Supplemental Information

Two ancient human genomes reveal Polynesian ancestry among the indigenous Botocudos of Brazil

A.-S. Malaspinas, O. Lao, H. Schroeder, M. Rasmussen, M. Raghavan, I. Moltke, P. F. Campos, F. Santana Sagredo, S. Rasmussen, V. F. Gonçalves, A. Albrechtsen, M. E. Allentoft, P. L. F. Johnson, M. Li, S. Reis, D. V. Bernardo, M. DeGiorgio, A. T. Duggan, M. Bastos, Y. Wang, J. Stenderup, J. Víctor Moreno-Mayar, S. Brunak, T. Sicheritz-Ponten, E. Hodges, G. J. Hannon, L. Orlando, T. D. Price, J. D. Jensen, R. Nielsen, J. Heinemeier, J. Olsen, C. Rodrigues-Carvalho, M. Mirazón Lahr, W. Neves, M. Kayser, T. Higham, M. Stoneking, S. D. J. Pena, E. Willerslev

Table of Contents

Supplementary 1
Botocudo individuals: sample description

Silvia Reis*, Anna-Sapfo Malaspinas, Murilo Bastos, Morten E. Allentoft, Marta Mirazón Lahr, Claudia Rodrigues-Carvalho*

*to whom correspondence should be addressed (sreis@mn.ufrj.br, claudia@mn.ufrj.br)

Botocudo is a generic name given by the Portuguese settlers to indigenous groups that were not from the Tupi linguistic family and that used lip plates and other forms of body modification. In general, the Botocudos are from the Macro-Jê linguistic family and lived in central-eastern Brazil. They were not a cohesive group but were formed by several cultural groups with linguistic variants and enmity. Examples of such groups are the Naknenuk and Aranã. Most groups are extinct but a few survived, such as the Krenak who live in the state of Minas Gerais.

The National Museum of the Federal University of Rio de Janeiro (Museu Nacional, Universidade Federal do Rio de Janeiro) founded in 1818, owns an important collection of skeletal remains from Botocudo Indians (circa 35 skulls). The first Botocudos were collected in 1874 and were from the "Babilônia Cave" (Netto 1875). A mummy from this cave is still on display today (see exhibition guide of the Museu Nacional ("Múmia em Território Brasileiro" 2013).

Samples from four individuals (MN00013, MN00015, MN00017 and MN00065) from this collection were included in the present study; the individuals will be referred to as MN000xx or Botxx interchangeably (i.e. Bot13 refer to MN00013 and so forth). A preliminary study of a few mtDNA SNPs and a deletion for these individuals was presented in earlier publications (Gonçalves et al. 2010; 2013). Pictures of the four Botocudo skulls included in this study are shown on Supplementary Figure 1.1 and Supplementary Figure 1.2.

Bot15 and Bot17

Bot15 and Bot17 are both males and among the first entries in the museum catalogue, started in 1906 (see a copy of the relevant page of the catalogue in Supplementary Figure 1.3). Although the catalogue entry for these remains does not specify the year of their accession in the museum, current evidence suggests that they were acquired around or before 1883. From 1882 and onwards, the museum organized a large exhibition of anthropological and archaeological material from Brazil (the "Exposição Anthropológica Brazileira"), and a large number of human remains of archaeological and ethnographic origin were displayed in the "Sala Lund". These were individually listed in the guide of the exhibition (Netto, 1882). The exhibition had a strong focus on Brazilian material, with the only exceptions noted in the guide being four skulls of Aymaras from Peru and Bolivia, and two Araucanian skulls, all with cranial modifications, which were most likely presented as comparative material. Although this list compiled by Netto does not contain individual accession numbers, it includes 22 Botocudo cranial (and some post-cranial) remains, *i.e.*, two-thirds of the whole Botocudo collection in the museum today. The original writing is still visible on some skulls (such as the label "Botocudo Rio Doce Minas", see Supplementary Figure 1.1) and traces of early exhibition tags are similar to those used in the anthropological exhibition. Therefore, we are confident that Bot15 and Bot17 were amongst the early, pre-1883, human remains acquired by the museum.

The geographic origin of the Bot15 and Bot17 skulls is stated in the catalogue, as well as written on the skulls themselves (see Supplementary Figure 1.1 and Supplementary Figure 1.3): Rio Doce, Minas Gerais. The inscription on the skulls was probably made in the field when the individuals were collected. The Doce River spans the states of Minas Gerais and Espirito Santo, and although the exact locality from which the skulls originated remains unknown, it can be assumed that both individuals were collected somewhere in the Rio Doce valley.

An upper left first molar from each of Bot15 and Bot17 was sent to the dedicated ancient DNA laboratory at the Centre for GeoGenetics in Copenhagen, Denmark for ancient DNA analysis.

Bot13

Bot13 is a female from the settlement of Mutum (Aldeamento do Mutum). Such settlements were created by the Brazilian government of the time to "civilize" the Indians. Bot13 was most likely a Naknenuk (or Nak-nanuk) indian. The Aldeamento do Mutum was founded in 1859 between the Rio Doce and Rio Mutum Preto (now the city of Baixo Guandu, see Moreira 2008). The diet of individuals from that settlement was based on hunting and fishing from the rivers in the region, supplemented by the products of the small-scale agriculture that was enforced by the government. Since the Aldeamento do Mutum soon became impoverished, the government sent provisions to supply the inhabitants with enough food. The provisions sent to all settlements in Espirito Santo were usually meat (from the nearby cities), beans, tobacco and dried fish (see Marinato 2007, p.109).

