score for number of clusters set to K -*means*
1285 ranging from 2 to 20. In all the three cases, K -means BIC was minimized when
1286 considering 4 clusters. **(Bottom)** Same as top with point colors corresponding to the
1287 assigned cluster.

1288 **S17 Fig. Pairwise genetic affinity among individuals assigned to different**
1289 **groups.**
1290 Boxplots for $1 - f_3(YRI; Ind1, Ind2)$, where $Ind1$ and $Ind2$ are two individuals
1291 belonging to Group 1 and Group 2, respectively. The groups are either the fourth
1292 Native American components identified or ancient Middle Holocene Southern Cone
1293 groups. For clarity, boxplot outliers are not shown. YRI: Yoruba from 1KGP.

1294 **S18 Fig. Graphical visualization of pairwise genetic distances among modern**
1295 **and ancient groups in South America**
1296 **(A)** Neighbor-joining tree from distances of the form $1/f_3(YRI; X, Y)$. USR1 from
1297 Ancient Beringia was used as *outgroup* **(B)** Multidimensional-scaling from distances
1298 of the form $1 - f_3(YRI; X, Y)$. Each group is represented as appearing in the leaf of (A).
1299 USR1 and Anzick-1 were not considered in (B). YRI: Yoruba from 1KGP

1300 **S19 Fig. Genetic affinity of the four Native American components with ancient**
1301 **groups.**
1302 **(A-D)** $f_3(YRI; X; Ancient)$. **(E-H)** $f_4(YRI, X; Ancient\ Beringia, Ancient)$. **(A)** and **(E)**: with
1303 Central Andes (CAN) as X . **(B)** and **(F)**: With Subtropical and Tropical Forests (STF)

as **X**. **(C)** and **(G)**: With Central Chile / Patagonia (CCP) as **X**. **(D)** and **(H)**: With Central Western Argentina (CWA) as **X**. YRI: Yoruba from 1KGP; *Ancient Beringia*: USR1 individual from [66]; **X**: Native American component in Argentina (one plot per **X**). *Ancient*: ancient group labeled on the *x*-axis and represented with a point/color scheme as in **Main Fig 5**. Vertical segments are the +/- 3 standard errors intervals.

S20 Fig. Changes across time of genetic affinity of the four Native American components with ancient groups.

Each point represents a f_4 score of the form $f_4(YRI, X; Ancient\ Beringia, Ancient)$ vs the age of ancient sample, where **X** is one of the four identified Native American components, and *Ancient* is an ancient group. **X** is represented by the color of the square while *Ancient* is represented by the point within the square. The point code of the ancient samples is shown in **Main Fig 5**. *Ancient Beringia*: USR1 individual from [66]. **(A)** f_4 vs age of ancient samples from Southern Cone. **(B)** f_4 vs age of ancient samples from Andes. **(C)** f_4 vs age of ancient samples from Southern Cone considering correction for both f_4 and age. **(D)** f_4 vs age of ancient samples from Andes considering correction for both f_4 and age. Linear regression slopes and the associated *P*-values are shown. CAN: Central Andes; STF: Subtropical and Tropical Forests; CCP: Central Chile / Patagonia; CWA: Central Western Argentina; YRI: Yoruba from 1KGP

S21 Fig. Comparison of genetic affinity of an ancient group to a Native American component relative to another.

$f_4(YRI, Ancient, X, Y)$ where **X** and **Y** are two of the four identified Native American components (one plot per **X**-**Y** combination), and *Ancient* is ancient group labeled on the *x*-axis and represented with a point/color scheme as in **Main Fig 5**. Vertical segments are the +/- 3 standard errors intervals. Note this setting for f_4 statistics is symmetrical when switching **X** and **Y**.

S22 Fig. Comparison of genetic affinity of a Native American component to another relative to an ancient group.

