Reconstructing Native American Population History

Article *in* Nature · July 2012

DOI: 10.1038/nature11258 · Source: PubMed

64 authors, including:

Desmond Campbell
University of Glasgow
76 PUBLICATIONS **3,948** CITATIONS
SEE PROFILE

Stephane Mazieres
Aix-Marseille Université
97 PUBLICATIONS **1,447** CITATIONS
SEE PROFILE

Nicolas Ray
University of Geneva
239 PUBLICATIONS **10,203** CITATIONS
SEE PROFILE

Maria V Parra
Tecnologico de Antioquia
77 PUBLICATIONS **2,571** CITATIONS
SEE PROFILE

LETTER

doi:10.1038/nature11258

Reconstructing Native American population history

David Reich[1,2], Nick Patterson[2], Desmond Campbell[3,4], Arti Tandon[1,2], Stéphane Mazieres[3,5], Nicolas Ray[6], Maria V. Parra[3,7], Winston Rojas[3,7], Constanza Duque[3,7], Natalia Mesa[3,7], Luis F. García[7], Omar Triana[7], Silvia Blair[7], Amanda Maestre[7], Juan C. Dib[8], Claudio M. Bravi[3,9], Graciela Bailliet[9], Daniel Corach[10], Tábita Hünemeier[3,11], Maria Cátira Bortolini[11], Francisco M. Salzano[11], María Luiza Petzl-Erler[12], Victor Acuña-Alonzo[13], Carlos Aguilar-Salinas[14], Samuel Canizales-Quinteros[15,16], Teresa Tusié-Luna[15], Laura Riba[15], Maricela Rodríguez-Cruz[17], Mardia Lopez-Alarcón[17], Ramón Coral-Vazquez[18], Thelma Canto-Cetina[19], Irma Silva-Zolezzi[20]†, Juan Carlos Fernandez-Lopez[20], Alejandra V. Contreras[20], Gerardo Jimenez-Sanchez[20]†, Maria José Gómez-Vázquez[21], Julio Molina[22], Ángel Carracedo[23], Antonio Salas[23], Carla Gallo[24], Giovanni Poletti[24], David B. Witonsky[25], Gorka Alkorta-Aranburu[25], Rem I. Sukernik[26], Ludmila Osipova[27], Sardana A. Fedorova[28], René Vasquez[29], Mercedes Villena[29], Claudia Moreau[30], Ramiro Barrantes[31], David Pauls[32], Laurent Excoffier[33,34], Gabriel Bedoya[7], Francisco Rothhammer[35], Jean-Michel Dugoujon[36], Georges Larrouy[36], William Klitz[37], Damian Labuda[30], Judith Kidd[38], Kenneth Kidd[38], Anna Di Rienzo[25], Nelson B. Freimer[39], Alkes L. Price[2,40] & Andrés Ruiz-Linares[3]

The peopling of the Americas has been the subject of extensive genetic, archaeological and linguistic research; however, central questions remain unresolved[1–5]. One contentious issue is whether the settlement occurred by means of a single[6–8] migration or multiple streams of migration from Siberia[9–15]. The pattern of dispersals within the Americas is also poorly understood. To address these questions at a higher resolution than was previously possible, we assembled data from 52 Native American and 17 Siberian groups genotyped at 364,470 single nucleotide polymorphisms. Here we show that Native Americans descend from at least three streams of Asian gene flow. Most descend entirely from a single ancestral population that we call 'First American'. However, speakers of Eskimo–Aleut languages from the Arctic inherit almost half their ancestry from a second stream of Asian gene flow, and the Na-Dene-speaking Chipewyan from Canada inherit roughly one-tenth of their ancestry from a third stream. We show that the initial peopling followed a southward expansion facilitated by the coast, with sequential population splits and little gene flow after divergence, especially in South America. A major exception is in Chibchan speakers on both sides of the Panama isthmus, who have ancestry from both North and South America.

The settlement of the Americas occurred at least 15,000 years ago through Beringia, a land bridge between Asia and America that existed during the ice ages[1–5]. Most analyses of Native American genetic diversity have examined single loci, particularly mitochondrial DNA or the Y chromosome, and some interpretations of these data model the settlement of America as a single migratory wave from Asia[6–8]. We assembled native population samples from Canada to the southern tip of South America, genotyped them on single nucleotide polymorphism (SNP) microarrays, and merged our data with six other data sets. The combined data set consists of 364,470 SNPs genotyped in 52 Native American populations (493 samples; Fig. 1a and Supplementary Table 1), 17 Siberian populations (245 samples; Supplementary Fig. 1 and Supplementary Table 2) and 57 other populations (1,613 samples) (Supplementary Notes).

A complication in studying Native American genetic history is admixture with European and African immigrants since 1492. Cluster analysis[16] shows that many of the samples we examined have some non-native admixture (an average of 8.5%; Fig. 1b and Supplementary Tables 1 and 3). This admixture is a challenge for learning about the historical relationships among the populations, and to address this complication we used three independent approaches. First, we restricted analyses to 163 Native Americans from 34 populations without evidence of admixture (Supplementary Notes). Second, we subtracted the expected contribution of European and African ancestry to the statistics we used to learn about population relationships (Supplementary Notes). Third, we inferred the probability of non-native ancestry at each genomic segment and 'masked' segments with more than a negligible probability of this ancestry (Fig. 1b,