The lower right third molar from Bot13 was sent to the Centre for GeoGenetics in Copenhagen, Denmark for genetic analysis.

Bot65

Bot65 is a male originating from Cachoeiro de Itapemirim, Espírito Santo. The remains were found in a cave between June 30 and July 5, 1882, by Antonio Moreira Penedo and Casimiro Ribeiro da Silva, two farmers from Cachoeiro de Itapemirim who came across two caves while hunting. The first cave was found on June 30 and the second on July 5, just 4 meters away. Both contained human bones. In one cave, the farmers excavated two skulls (one from a child) and in the second cave they found one skull, a semi articulated foot with ligaments, and some bones (Noticiário: Achado Importante, 1882).

The skulls and the foot were sent by Deolindo José Vieira Maciel (an engineer) to the Museu Nacional in July 1882, to be part of the Anthropological Exhibition. Thus, Bot65 became part of the museum collection in 1882.

The upper left third molar from Bot65 was sent to the Centre for GeoGenetics in Copenhagen, Denmark for genetic analysis.

A sample from each of the four Botocudo teeth was sent to Oxford and Aarhus University for ^{14}C dating and isotopic analysis, and to the University of Wisconsin for strontium measurements (S8).

Addressing the possibility of mislabeling

The discovery of two individuals with Polynesian ancestry among a museum collection of Native American remains from Brazil obviously represents a surprising and potentially significant finding, as it could represent the first genomic evidence for Polynesians reaching South America. However, before drawing conclusions, several possible ways to explain their presence in the collection have to be explored. The first explanation to consider is that they are, in fact, Polynesian in origin and were mislabeled at the museum. We investigated this possibility by reviewing all the relevant material in the archives of the museum and nothing suggests a mislabeling, or makes those two skulls stand out in terms of how the collection was assembled.

We also reviewed the Polynesian collection. A total of four Polynesian skulls are present in the collection of the Museu Nacional today. Two are from the Marchese Islands (Fatu Hiva), one is from the Chatham Island, and one from New Zealand (Supplementary Figure 1.4). These Polynesian remains were acquired in the 19th century. So far no report or document has been identified in the Museum's archive concerning these skulls, aside from the entry in the catalogue (compiled since 1906). The skulls appear to have been cleaned, and from a taphonomic perspective, are very different from the Botocudo skulls, which have varying degrees of sediment attached to the surface and/or crevices.

If Bot15 and Bot17 were indeed mislabeled, the error would have had to happen before the skulls arrived at the museum, or in the short time period in which they were at the museum but before being shown at the Anthropological Exhibition, as there is no doubt that they were on display as "Botocudos", and analyzed as such. Additionally, at the time, the collection was small and the museum researchers were particularly interested in the fascinating Botocudo Indians, trying to establish their ”racial” characteristics, while working extensively with craniometry. It is interesting that the hand written labels on the two Bot15 and Bot17 skulls have different characteristics (Supplementary Figure 1.1), suggesting that if they were mislabeled, it must then have happened independently by two different people. Another relevant point is that Bot17 has a similar calligraphic inscription as MN00064, another Botocudo skull at the museum, carrying mtDNA haplogroup C, which is found in high frequency in Native American populations (Vanessa F Gonçalves et al. 2010). Since it seems clear that the same person labeled Bot17 and MN00064, it seems unlikely that this person would have mislabeled one of them.

Finally, as is true for all 19th century collections around the world, human crania were acquired by different means – through rare excavations, expeditions, donations, sales and auctions. Thus proving without doubt the exact provenance of any one cranium beyond the point of museum accession is often not possible. However, as in all other studies of material from historical museum collections, if the record of accession is not questionable, the accuracy of the information they portray has to be the baseline for interpretation.

References

Gonçalves, Vanessa F, Flavia C Parra, Higgor Gonçalves-Dornelas, Claudia Rodrigues-Carvalho, Hilton P Silva, and Sergio Dj Pena. 2010. "Recovering Mitochondrial DNA Lineages of Extinct Amerindian Nations in Extant Homopatric Brazilian Populations." *Investigative Genetics* 1 (1): 13. doi:10.1186/2041-2223-1-13.

Gonçalves, Vanessa Faria, Jesper Stenderup, Cláudia Rodrigues-Carvalho, Hilton P. Silva, Higgor Gonçalves-Dornelas, Andersen Líryo, Toomas Kivisild, et al. 2013. "Identification of Polynesian mtDNA Haplogroups in Remains of Botocudo Amerindians from Brazil." *Proceedings of the National Academy of Sciences* (April 1). doi:10.1073/pnas.1217905110. http://www.pnas.org/content/early/2013/03/28/1217905110.

Marinato, Francieli Aparecida. 2007. "Índios Imperiais: Os Botocudos, Os Militares e a Colonização Do Rio Doce (Espírito Santo, 1824-1845)". Dissertação (mestrado), ES, Brasil: Centro de Ciências Humanas e Naturais.