$f_4(Ancient, X; Y, YRI)$ where **X** and **Y** are two of the four identified Native American components (one plot per **X**-**Y** combination), and *Ancient* is ancient group labeled on the *x*-axis and represented with a point/color scheme as in **Main Fig 5**. Vertical segments are the +/- 3 standard errors intervals. CAN: Central Andes; STF: Subtropical and Tropical Forests; CCP: Central Chile / Patagonia; CWA: Central Western Argentina; YRI: Yoruba from 1KGP

S23 Fig. Removing admixed Santiago de Chile individuals to compute *F*-statistics does not affect the results.

Admixed individuals from Santiago de Chile were removed to perform the analyses presented in this figure. **(A)** $f_3(Target; S1, S2)$ to test for treeness; **(B)** $f_4(YRI, Target; S1, S2)$ to test whether Target shares more ancestry with S1 or S2; **(C)** $f_3(YRI; CWA; Ancient)$; **(D)** $f_4(YRI, CWA; Ancient\ Beringia, Ancient)$; **(E)** $f_4(Ancient, CCP; CWA, YRI)$; **(F)** $f_4(Ancient, CWA; CCP\ YRI)$; **(G)** $f_4(YRI, Ancient; CWA, CCP)$. CCP: Central Chile / Patagonia.

S24 Fig. Schematic routes for the main population arrivals in the Southern Cone.

Each arrow represents one of the four components discussed throughout the article. Neither the time and place of the splits among these components nor gene flow among them have been addressed in this study.

S25 Fig. Admixture analyses in DS2 to define European, African and Native American reference individuals for local ancestry analyses

(A) Cross-validation scores from $K=3$ to $K=10$. **(B)** Admixture for $K=7$

S26 Fig. Admixture analyses in DS3 to define European, African and Native American reference individuals for local ancestry analyses

(A) Cross-validation scores from $K=3$ to $K=10$. **(B)** Admixture for $K=5$

S27 Fig. Comparison of admixture and RFMix ancestry proportion estimates in DS2p and DS3p.

(A-C) For DS2p: Argentinean samples from the present study with reference panel that consists in 1KGP individuals from Africa, Europe and America [42] and Chilean individuals from [37]. Native American, European and African ancestry proportions estimates with RFMix vs with Admixture with $K=7$. **(D-F) For DS3p:** Argentinean samples from [31] with reference panel that consists in 1KGP individuals from Africa, Europe and America [42]. Native American, European and African ancestry proportions estimates with RFMix vs with Admixture with $K=5$.

S28 Fig. Consistency of the masking procedure applied to DS2p and DS3p.

We compared the percentage of variants with same ancestry ditypes in DS2p and DS3p for American admixed individuals from the 1000 Genomes. Project.

Supplementary Table Captions

S1 Table. Sample information.

Sampling location, gender, uniparental lineages, Affymetrix QC metrics, color and point coding used for plots.

S2 Table. Data sets (DS) analyzed throughout the article.

S3 Table. Ancestry proportion estimates in a worldwide context.

Ancestry proportion estimates from Admixture analyses with $K=3$ and $K=8$ at the worldwide level. The column names describe the labels attributed to each ancestry detecting for both Admixture analyses, as well as the hexadecimal code for the color used to represent it in the corresponding admixture plot. The columns "Point", "Color" and "cex" list the graphical parameters used to represent each individual in the different plots throughout the article.

S4 Table. Native American cluster assignation.

Individual Native American cluster assignation is given for each of the three K -means procedures and for the consensus call (columns "F3", "PCA", "Admixture" and "Consensus"). The ancestry proportion estimates from Admixture analyses with $K=3$ on the masked data for Native American ancestry are also provided. The column names explicit the labels attributed to each ancestry detecting for both Admixture analyses as well as the hexadecimal code for the color. For admixed individuals (from the present study and [31]), the "Population" and "Region" columns list the locality and province, respectively, while for Native American population (from [37,38]) the "Population" and "Region" columns list the ethnic and main ethnic groups, respectively.