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [3]Department of Genetics, Evolution and Environment, University College London WC1E 6BT, UK. [4]Department of Psychiatry and Centre for Genomic Sciences, The University of Hong Kong, Pokfulam, Hong Kong SAR. [5]Anthropologie Bio-culturelle, Droit, Ethique et Santé (ADES), UMR 7268, Aix-Marseille Université/CNRS/EFS, Marseille 13344, France. [6]Institute for Environmental Sciences, and Forel Institute, University of Geneva, Geneva 1227, Switzerland. [7]Universidad de Antioquia, Medellín, Colombia. [8]Fundación Salud para el Trópico, Santa Marta, Colombia. [9]Instituto Multidisciplinario de Biología Celular (CCT La Plata-CONICET, CICPCA), 1900 La Plata, Argentina. [10]Servicio de Huellas Digitales Genéticas and CONICET, Universidad de Buenos Aires, Argentina. [11]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio Grande do Sul, Porto Alegre 91501-970, Brazil. [12]Departamento de Genética, Universidade Federal do Paraná, Curitiba 81531-980, Brazil. [13]National Institute of Anthropology and History, México City 06100, México. [14]Departamento de Endocrinología y Metabolismo, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, México City 14100, México. [15]Unidad de Biología Molecular y Medicina Genómica, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán/Universidad Nacional Autónoma de México, México City 14000, México. [16]Departamento de Biología, Facultad de Química, Universidad Nacional Autónoma de México, México City 04510, México. [17]Unidad de Investigación Médica en Nutrición, Hospital de Pediatría, CMNSXXI, Instituto Mexicano del Seguro Social, México City 06720, México. [18]Sección de Posgrado, Escuela Superior de Medicina del Instituto Politécnico Nacional, México City 11340, México. [19]Laboratorio de Biología de la Reproducción, Departamento de Salud Reproductiva y Genética, Centro de Investigaciones Regionales, Mérida Yucatán 97000, México. [20]Instituto Nacional de Medicina Genómica, México City 14610, México. [21]Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León 66451, México. [22]Centro de Investigaciones Biomédicas de Guatemala, Ciudad de Guatemala, Guatemala. [23]Instituto de Ciencias Forenses, Universidade de Santiago de Compostela, Fundación de Medicina Xenómica (SERGAS), CIBERER, Santiago de Compostela, Galicia 15782, Spain. [24]Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima 15102, Peru. [25]Department of Human Genetics, University of Chicago, Chicago 60637, USA. [26]Laboratory of Human Molecular Genetics, Institute of Molecular and Cellular Biology, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia. [27]Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia. [28]Department of Molecular Genetics, Yakut Research Center of Complex Medical Problems and North-East Federal University, Yakutsk, Sakha (Yakutia) 677010, Russia. [29]Instituto Boliviano de Biología de la Altura, Universidad Autonoma Tomás Frías, Potosí, Bolivia. [30]Département de Pédiatrie, Centre de Recherche du CHU Sainte-Justine, Université de Montréal, Montréal, Quebec H3T 1C5, Canada. [31]Escuela de Biología, Universidad de Costa Rica, San José, Costa Rica. [32]Center for Human Genetic Research, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02114, USA. [33]Computational and Molecular Population Genetics Laboratory, Institute of Ecology and Evolution, University of Bern, 3012 Bern, Switzerland. [34]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [35]Instituto de Alta Investigación, Universidad de Tarapacá, Programa de Genética Humana ICBM Facultad de Medicina Universidad de Chile and Centro de Investigaciones del Hombre en el Desierto, Arica 1001236, Chile. [36]Anthropologie Moléculaire et Imagerie de Synthèse, CNRS UMR 5288, Université Paul Sabatier Toulouse III, Toulouse 31000, France. [37]School of Public Health, University of California, Berkeley, California 94720, USA. [38]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA. [39]Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, California 90095, USA. [40]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. †Present addresses: BioAnalytical Science Department Nestec Ltd, Nestlé Research Center, 1000 Lausanne, Switzerland (I.S.-Z.); Global Biotech Consulting Group, México City 09010, México (G.J.-S.).