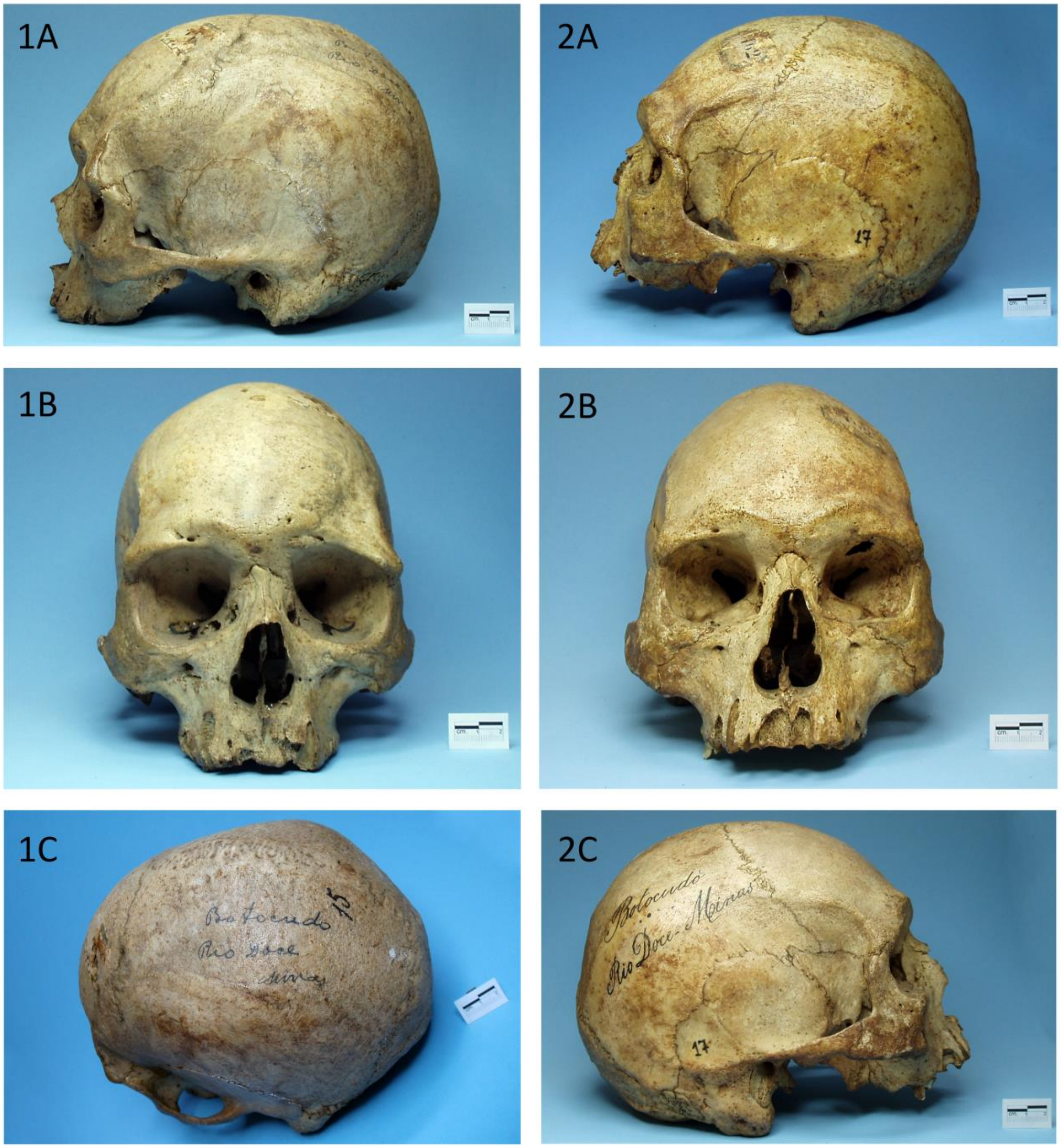
Moreira, Vânia Maria Losada. 2008. "História, Evangelização e Política Indigenista: a Missão Do Mutum. In: 26 Reuniaão Brasileira de Antropologia, 2008". CD virtual da 26 RBA. (Des)igualdades Na Diversidade. Porto Seguro, Bahia, Brasil.

"Múmia Em Território Brasileiro." 2013. Guia de Visitação Ao Museu Nacional. Rio de Janeiro, Brasil: Museu Nacional. http://www.museunacional.ufrj.br/guiaMN/Guia/paginas/6/indiabras.htm.