S5 Table. F -statistics

(A) $f_3(\text{Target}; S1, S2)$ only for comparisons including Native American components
(B) $f_4(YRI, \text{Target}; S1, S2)$ only for comparisons including Native American components
(C) $f_3(YRI; X, Y)$ only for comparisons including Ancient Beringia, Mixe and Native American components
(D) $f_3(YRI; X, Y)$ where X and Y can be either an ancient group or one of the four Native American components.
(E) $f_4(YRI, X; \text{Ancient Beringia}, \text{Ancient})$ only for comparisons including between a Native American components (X) and an ancient group (Ancient).
(F) $f_4(YRI, \text{Ancient}, X, Y)$ only for comparisons including two Native American components (X, Y) and an ancient group (Ancient).
(G) $f_4(\text{Ancient}, X; Y, YRI)$ only for comparisons including two Native American components (X, Y) and an ancient group (Ancient). YRI: Yoruba from 1KGP

Fig. 1. Sample ocations from the present study.

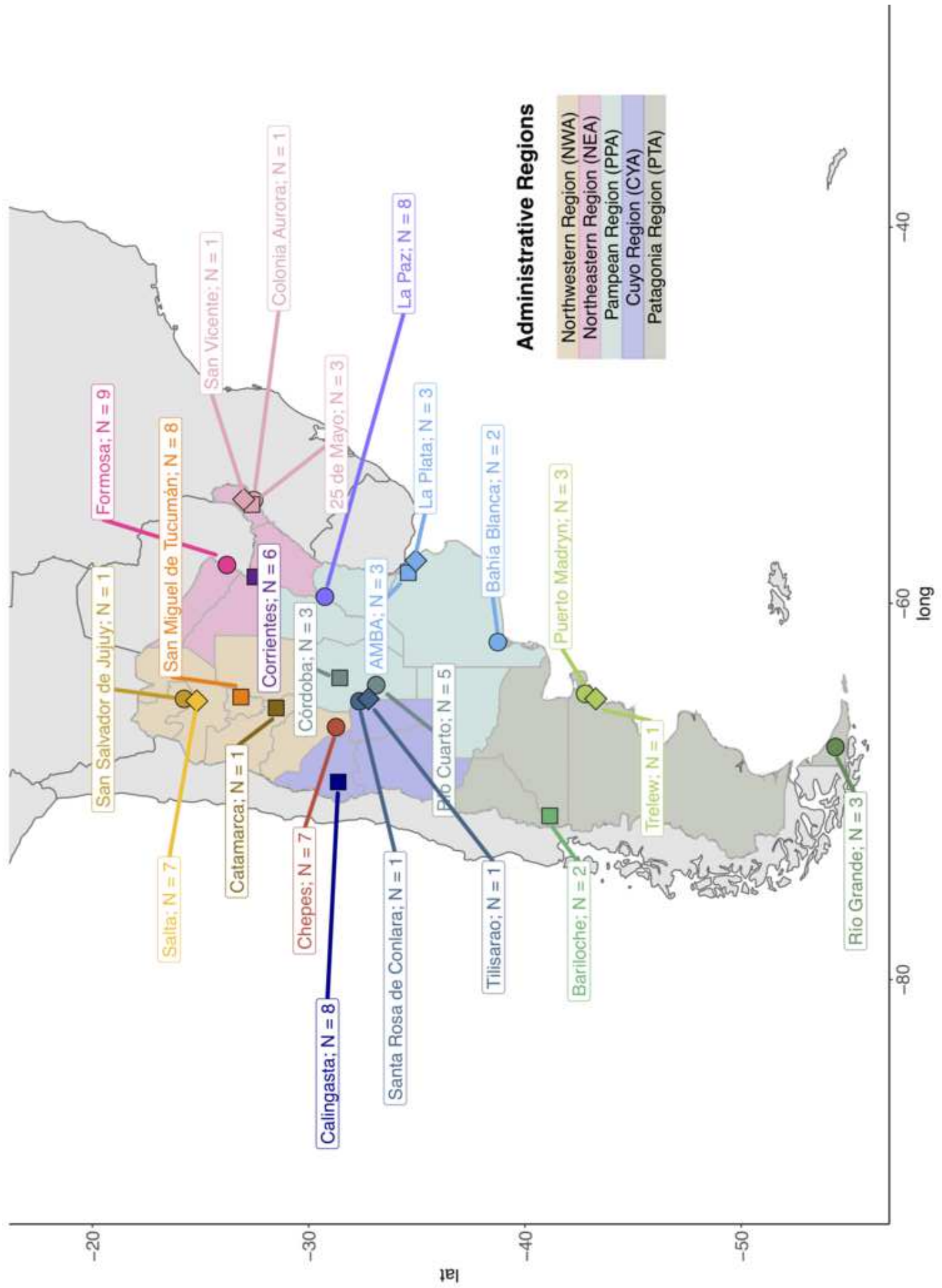


Fig. 2. Admixture analyses in a worldwide context.