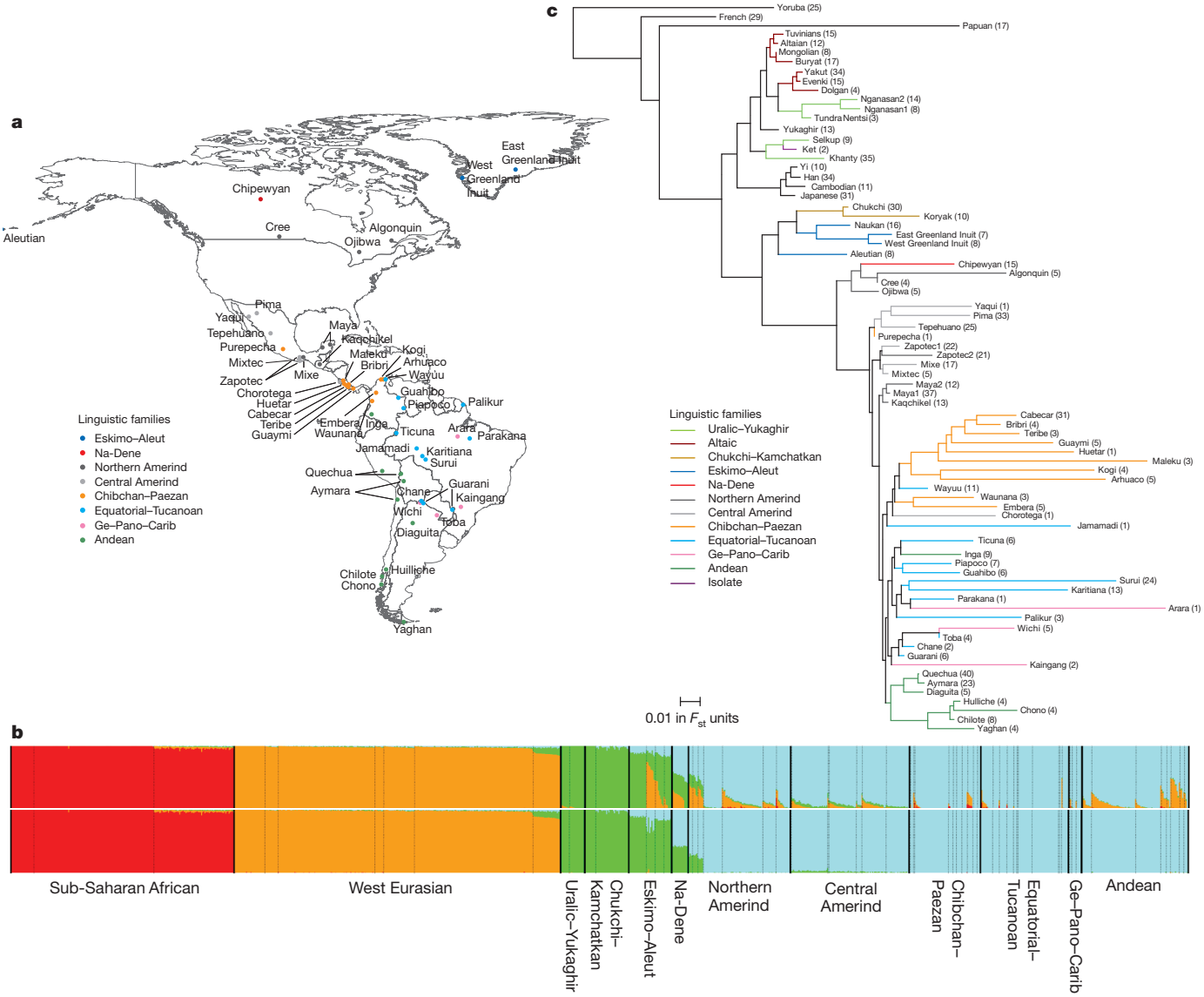

Figure 1 | Geographic, linguistic and genetic overview of 52 Native American populations. **a**, Sampling locations of the populations, with colours corresponding to linguistic groups. **b**, Cluster-based analysis ($k = 4$) using ADMIXTURE shows evidence of some West-Eurasian-related and sub-Saharan-African-related ancestry in many Native Americans before masking (top), but little afterwards (bottom). Thick vertical lines denote major linguistic groupings, and thin vertical lines separate individual populations. **c**, Neighbour-joining tree based on F_{st} distances relating Native American to selected non-American populations (sample sizes in parentheses). Native American and Siberian data were analysed after masking, but consistent trees were obtained on a subset of completely unadmixed samples (Supplementary Fig. 3). Some populations have evidence for substructure, and we represent these as two different groups (for example Maya1 and Maya2).

Supplementary Notes and Supplementary Fig. 2). Our inferences from these three approaches are concordant (Supplementary Figs 3 and 4).

We built a tree (Fig. 1c) using F_{st} distances between pairs of populations, which broadly agrees with geography and linguistic categories[17] (trees based on masked and unmasked data were similar; Supplementary Fig. 3). An early split separates Asians from Native Americans and extreme northeastern Siberians (Chukchi, Naukan, Koryak), which is consistent with studies that have identified pan-American variants shared with some northeastern Siberians[6,7,10,18]. Eskimo–Aleut speakers and far-northeastern Siberians form a cluster that is separated from other Native American populations by a long internal branch. Within America the tree shows a series of splits in an approximate north–south sequence beginning with the Arctic, followed by northern North America, northern/central and southern Mexico and lower Central America/Colombia, and ending in three South American clusters (the Andes, the Chaco region and eastern South America). This pattern of splits is consistent with a north–south population expansion, an inference that is also supported by the negative correlation between heterozygosity and distance from the Bering Strait ($r = -0.48$, $P = 0.007$). This correlation increases if we use 'least cost distances' that consider the coasts as facilitators of migration[19–21], and persists if we exclude four Native North American populations with ancestry from later streams of Asian gene flow (Supplementary Notes and Supplementary Fig. 5).

Trees provide a simplified model of history that does not accommodate the possibility of gene flow after population separation. Circumstantial evidence that some Native American populations may not fit a simple tree comes from cluster analysis, which infers Siberian-related ancestry in some northern North Americans (Fig. 1b), and from single-locus studies that have identified genetic variants shared between Eurasia and North America that are absent from South America[11,22,23]. The advent of genome-wide data sets has allowed the development of a formal four-population test for whether sets of four populations are consistent with a tree. This test is robust to the

Table 1 | Native Americans descend from at least three streams of Asian gene flow

Population groupings tested	*P* value for this many Asian streams being enough to explain the data			Minimum number of streams of Asian gene flow needed to explain the data
	1	2	3	
East Greenland Inuit/West Greenland Inuit/First American	$<10^{-9}$	0.64	1	2
East Greenland Inuit/Aleutian/First American	$<10^{-9}$	0.57	1	2
West Greenland Inuit/Aleutian/First American	$<10^{-9}$	0.41	1	2
Chipewyan/East Greenland Inuit/First American	$<10^{-9}$	0.02	1	3
Chipewyan/West Greenland Inuit/First American	$<10^{-9}$	0.006	1	3
Chipewyan/Aleutian/First American	$<10^{-9}$	0.03	1	3
Saqqaq/East Greenland Inuit/First American	$<10^{-9}$	6×10^{-6}	1	3
Saqqaq/West Greenland Inuit/First American	$<10^{-9}$	2×10^{-6}	1	3
Saqqaq/Aleutian/First American	$<10^{-9}$	0.17	1	2
Saqqaq/Chipewyan/First American	$<10^{-9}$	0.29	1	2
Saqqaq/Eskimo–Aleut/Chipewyan/First American	$<10^{-9}$	8×10^{-6}	0.27	3