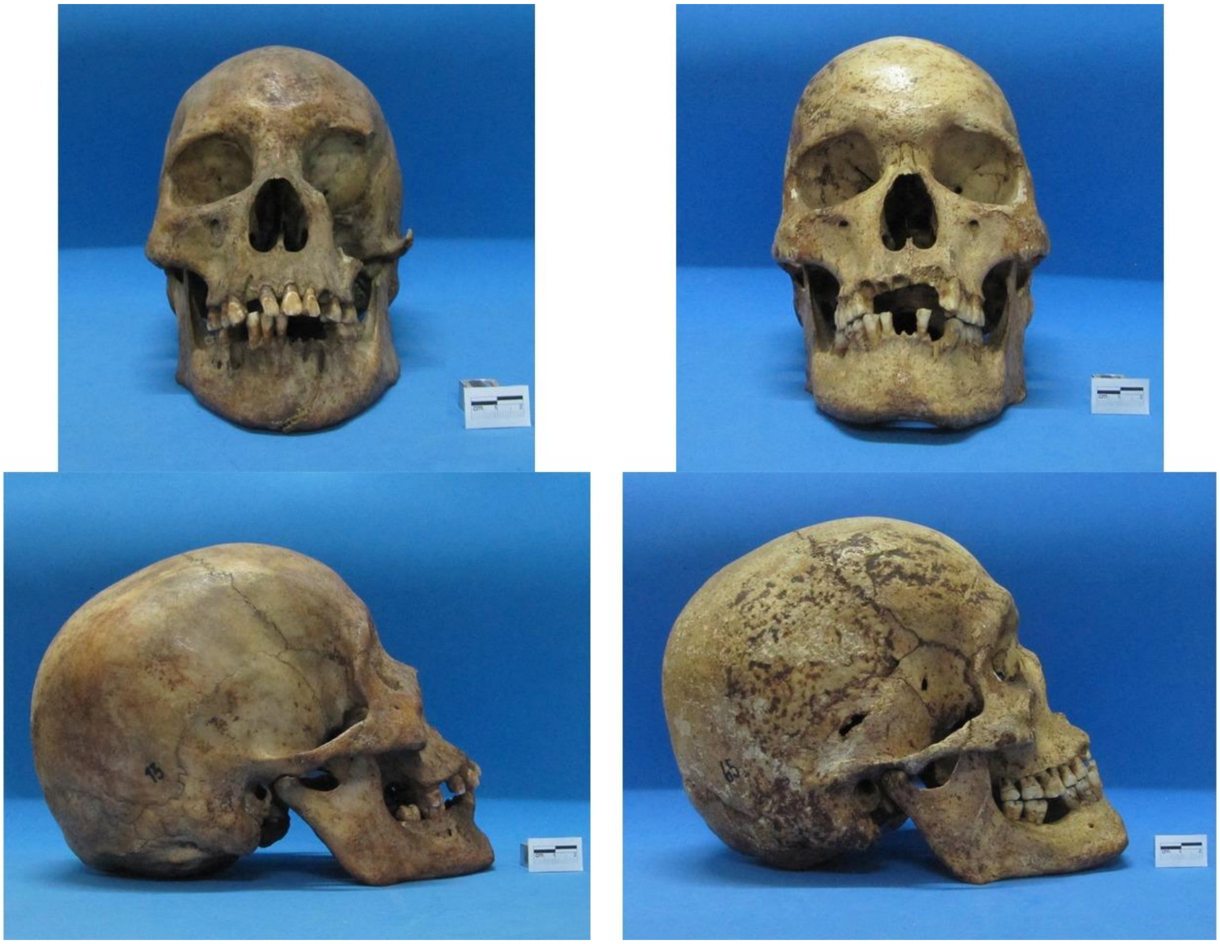
Netto, Ladislau. 1875. "Relatório Anual Do Museu Nacional". Rio de Janeiro, Brasil.

———. 1882. "Guia Da Exposição Anthropológica Brazileira Realisada Pelo Museu Nacional Do Rio de Janeiro a 29 de Julho de 1882." Rio de Janeiro.

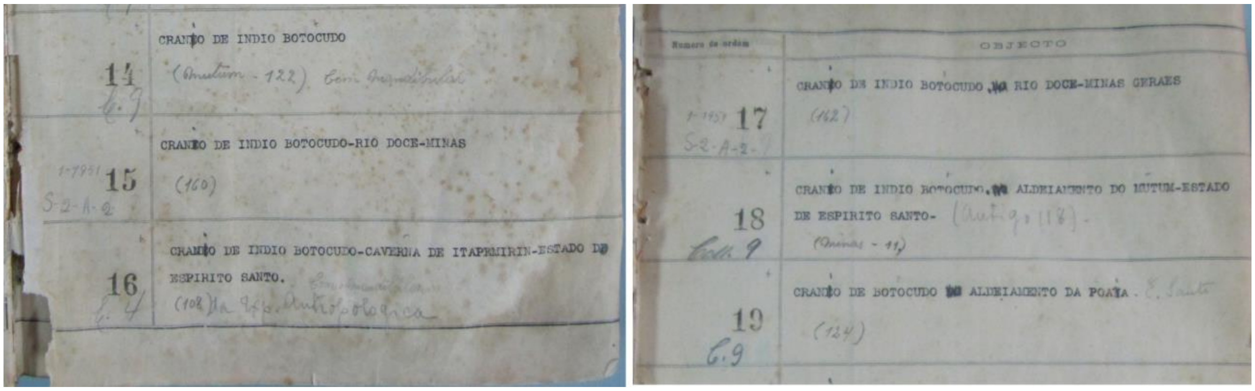
Noticiário: Achado Importante. 1882. *O Cachoeirano*, July 16, Anno V, n. 29.



Supplementary Figure 1.1 Bot15 and Bot17 skulls. Photographs of Bot15 (left) and Bot17 (right) including lateral (1A and 2A) and frontal (1B and 2B) views of the two skulls. 1C and 2C show a close up of the labeling of Bot15 skull (left) and Bot17 skull (right).