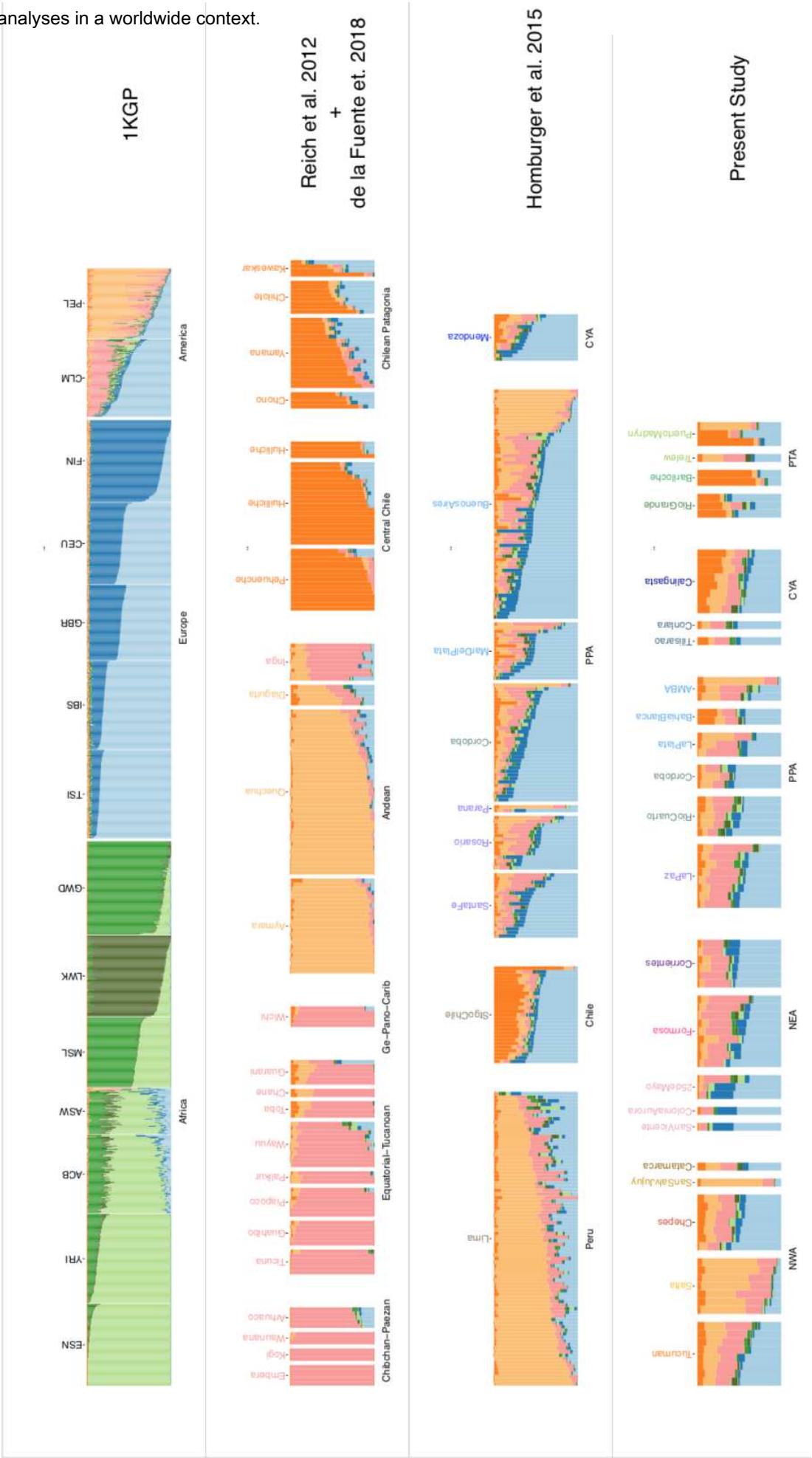


Fig. 3. European ancestry specific principal component analysis.