We use the method described in Supplementary Notes to test formally whether specified groupings of Native American populations are consistent with descending from one, two or three streams of gene flow from Asia. We use 'First American' to refer to a pool of 43 populations from Meso-America southward, and 'Eskimo–Aleut' to refer to a pool of East and West Greenland Inuit and Aleuts. We test either three or four population groupings (when there are three groupings, the maximum number of streams we can reject is two, and so the *P* value for three streams is always 1). At least two streams of Asian gene flow are required to explain all rows ($P<10^{-9}$). The Chipewyan, Eskimo–Aleut and First Americans can only be jointly explained by at least three streams. Analysis of the Saqqaq Palaeo-Eskimo (using about sixfold fewer SNPs than for the other analyses) show that the Asian ancestry in this individual has a component that is different from that in First Americans and Greenland Inuit, but indistinguishable from the Chipewyan.

ascertainment bias affecting SNP arrays[24]. For each of the 52 Native American populations in turn, we tested the hypothesis that they conform to the tree: ((test population, southern Native American), (outgroup1, outgroup2)) for 45 pairs of ten Asian outgroups. We used a Hotelling *T*-test to evaluate whether all four-population test f_4 statistics of this form are consistent with the expectation of zero (Supplementary Notes). The test is not significant for 47 populations, which is consistent with their stemming from the same, presumably first, wave of American settlement; we call this ancestry 'First American' (Table 1). In contrast, four populations from northern North America show highly significant evidence of ancestry from additional streams of gene flow from Asia, subsequent to the initial peopling of America, which we confirm through the Hotelling *T*-test and a complementary test (Supplementary Notes): East Greenland Inuit ($P<10^{-9}$), West Greenland Inuit ($P<10^{-9}$), Aleutian Islanders ($P=9\times10^{-5}$) and Chipewyan ($P<10^{-9}$). The recently sequenced genome of a 4,000-year-old Saqqaq Palaeo-Eskimo from Greenland[25] also has evidence of ancestry that is distinct from more southern Native Americans ($P=2\times10^{-9}$) (Supplementary Notes).

Examination of the values of the f_4 statistics allows us to infer the minimum number of gene flow events from Asia into America consistent with the data. Each stream of gene flow is expected to produce a distinct vector of f_4 statistics, constituting a 'signature' of how the ancestral migrating population relates to present-day Asian populations. By finding the minimum number of vectors whose linear combinations are necessary to produce the vector observed in each population, we infer that a minimum of three gene flow events from Asia are necessary to explain the data from all Native American populations jointly, including the Saqqaq Palaeo-Eskimo (Supplementary Notes). These three episodes correspond to First American ancestry (distributed throughout the Americas) and to two additional streams of gene flow detected in a subset of northern North Americans (East Greenland Inuit, West Greenland Inuit, Aleutian Islanders, Chipewyan and Saqqaq). Table 1 shows that f_4 statistics in the Inuit and Aleutian islanders are consistent with deriving the non-First-American portions of their ancestry from the same later stream of Asian gene flow, providing support for deep shared ancestry between these linguistically linked groups[12,26]. The Na-Dene-speaking Chipewyan have a different pattern of f_4 statistics from Eskimo–Aleut speakers, implying that they descend at least in part from a separate stream of Asian gene flow ($P<10^{-9}$ for comparisons with the Greenland Inuit; Table 1). This is consistent with the hypothesis that Na-Dene languages mark a distinct migration from Asia[9,17]. Because we only have data from one Na-Dene-speaking group, an important direction for future work will be to test whether the distinct Asian ancestry that we detect in the Chipewyan is a shared signature throughout Na-Dene speakers. Finally, the Saqqaq[25] have a vector of f_4 statistics consistent with that in the Chipewyan, raising the possibility that the Saqqaq and Chipewyan both carry genetic material from the same later stream of Asian gene flow into the Americas, postdating the First American migration (Supplementary Notes).

To develop an explicit model for the settlement of the Americas, we used the admixture graph (AG) framework[24]. AGs are generalizations of trees that accommodate the possibility of a limited number of unidirectional gene flow events. They are powerful tools for learning about history because they make predictions about the values of *f*-statistics (such as f_4) that can be used to test the fit of a proposed model[24] (Supplementary Notes). Figure 2 presents an AG relating selected Native American and Old World populations that is a good fit to the data in the sense that none of the *f*-statistics predicted by the

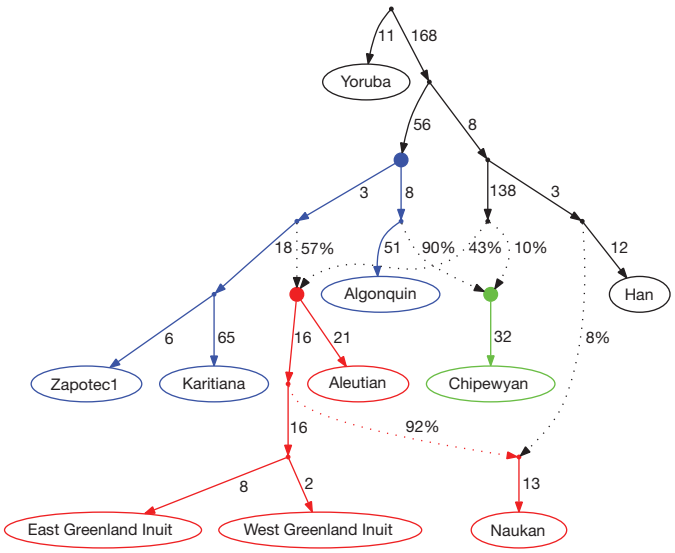

Figure 2 | Distinct streams of gene flow from Asia into America. We present an AG that gives no evidence of being a poor fit to the data and is consistent with three streams of Asian gene flow into America. Solid points indicate inferred ancestral populations, drift on each lineage is given in units proportional to $1{,}000\times F_{st}$, and mixture events (dotted lines) are denoted by the percentage of ancestry. The Asian lineage leading to First Americans is the most deeply diverged, whereas the Asian lineages leading to Eskimo–Aleut speakers and the Na-Dene-speaking Chipewyan are more closely related and descend from a common Siberian ancestral population that is a sister group to the Han. The inferred ancestral populations are indicated by filled circles, and the lineages descending from them are coloured: First American (blue), ancestors of the Na-Dene-speaking Chipewyan (green), and Eskimo–Aleut (red). The model also infers a migration of people related to Eskimo–Aleut speakers across the Bering Strait, thus bringing First American genes to Asia (the Naukan are shown, but the Chukchi show a similar pattern; Supplementary Notes).