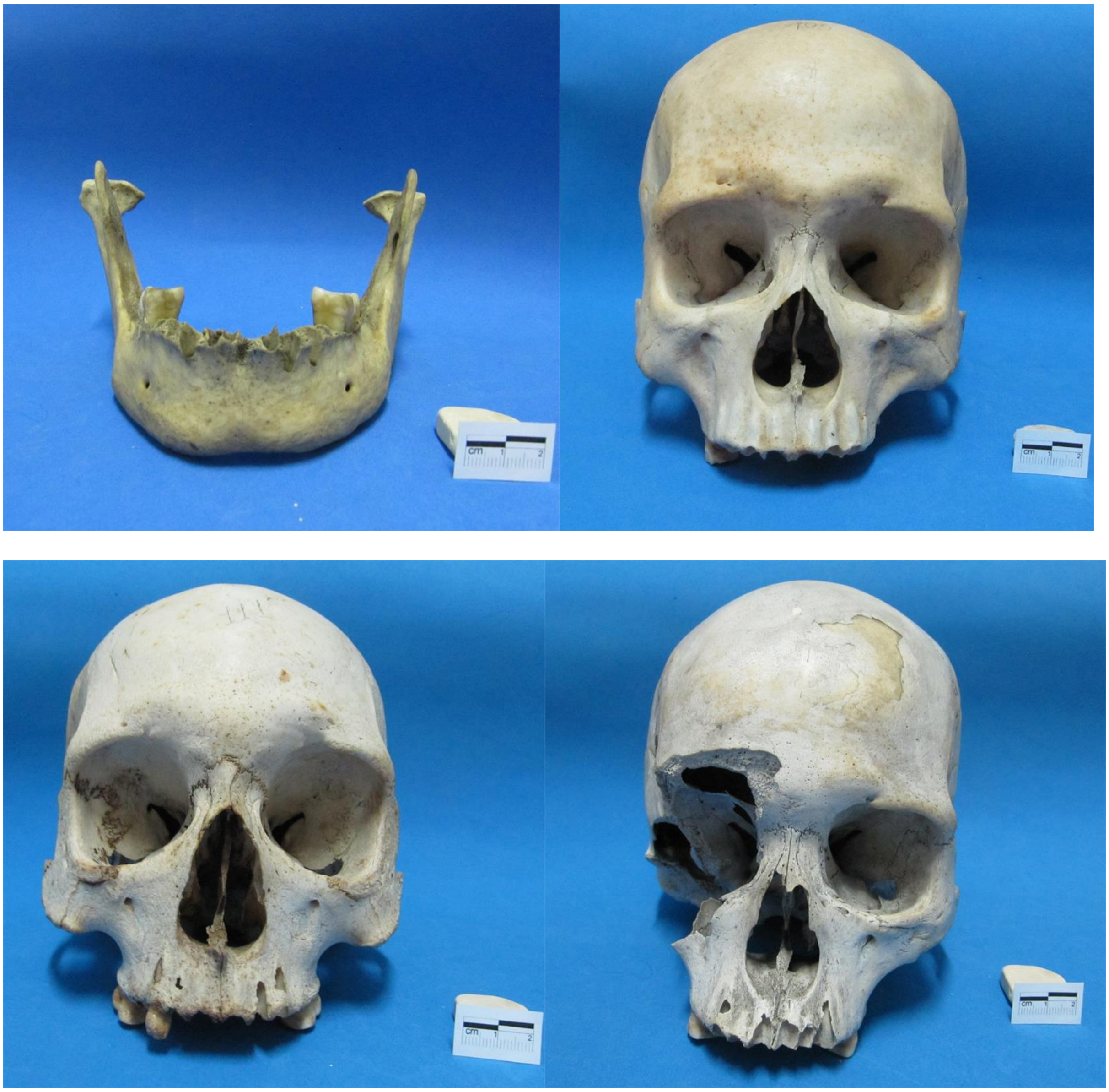
Supplementary Figure 1.2 Bot13 and Bot65 skulls. Front and lateral views of Bot13 (left) and Bot65 (right).



14 C. 9	CRANEO DE INDIO BOTOCUDO (Amutum - 122) Com mandibula
1-1951 15 S-2-A-2	CRANEO DE INDIO BOTOCUDO-RIO DOCE-MINAS (160)
16 C. 4	CRANEO DE INDIO BOTOCUDO-CAVERNA DE ITAPEMIRIN-ESTADO DO ESPIRITO SANTO. Com mandibula (108) da Sp. Antropologica

Numero de ordem	OBJECTO
1-1951 17 S-2-A-2	CRANEO DE INDIO BOTOCUDO, DO RIO DOCE-MINAS GERAES (162)
18 Coll. 9	CRANEO DE INDIO BOTOCUDO, DO ALDEIAMENTO DO MUTUM-ESTADO DE ESPIRITO SANTO- (Antigo, 118). (Minas - 11)
19 C.9	CRANEO DE BOTOCUDO DO ALDEIAMENTO DA POAYA. E. Santo (124)

Supplementary Figure 1.3 Excerpts of the relevant pages of the general catalog for Bot15 and Bot17.



Supplementary Figure 1.4 Skulls from the Marchese Island (upper left and right), Chatham Island (lower left) and New Zealand (lower right), currently at the Museu Nacional.

Supplementary 2

DNA extraction, library preparation, capture experiments and sequencing of the Botocudo samples

Maanasa Raghavan*, Paula F. Campos, Hannes Schroeder, Vanessa F. Gonçalves, Jesper Stenderup, Morten Rasmussen, Morten E. Allentoft

*to whom correspondence should be addressed (mraghavan@snm.ku.dk)

Botocudos

The lab work for the Botocudo samples took place over several years and was conducted by different researchers; hence the protocols followed for each sample are somewhat different. All laboratory work was conducted in sterile, clean laboratories dedicated to ancient DNA research at the Centre for GeoGenetics, University of Copenhagen, Denmark.