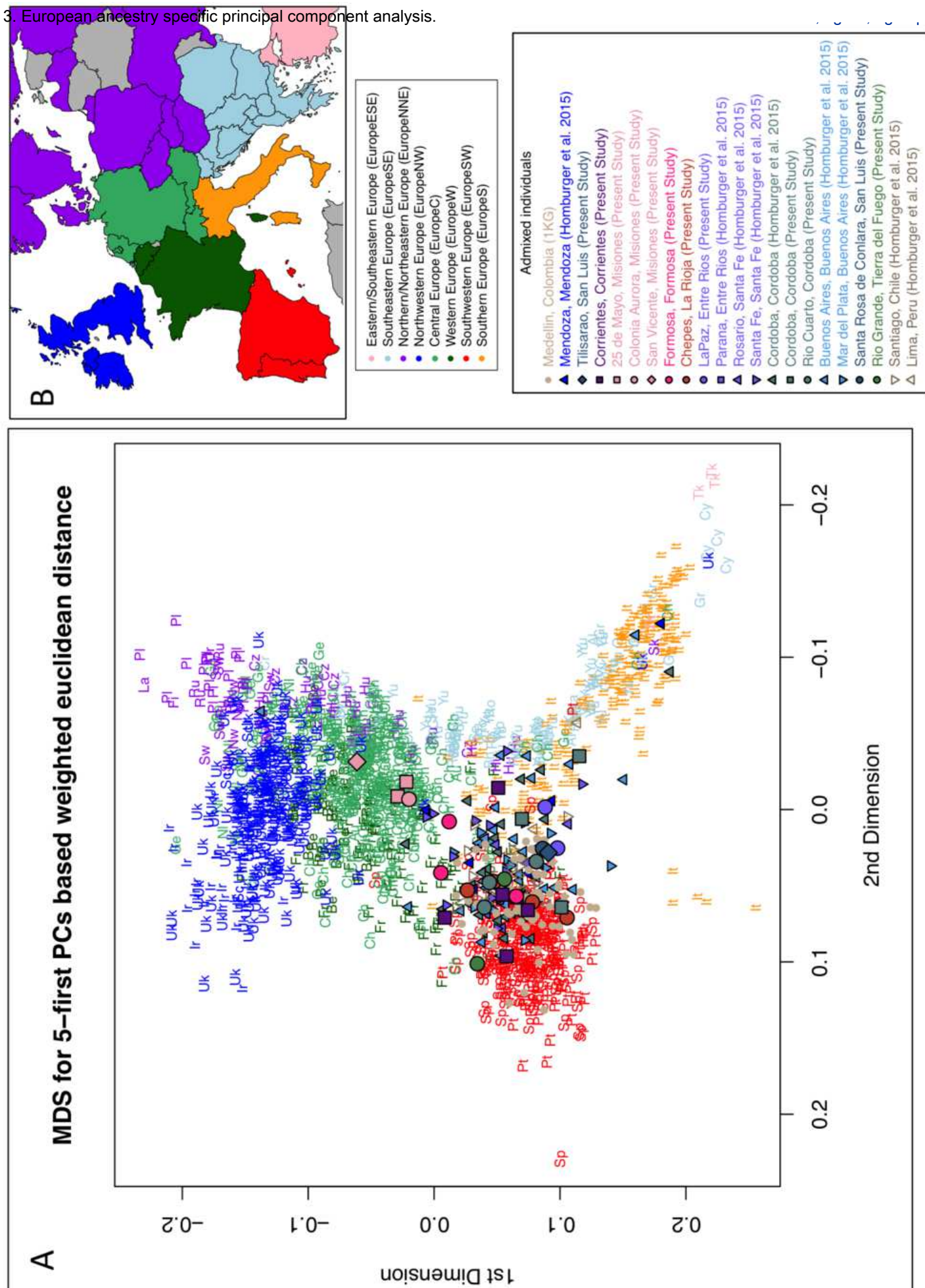


Fig. 4. African ancestry-specific admixture analysis.