model are more than three standard errors from what is observed. This supports the hypothesis of three deep lineages in Native Americans: the Asian lineage leading to First Americans is the most deeply diverged, whereas the Asian lineages leading to Eskimo–Aleut speakers and the Na-Dene-speaking Chipewyan are more closely related and descend from a putative Siberian ancestral population more closely related to Han (Fig. 2). We also arrive at the finding that Eskimo–Aleut populations and the Chipewyan derive large proportions of their genomes from First American ancestors: an estimated 57% for Eskimo–Aleut speakers, and 90% in the Chipewyan, probably reflecting major admixture events of the two later streams of Asian migration with the First Americans that they encountered after they arrived (Supplementary Notes). The high proportion of First American ancestry explains why Eskimo–Aleut and Chipewyan populations cluster with First Americans in trees like that in Fig. 1c despite having some of their ancestry from later streams of Asian migration, and explains the observation of some genetic variants that are shared by all Native Americans but are absent elsewhere[6,7,10,18]. We also infer back-migration of populations related to the Eskimo–Aleut from America into far-northeastern Siberia (we obtain an excellent fit to the data when we model the Naukan and coastal Chukchi as mixtures of groups related to the Greenland Inuit and Asians (Fig. 2 and Supplementary Notes)). This explains previous findings of pan-American alleles also in far-northeastern Siberia[6,7,10,18].

We next used AGs to develop a model for the history of populations who derive all their ancestry from the First American migration, with no ancestry from subsequent streams of Asian gene flow. Figure 3 presents an AG we built for 16 selected Native American populations and two outgroups, which is a good fit to the data in that the largest $|Z|$-score for a difference between the observed and predicted f-statistics is 3.2 from among the 11,781 statistics we tested (Supplementary Notes) (The AG of Fig. 3 used masked data; however, a consistent set of relationships is inferred for unadmixed samples (Supplementary Fig. 4).) This model provides a greatly improved statistical fit to the data compared with the tree of Fig. 1c and leads to several novel inferences. First, a relatively large fraction of South American populations fit the AG without a need for admixture events, which we speculate reflects a history of limited gene flow among these populations since their initial divergence. In contrast, only a small fraction of Meso-American populations fit into the AG, which could reflect either a higher rate of migration among neighbouring groups or our denser sampling in Meso-America allowing us to detect more subtle gene flow events. Second, some Meso-American populations have experienced very little genetic drift since divergence from the common ancestral population with South Americans (adding up the genetic drifts along the relevant edges of Fig. 3, we infer $F_{st} = 0.014$ between the Zapotec and a hypothetical population ancestral to all of Central and South America), suggesting that effective population sizes in Meso-America have been relatively large since settlement of the region. Third, the model infers three admixture events consistent with geographic locations and linguistic affiliations (Supplementary Notes). The Inga have both Amazonian and Andean ancestry, which is consistent with their speaking a Quechuan language but living in the eastern Andean slopes of Colombia and thus interacting with groups in the neighbouring Amazonian lowlands. The Guarani stem from two distinct strands of ancestry within eastern South America. The most striking admixture event is in the Costa Rican Cabecar (Fig. 3) and other Chibchan-speaking populations (Supplementary Notes) from the Isthmo-Colombian area. One of the lineages that we detect in these populations occurs definitively within the radiation of South American populations, and so the presence of these populations in lower Central America suggests that there was reverse gene flow across the Panama isthmus after the initial settlement of South America. There has been controversy about whether Chibchan speakers of lower Central America represent direct descendants of the first settlers in the region or more recent migration across the isthmus, and our results support