DNA extraction

Dentine powder was obtained for all samples by drilling into the teeth with a Dremel drill. For sample Bot15, total cellular DNA was extracted according to the following protocol: 200 mg of tooth powder was incubated overnight at 55ºC in Yang buffer (Yang et al. 1998) and proteinase K. To pellet the undigested powder, the solution was centrifuged at 12,000 g for 5 minutes. The liquid fraction was then transferred into an Amicon ultrafiltration device (30-kDa cutoff, Millipore, Billerica, MA), and spun at 4,000 g for 10 minutes. Following liquid concentration to about 200 µl, DNA was purified using the Qiaquick purification kit (Qiagen, Hilden, Germany).

The sample Bot17 was extracted using a silica suspension-based method (Rohland and Hofreiter 2007). Around 250 mg of powder was drilled from the tooth and then digested for 48 hours at 37°C in lysis buffer containing 0.45 M EDTA and 0.25 mg/ml proteinase K. Following digestion, the supernatant was transferred to 4 volumes of binding buffer containing 5 M GuSCN, 25 mM NaCl and 50 mM Tris. After adding 50 µl of silica suspension, the pH of the solution was adjusted to ~4.0-5.0 by adding *ca*. 300 µl of 30% w/v HCl. The binding solution was then left to incubate at RT for 3 hours. Subsequently, the supernatant was removed and the silica pellets were resuspended in 1 ml of binding buffer. After carefully removing all of the binding buffer, the pellets were rinsed twice with wash buffer containing 50% v/v ethanol, 125 mM NaCl, 10 mM Tris and 1 mM EDTA, pH 8.0. Then the silica pellets were left to dry. Finally, the pellets were resuspended in 60 µl EB buffer and left to incubate for 10 minutes at 37°C, after which the supernatant was removed and stored in low-bind safe-lock tubes.

The samples Bot13 and Bot65 (215 mg and 285 mg of drilled powder, respectively) were incubated at 55°C in a buffer consisting of 0.45 M EDTA, Triton X-100 (1%) and 0.25 mg/ml Proteinase K. After 48 hours, the supernatant was concentrated in 0.5 ml Amicon ultrafiltration device (30-kDa cut-off) to a volume of *ca*. 100 µl. The concentrate was purified using the Qiagen MinElute PCR Purification Kit (Qiagen, Hilden, Germany) with 10x volume PN buffer and AW1/AW2 wash buffers from the

Qiagen Blood and Tissue kit (Qiagen, Hilden, Germany), followed by a final elution in 60 µl EB buffer.

Library preparation

For Bot15, libraries were built using the GS FLX Titanium Rapid Library Preparation Kit (454 Life Sciences, Roche, Branford, CO), according to the manufacturer's protocol with few modifications. Primarily, the extract was not subject to nebulization due to ancient DNA being fragmented in nature. Two libraries (B3a and B3b) were constructed for shotgun sequencing. For each library, 16µl of DNA extract was mixed with 2.5 µl RL, 10x PNK buffer, 2.5 µl RL ATP, 1 µl RL dNTP, 1 µl RL T4 polymerase, 1 µl RL PNK and 1 µl RL Taq polymerase. The mix was incubated at 25°C for 20 minutes, 72°C for 20 minutes and then placed at 4°C. One µl of Index PE Adaptor Oligo Mix (Illumina Multiplexing Sample Preparation Oligonucleotide Kit) and one µl RL ligase were added and the sample was incubated for 10 minutes at 25°C. Each library was purified through a Qiagen Qiaquick column and eluted in 60 µl of Qiagen Buffer EB. The purified libraries were amplified as follows (three parallel reactions for each, with each PCR reaction receiving a different index): 5 µl DNA library, 1X High Fidelity Platinum Taq buffer, 2 mM $MgSO_4$, 200 µM dNTPs each (Invitrogen, Carlsbad, CA), 0.2 µM each of library enrichment PCR1 primers (Kampmann et al. 2011), 0.5 U of High Fidelity Platinum Taq (Invitrogen, Carlsbad, CA) and water to 50 µl. Cycling conditions were: initial denaturing 94°C for 4 minutes, 12 cycles of: 94°C for 30 seconds, 57°C for 20 seconds, 68°C for 20 seconds, and a final extension at 72°C for 7 minutes. PCR products were purified through MinElute spin columns and eluted in 60 µl of Qiagen Buffer EB. A second round of PCR (three parallel reactions for each library) was set up as follows: 5 µl of purified product from first PCR, 25 µl of 2X Phusion master mix (Finnzymes, Espoo, Finland), 0.2 µM each of library enrichment PCR2 primers, and water to 50 µl. Cycling conditions included an initial denaturing at 98°C for 30 seconds, 20 cycles of: 98°C for 10 seconds, 57°C for 20 seconds, 72°C for 20 seconds, and a final extension at 72°C for 7 minutes. The PCR products were purified through a Qiagen Qiaquick column and eluted in 60 µl of water.