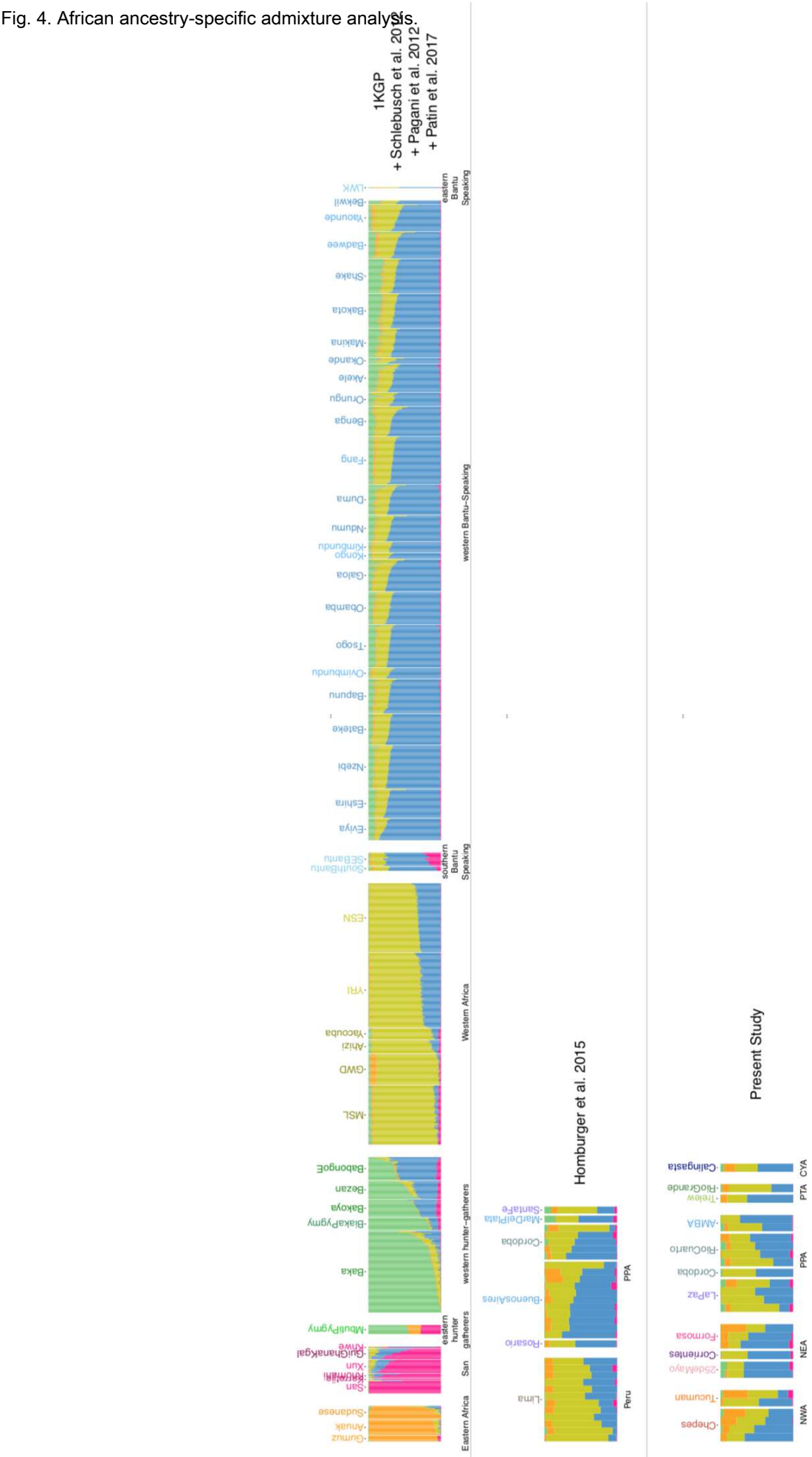


Fig. 5. Native American ancestry-specific principal component analysis.