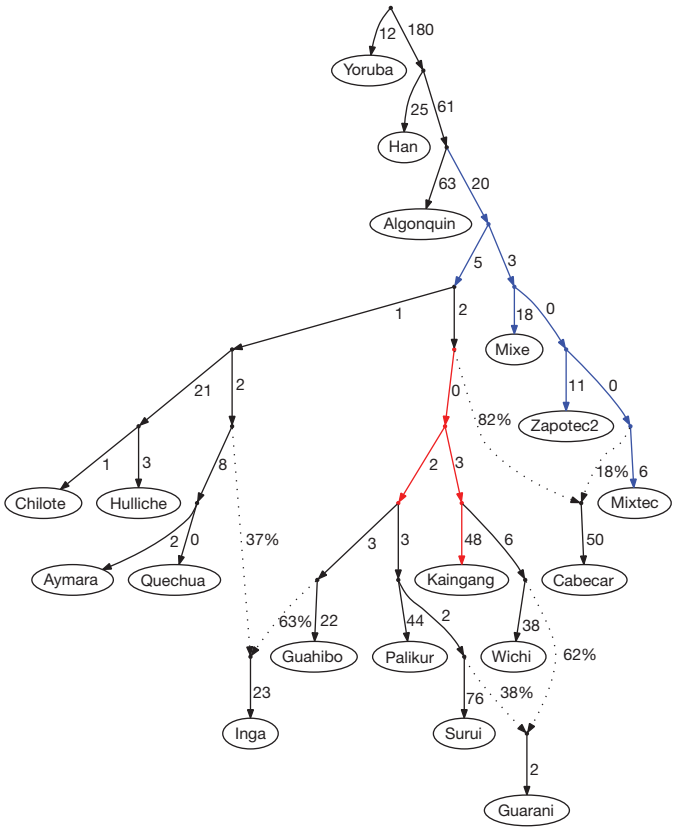

Figure 3 | A model fitting populations of entirely First American ancestry. We show an AG depicting the relationships between 16 selected Native American populations with entirely First American ancestry along with two outgroups (Yoruba and Han). The Colombian Inga are modelled as a mixture of Andean and Amazonian ancestry. The Paraguayan Guarani are fitted as a mixture of separate strands of ancestry from eastern South America. The Central American Cabecar are modelled as a mixture of strands of ancestry related to South Americans and to North Americans, supporting back-migration from South into Central America. The colouring of edges indicates alternative insertion points for the admixing lineages leading to the Cabecar that produce a similar fit to the data in the sense that the χ^2 statistic is within 3.84 of the AG shown. The red colouring shows that the South American lineage contributing to the Cabecar split off after the divergence of the Andean populations, and the blue colouring shows that the other lineage present in the Cabecar diverged before the separation of Andeans. Estimated admixture proportions are shown along the dotted lines, and lineage-specific drift estimates are in units proportional to $1{,}000 \times F_{st}$.

the view that more recent migration has contributed most of these populations' ancestry[27].

This is the most comprehensive survey of genetic diversity in Native Americans so far. Our analyses show that the great majority of Native American populations—from Canada to the southern tip of Chile—derive their ancestry from a homogeneous 'First American' ancestral population, presumably the one that crossed the Bering Strait more than 15,000 years ago[6–8]. We also document at least two additional streams of Asian gene flow into America, allowing us to reject the view that all present-day Native Americans stem from a single migration wave[6–8], and supporting the more complex scenarios proposed by some other studies[9–15]. In particular, the three distinct Asian lineages we detect—'First American', 'Eskimo–Aleut' and a separate one in the Na-Dene-speaking Chipewyan—are consistent with a three-wave model proposed[9] mostly on the basis of dental morphology and a controversial interpretation of the linguistic data. However, our analyses also document extensive admixture between First Americans and the subsequent streams of Asian migrants, which was not predicted by that model, such that Eskimo–Aleut speakers and the Chipewyan

derive more than half their ancestry from First Americans. Further insights into Native American history will benefit from the application of analyses similar to those performed here to whole-genome sequences and to data from the many admixed populations in the Americas that do not self-identify as native[28–30].

METHODS SUMMARY

The DNA samples we analysed were collected over several decades. For each sample we verified that informed consent was obtained consistent with studies of population history and that institutional approval had been obtained in the country of collection. Ethical oversight and approval for this project was provided by the National Health Service National Research Ethics Service, Central London committee (reference no. 05/Q0505/31). The data set is based on merging Illumina SNP array data newly generated for this study (including 273 Native American samples) with data from six other studies. We applied stringent data curation and validation procedures to the merged data set. We used local ancestry inference software to identify genome segments in each Native American and Siberian sample without evidence of recent European or African admixture, and created a data set that masked segments of potentially non-native origin. Most analyses are performed on the masked data set; however, we confirmed major inferences on a subset of 163 Native American samples that had no evidence of European or African admixture. We used model-based clustering and neighbour-joining trees to obtain an overview of population relationships, and then tested whether proposed sets of four populations were consistent with having a simple tree relationship using the four-population test, which we generalized by means of a Hotelling T-test. We analysed the correlation in allele frequency differences across populations to infer the minimum number of gene flow events that occurred between Asia and America. We fitted the patterns of correlation in allele frequency differences to proposed models of history—AGs—that can incorporate population splits and mixtures.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 September 2011; accepted 25 May 2012.
Published online 11 July 2012.

1. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, 1994).
2. Meltzer, D. J. *First Peoples in a New World: Colonizing Ice Age America* (Univ. of California Press, 2009).
3. Goebel, T., Waters, M. R. & O'Rourke, D. H. The late Pleistocene dispersal of modern humans in the Americas. *Science* **319,** 1497–1502 (2008).
4. Dillehay, T. D. Probing deeper into first American studies. *Proc. Natl Acad. Sci. USA* **106,** 971–978 (2009).
5. O'Rourke, D. H. & Raff, J. A. The human genetic history of the Americas: the final frontier. *Curr. Biol.* **20,** R202–R207 (2010).
6. Tamm, E. *et al.* Beringian standstill and spread of Native American founders. *PLoS ONE* **2,** e829 (2007).
7. Kitchen, A., Miyamoto, M. M. & Mulligan, C. J. A three-stage colonization model for the peopling of the Americas. *PLoS ONE* **3,** e1596 (2008).
8. Fagundes, N. J. *et al.* Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. *Am. J. Hum. Genet.* **82,** 583–592 (2008).
9. Greenberg, J. H., Turner, C. G. & Zegura, S. L. The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence. *Curr. Anthropol.* **27,** 477–497 (1986).
10. Lell, J. T. *et al.* The dual origin and Siberian affinities of Native American Y chromosomes. *Am. J. Hum. Genet.* **70,** 192–206 (2002).
11. Bortolini, M. C. *et al.* Y-chromosome evidence for differing ancient demographic histories in the Americas. *Am. J. Hum. Genet.* **73,** 524–539 (2003).
12. Volodko, N. V. *et al.* Mitochondrial genome diversity in arctic Siberians, with particular reference to the evolutionary history of Beringia and Pleistocenic peopling of the Americas. *Am. J. Hum. Genet.* **82,** 1084–1100 (2008).
13. Ray, N. *et al.* A statistical evaluation of models for the initial settlement of the American continent emphasizes the importance of gene flow with Asia. *Mol. Biol. Evol.* **27,** 337–345 (2010).
14. de Azevedo, S. *et al.* Evaluating microevolutionary models for the early settlement of the New World: the importance of recurrent gene flow with Asia. *Am. J. Phys. Anthropol.* **146,** 539–552 (2011).
15. Perego, U. A. *et al.* Distinctive Paleo-Indian migration routes from Beringia marked by two rare mtDNA haplogroups. *Curr. Biol.* **19,** 1–8 (2009).
16. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).
17. Ruhlen, M. *A Guide to the World's Languages* (Stanford Univ. Press, 1991).
18. Schroeder, K. B. *et al.* A private allele ubiquitous in the Americas. *Biol. Lett.* **3,** 218–223 (2007).
19. Ray, N. PATHMATRIX: a geographical information system tool to compute effective distances among samples. *Mol. Ecol. Notes* **5,** 177–180 (2005).
20. Wang, S. *et al.* Genetic variation and population structure in native Americans. *PLoS Genet.* **3,** e185 (2007).
21. Yang, N. N. *et al.* Contrasting patterns of nuclear and mtDNA diversity in Native American populations. *Ann. Hum. Genet.* **74,** 525–538 (2010).
22. Brown, M. D. *et al.* mtDNA haplogroup X: an ancient link between Europe/Western Asia and North America? *Am. J. Hum. Genet.* **63,** 1852–1861 (1998).
23. Karafet, T. M. *et al.* Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am. J. Hum. Genet.* **64,** 817–831 (1999).
24. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461,** 489–494 (2009).
25. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463,** 757–762 (2010).
26. Balter, M. Archaeology. The peopling of the Aleutians. *Science* **335,** 158–161 (2012).
27. Cooke, R. Prehistory of native Americans on the Central American land bridge: Colonization, dispersal, and divergence. *J. Archaeol. Res.* **13,** 129–187 (2005).
28. Wang, S. *et al.* Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet.* **4,** e1000037 (2008).
29. Bryc, K. *et al.* Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc. Natl Acad. Sci. USA* **107** (Suppl 2), 8954–8961 (2010).
30. Wall, J. D. *et al.* Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. *Mol. Biol. Evol.* **28,** 2231–2237 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the volunteers who provided the samples that made this study possible. We thank E. D. Ruiz for assistance in the collection involving the Mixtec, Zapotec and Mixe; and P. Herrera for assistance in the collection involving the Quechua; A. Carnevale, M. Crawford, M. Metspalu, F. C. Nielsen, X. Soberon, R. Villems and E. Willerslev for facilitating sharing of data from Mexican, Siberian and Arctic populations; C. Stevens and A. Crenshaw for assistance with genotyping; and P. Bellwood, D. Bolnick, K. Bryc, J. Diamond, T. Dillehay, R. Gonzalez-José, M. Hammer, J. Hill, B. Kemp, S. LeBlanc, D. Meltzer, P. Moorjani, A. Moreno-Estrada, B. Pakendorf, J. Pickrell, M. Ruhlen, D. G. Smith, M. Stoneking, N. Tuross and A. Williams for critiques and discussions. Support was provided by National Institutes of Health grants NS043538 (A.R.-L.), NS037484 and MH075007 (N.B.F.), GM079558 (A.D.), GM079558-S1 (A.D.), GM057672 (K.K.K. and J.R.K.), and HG006399 (D.R., N.P. & A.L.P); by a Biotechnology and Biological Sciences Research Council grant BB/1021213/1; by a National Science Foundation HOMINID grant BCS-1032255 (D.R. and N.P.); by a Canadian Institutes of Health Research grant (D.L.); by a Universidad de Antioquia CODI grant (G.B.); by a Fondo de Investigación Sanitaria grant PS 09/2368 (A.C.); by a Ministerio de Ciencia e Innovacion grant SAF2011-26983 (A.S.); by a Wenner-Gren Foundation grant ICRG-65 (A.D. and R.S.); by Russian Foundation for Basic Research grants 06-04-048182 (R.S.) and 02-06-80524a (L.O.); by a Siberian Branch Russian Academy of Sciences field grant (L.O.); by a Centre National de la Recherche Scientifique Programme Interdisciplinaire de Recherche Amazonie grant (J.-M.D.); and by startup funds from Harvard Medical School (D.R.) and the Harvard School of Public Health (A.L.P.).

Author Contributions D.R., N.B.F., A.L.P. and A.R-L. conceived the project. D.R., N.P., D.C., A.T., S.M., N.R. and A.R-L. performed analyses. D.R. and A. R.-L. wrote the paper with input from all the co-authors. A.R.-L. assembled the sample collection, directed experimental work and coordinated the study. All other authors contributed to the collection of samples and data.

Author Information The data analysed here are available for non-profit research on population history under an inter-institutional data access agreement with the Universidad de Antioquia, Colombia; queries regarding data access should be sent to A.R.-L. (a.ruizlin@ucl.ac.uk). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.R. (reich@genetics.med.harvard.edu) or A.R.-L. (a.ruizlin@ucl.ac.uk).

METHODS

DNA samples. The samples analysed here were collected for previous studies over several decades. We reviewed the documentation available for each population to confirm that all samples were collected with informed consent encompassing genetic studies of population history. Institutional approval for use of each set of samples in such research was obtained before this study in the country of collection. Approval for this study was also provided by the National Research Ethics Service, Central London REC 4 (reference no. 05/Q0505/31).

Genotyping. All samples were genotyped by using Illumina arrays, and the data set analysed here is the result of merging data from seven different sources (Supplementary Notes). The genotyping conducted specifically for this study was performed at the Broad Institute of Harvard and Massachusetts Institute of Technology, with the exception of ten Chipewyan samples that were genotyped at McGill University (no systematic differences were observed between these and the five Chipewyan samples genotyped at the Broad Institute). Supplementary Table 3 specifies details for each of the 493 Native American samples. A total of 419 samples were genotyped from genomic DNA, and 74 from whole-genome-amplified material prepared using the Qiagen REPLI-g midi kit.