For Bot17, 30 µl of the extract was built into one blunt-end library using the NEBNext DNA Sample Prep Master Mix Set 2 (New England Biolabs, Ipswich, MA E6070) and Illumina specific adapters (Meyer and Kircher 2010). The library preparation was carried out as per manufacturer's instructions, bypassing the fragmentation step. The end-repair step was performed in 50 µl reactions using 30 µl of DNA extract, 1X End Repair Reaction Buffer and 0.1 U/µl End Repair Enzyme Mix. The reactions were incubated for 15 minutes at 12°C and 15 minutes at 37°C and purified using Qiagen MinElute spin columns. The elution volume was 30 µl. Following end-repair, Illumina-specific adapters were added to the end-repaired DNA in 50 µl ligation reactions containing 30 µl of DNA, 1X Quick Ligation Reaction Buffer, 0.1 U/µl Quick T4 DNA Ligase and 0.25 µM of adapter mix (Meyer and Kircher 2010). Following an incubation time of 15 minutes at 20°C, the reactions were purified over Qiagen QiaQuick columns and eluted in 30 µl EB Buffer. The adapter fill-in reaction was performed without prior size-selection using 30 µl of adapter-ligated DNA, 1X of Adapter Fill-in Reaction Buffer and 2 µl of Bst polymerase in a final volume of 50 µl. The mix was incubated for 30 minutes at 37°C followed by 20 minutes at 80°C to inactivate the enzyme. Following heat inactivation, 10 µl of library DNA was amplified and indexed using 1x AccuPrime reaction mix, 0.2 µM custom-made index primer (AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTG, where N's represent the index sequence), 0.2 µM Illumina inPE1.0 primer and 1.25 units AccuPrime Pfx polymerase (Invitrogen, Carlsbad, CA). The thermocycling profile was 2 minutes at 95°C, followed by 10 cycles of 15 seconds at 95°C, 30 seconds at 60°C and 30 seconds at 68°C. The amplified library was then purified using a Qiagen QiaQuick column and eluted in 30 µl EB.

For Bot13 and Bot65, one library each was built as outlined above for Bot15 (GS FLX Titanium Rapid Library Preparation Kit, 454 Life Sciences, Roche, Branford, CO), with the following modifications. Ligation was performed for 15 minutes at 25°C. The libraries were purified through Qiagen MinElute columns and eluted in 25 µl of Qiagen Buffer EB after a 10 minute incubation at 37°C. The purified libraries were amplified as follows: 25 µl DNA library, 1X High Fidelity Platinum Taq buffer, 2 mM $MgSO_4$, 200 µM dNTPs each (Invitrogen), 200 nM Illumina Multiplexing PCR primer 1.0, 4 nM Illumina Multiplexing PCR primer 2.0, 200 nM Illumina Index PCR primer, 1 U of High Fidelity Platinum Taq (Invitrogen) and water to 50 µl. Cycling conditions were: initial denaturing at 94°C for 4 minutes, 10 cycles of: 94°C for 30 seconds, 60°C for 30 seconds, 68°C for 20 seconds, and a final extension at 72°C for 7 minutes. PCR products were purified through Qiagen MinElute spin columns and eluted in 20 µl of Qiagen Buffer EB (10 minutes incubation at 37°C)**.** A second round of PCR (four parallel reactions for each library) was set up as follows: 5 µl amplified library from first round PCR, 1X High Fidelity Platinum Taq buffer, 2 mM $MgSO_4$, 200 µM dNTPs each (Invitrogen, Carlsbad, CA), 500 nM Illumina Multiplexing PCR primer 1.0, 10 nM Illumina Multiplexing PCR primer 2.0, 500 nM Illumina Index PCR primer, 1 U of High Fidelity Platinum Taq (Invitrogen, Carlsbad, CA) and water to 50 µl. Cycling conditions were: initial denaturing at 94°C for 4 minutes, 8 cycles of: 94°C for 30 seconds, 60°C for 30 seconds, 68°C for 20 seconds, and a final extension at 72°C for 7 minutes. PCR products originating from the same library were purified through one MinElute spin column each and eluted in 20 µl of Qiagen Buffer EB (10 minutes incubation at 37°C). Due to the lack of any observable peaks corresponding to a library profile on the Bioanalyzer (Agilent 2100 Bioanalyzer High Sensitivity DNA chip), a third round of amplification was carried out. The set up was as follows (four parallel reactions for each library): 5 µl amplified library from second round PCR, 1X High Fidelity Platinum Taq buffer, 2 mM $MgSO_4$, 200 µM dNTPs each (Invitrogen, Carlsbad, CA), 200 nM each of Sol_bridge_P5 and Sol_bridge_P7 (Maricic, Whitten, and Pääbo 2010), 1 U of High Fidelity Platinum Taq (Invitrogen, Carlsbad, CA) and water to 50 µl. Cycling conditions were: initial denaturing at 94°C for 4 minutes, 10 cycles of: 94°C for 30 seconds, 58°C for 30 seconds, 68°C for 20 seconds, and a final extension at 72°C for 7 minutes. PCR products originating from the same library were purified through one MinElute spin column each and eluted in 20 µl of Qiagen Buffer EB (10 minutes incubation at 37°C).