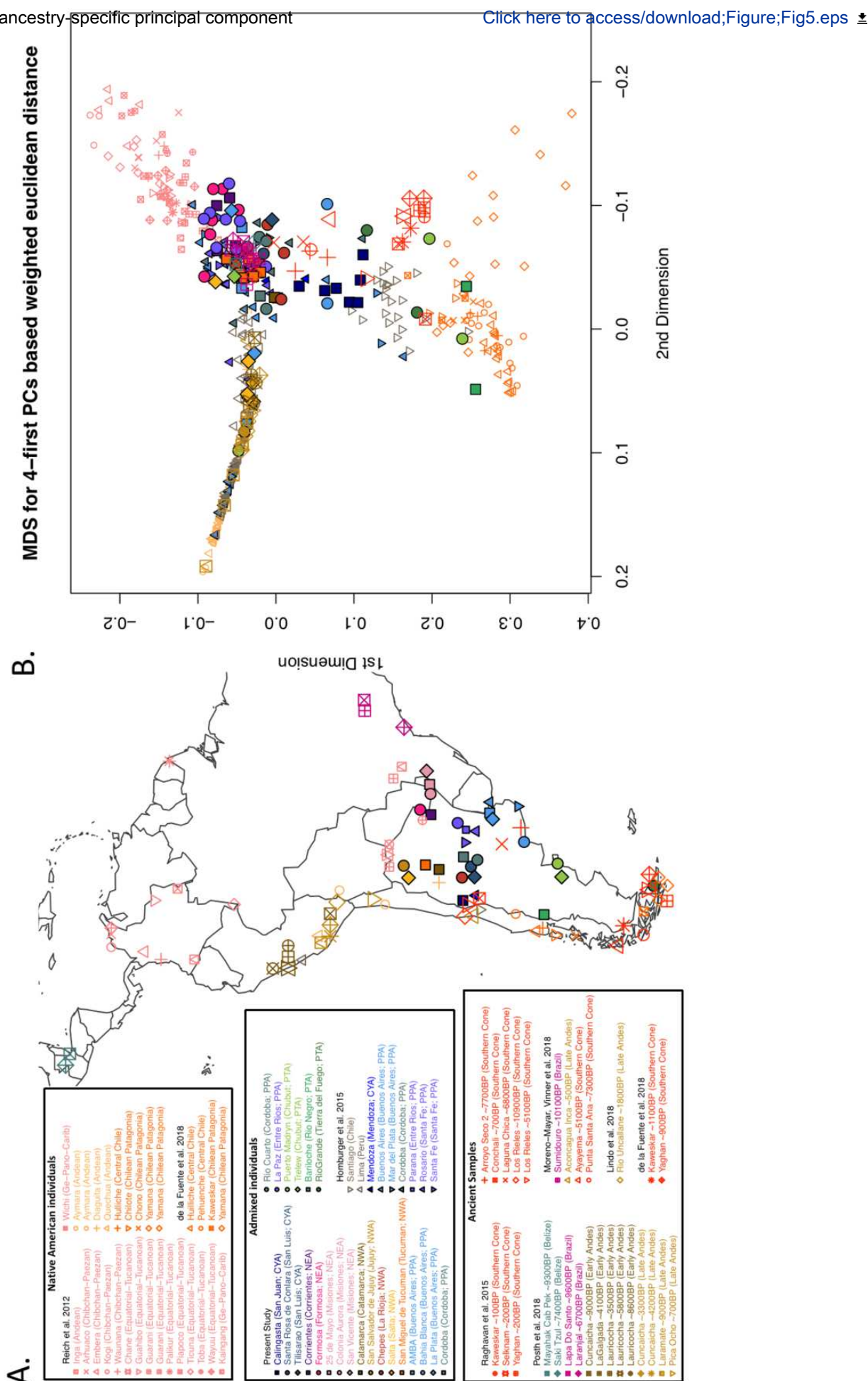




Fig. 7. Relationship among the four Native American groups identified.