Data curation. We required more than 95% genotyping completeness for each SNP and sample. We merged the data specifically obtained for this study with six other data sets. We further removed samples that were outliers in principal-component analysis relative to others from their group, showed an excess rate of heterozygotes in comparison with the expected rate from the allele frequencies in the population, or had evidence of being a second-degree relative or closer to another sample in the study (Supplementary Notes). Genetic analyses summarized in the Supplementary Notes found substructure in some populations (Maya, Zapotec and Nganasan); we use labels such as 'Maya1' and 'Maya2' to indicate the subgroups.

Masking of genomic segments containing non-Native American ancestry. For each Native American individual, we used HAPMIX[31] to model their haplotypes with two ancestral panels: first, 'Old World' populations (a pool of 408 Europeans and 130 West Africans) and second, 'Native' populations, a pool of all Native American and Siberian populations. Haplotype phase in the ancestral panel, which is necessary for HAPMIX, was determined by phasing both pools of samples together with Beagle[32]. We masked genome segments that had an expected number of more than 0.01 non-Native American chromosomes according to HAPMIX, thus retaining only segments with an extremely high nominal probability of being homozygous for native ancestry. Multiple analyses reported in Supplementary Information indicate that our masking procedure produces inferences about history that are consistent with those based on unadmixed samples.

Population structure analysis, F_{st} and neighbour-joining tree. We used EIGENSOFT to perform PCA and compute pairwise population F_{st} (ref. 33). Clustering was performed with ADMIXTURE[16]. A neighbour-joining[34] tree based on F_{st} was built with POWERMARKER[35].

Linguistic categories. We used Greenberg's classification[17,36]. We considered using alternative classifications; however, others (for example that in ref. 37) do not propose links between languages at a deep enough level to compare with genetic relationships on a continent-wide scale.

Correlating geography with population diversity. Euclidean distances from the Bering Strait (64.8° N, 177.8° E) and the location of each population (Supplementary Table 1) were calculated by using great arc distances based on a Lambert azimuthal equal-area projection. Least-cost distances between the same points were computed with PATHMATRIX[19], which allowed us to build a spatial cost map incorporating the coastal outline of the Americas. We compared the following coastal/inland relative costs: 1:2, 1:5, 1:10, 1:20, 1:30, 1:40, 1:50, 1:100, 1:200, 1:300, 1:400 and 1:500. We computed a Pearson correlation coefficient between heterozygosity for each population and their least-cost distance from the Bering Strait (Supplementary Notes).

Documentation of at least three streams of gene flow from Asia to America. We used the four-population test to assess whether proposed sets of four populations were consistent with a tree. For each of 52 test populations, we assessed their consistency with deriving from the same Asian source population as southern Native Americans by studying statistics of the form f_4 (southern Native American, test population; outgroup1, outgroup2), where the two outgroups are the 45 ($=10 \times 9/2$) possible pairs of ten Asian outgroups (Han Chinese and nine Siberian populations with at least ten samples each, and not including the Naukan and Chukchi whom we showed to have some First American ancestry as a result of back-migration across the Bering Strait, making them inappropriate as outgroups (Supplementary Notes)). We applied a Hotelling T-test to assess whether the ensemble of all possible f_4 statistics was consistent with zero after taking into account their correlation structure, resulting in a single hypothesis test for whether the test population was consistent with having the same relationship to the panel of Asian populations as the set of southern Native American samples used as a reference group. We also generalized this test by studying the matrix of all f_4 statistics simultaneously and computing statistics that measured whether the f_4 statistics seen in proposed sets of Native American populations were consistent with deriving from a specified number of Asian migrations. In Supplementary Notes we show that if there have been N distinct streams of gene flow from Asia into the Americas, then the matrix of all possible f_4 statistics can have rank no more than $N-1$ (ignoring sampling noise). The case $N=1$ reduces to calculating a Hotelling T^2 statistic. We also developed a likelihood ratio test, generalizing the Hotelling T-test, to evaluate the statistical evidence for larger values of N, allowing us to estimate the minimum number of exchanges between Asia and America that are needed to explain the genetic data.

Admixture graphs. We used the AG framework[24] to fit models of population separation followed by mixture to the data. An AG makes predictions about the correlations in allele frequency differentiation statistics (f-statistics) that will be observed between all pairs, triples and quadruples of populations[24], and these can be compared with the observed values (along with a standard error from a Block Jackknife) to test hypotheses about population relationships (Supplementary Notes). We do not have a formal goodness-of-fit test for whether a given AG fits the data correcting for the number of hypotheses tested and number of degrees of freedom, but use two approximations. First, we examine individual f-statistics, searching for those that are more than three standard errors from expectation indicative of a poor fit. Second, we compute a χ^2 statistic for the match between the observed and predicted f-statistics, taking into account the empirical covariance matrix among the f-statistics computed on the basis of a Block Jackknife. This results in a nominal P value, but it is unclear to us at present whether the empirical covariance matrix that we obtain can be equated with the theoretical covariance matrix that is needed to compute a formal P value. For a fixed graph complexity (number of drift edges and admixture weights), however, we can compare the χ^2 value for different admixture graphs to obtain a formal test for whether some topologies are significantly better fits; this results in the colouring of edges in Fig. 3, which shows alternative insertion points for admixture edges that are equally good fits.

31. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5,** e1000519 (2009).
32. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).
33. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2,** e190 (2006).
34. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).
35. Liu, K. & Muse, S. V. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21,** 2128–2129 (2005).
36. Greenberg, J. H. *Language in the Americas* (Stanford Univ. Press, 1987).
37. Campbell, L. *American Indian languages: the historical linguistics of Native America* (Oxford Univ. Press, 1997).