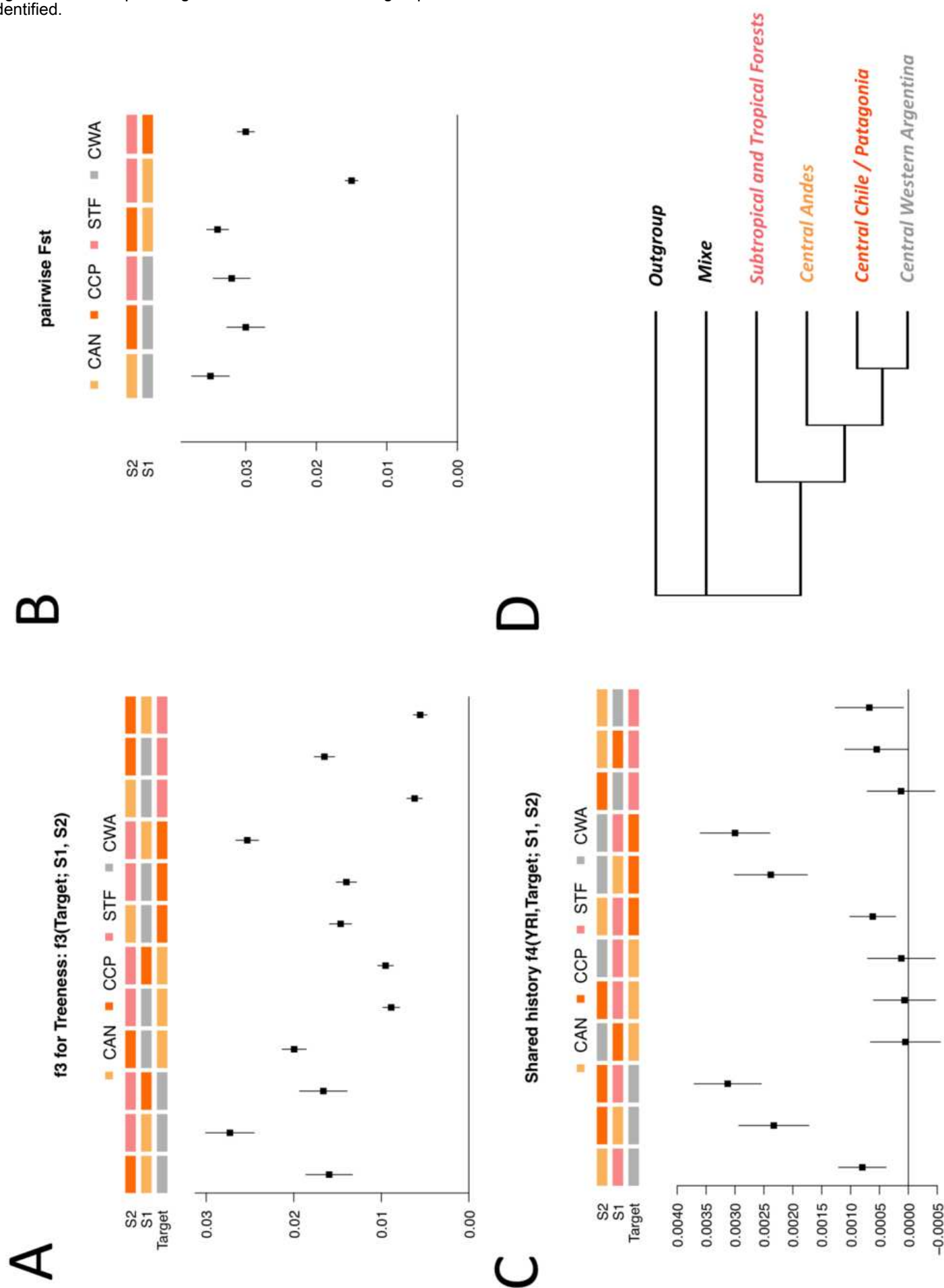


Fig. 8. Changes across time of genetic affinity of the four Native American with ancient samples.