Mitochondrial and SNP array captures

Capture experiments were performed for Bot15 only. Remainder volumes from the two pre-amplified libraries B3a and B3b were pooled together and purified through MinElute columns according to accompanying protocol and eluted in 50 µl of Qiagen Buffer EB. The purified libraries were subsequently amplified (seven parallel reactions for each) as follows: 5 µl DNA library, 1X High Fidelity Platinum Taq buffer, 2 mM $MgSO_4$, 200 µM dNTP each (Invitrogen, Carlsbad, CA), 0.2 µM each of library enrichment PCR1 primers (Kampmann et al., 2011), 0.5 U of High Fidelity Platinum Taq (Invitrogen, Carlsbad, CA) and water to 50 µl. Cycling conditions were: initial denaturing 94°C for 10 minutes, 12 cycles of: 94°C for 30 seconds, 57°C for 20 seconds, 68°C for 20 seconds, and a final extension at 72°C for 10 minutes. PCR products were purified through MinElute spin columns and eluted in 60 µl of water. A second round of PCR (five parallel reactions for each library) was set up as follows: 5 µl of purified product from first PCR, 25 µl of 2X Phusion master mix (Finnzymes, Espoo, Finland), 0.2 µM each of library enrichment PCR2 primers and water to 50 µl. Cycling conditions included an initial denaturing at 98°C for 30 seconds, 20 cycles of: 98°C for 10 seconds, 58°C for 20 seconds, 72°C for 20 seconds, and a final extension at 72°C for 7 minutes. PCR products were purified on MinElute columns and eluted in 60 µl water and quantified using Qubit 2.0 Fluorometer (Invitrogen, Carlsbad, CA).

Mitochondrial genome capture was carried out on the pool according to (Maricic, Whitten, and Pääbo 2010). A final round of amplification was performed on the enriched library prior to sequencing, as determined by quantitative PCR (qPCR), with four parallel PCR reactions set up as follows: 2 µl of enriched pool, 1x Phusion HF buffer (New England Biolabs), 200 µM dNTP each (Invitrogen, Carlsbad, CA), 0.5 µM Sol_bridge_P5, 0.5 µM Sol_bridge_P7, 2% DMSO, 0.02 U/ µl Phusion High-Fidelity DNA Polymerase (New England Biolabs) and water up to 50 µl. Cycling conditions were: initial denaturing at 98°C for 30 seconds, 18 cycles of: 98°C for 10 seconds, 57°C for 20 seconds, 72°C for 20 seconds, and a final extension at 72°C for 7 minutes. The four PCR reactions were pooled and purified through a MinElute column and eluted in 20 µl of Qiagen Buffer EB.

Array capture was performed on the above amplified library after two further amplification rounds (20 cycles each), in order to acquire the recommended 20 µg per library for starting the capture. Agilent 1 million-feature custom arrays were designed to target positions in the nuclear genome (See S8 for description of how the positions were selected) The protocol outlined in (Hodges et al. 2009) was followed from steps 29 through 61. Two successive rounds of hybridization were performed under identical conditions.

Sequencing

All libraries (captured or shotgun) were visualized using an Agilent 2100 Bioanalyzer High Sensitivity DNA chip. The libraries for samples Bot15, Bot17 were subsequently sequenced on Illumina HiSeq 2000, while those for Bot13 and Bot65 were sequenced on Illumina MiSeq. All libraries were sequenced at the National High-throughput DNA Sequencing Centre (http://seqcenter.ku.dk/).

References

Hodges, Emily, Michelle Rooks, Zhenyu Xuan, Arindam Bhattacharjee, D. Benjamin Gordon, Leonardo Brizuela, W. Richard McCombie, and Gregory J. Hannon. 2009. “Hybrid Selection of Discrete Genomic Intervals on Custom-Designed Microarrays for Massively Parallel Sequencing.” *Nature Protocols* 4 (6) (May): 960–974. doi:10.1038/nprot.2009.68.

Kampmann, Marie-Louise, Sarah L Fordyce, María C Avila-Arcos, Morten Rasmussen, Eske Willerslev, Lars P Nielsen, and M Thomas P Gilbert. 2011. “A Simple Method for the Parallel Deep Sequencing of Full Influenza A Genomes.” *Journal of Virological Methods* 178 (1-2) (December): 243–248. doi:10.1016/j.jviromet.2011.09.001.

Maricic, Tomislav, Mark Whitten, and Svante Pääbo. 2010. “Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products.” *PloS One* 5 (11): e14004. doi:10.1371/journal.pone.0014004.

Meyer, Matthias, and Martin Kircher. 2010. “Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing.” *Cold Spring Harbor Protocols* 2010 (6) (June): pdb.prot5448. doi:10.1101/pdb.prot5448.

Rohland, Nadin, and Michael Hofreiter. 2007. “Ancient DNA Extraction from Bones and Teeth.” *Nature Protocols* 2 (7) (July): 1756–1762. doi:10.1038/nprot.2007.247.

Yang, Dongya Y., Barry Eng, John S. Waye, J. Christopher Dudar, and Shelley R. Saunders. 1998. “Improved DNA Extraction from Ancient Bones Using Silica-Based Spin Columns.” *American Journal of Physical Anthropology* 105 (4): 539–543. doi:10.1002/(SICI)1096-8644(199804)105:4<539::AID-AJPA10>3.0.CO;2-1.

Supplementary 3

Processing and mapping of raw sequence data from the Botocudo individuals and present day human genomes

Simon Rasmussen*, Mingkun Li, Morten Rasmussen, Anna-Sapfo Malaspinas

*to whom correspondence should be addressed (simon@cbs.dtu.dk)

As detailed in the main text and in S2, for Bot15 three main experiments were performed: shotgun sequencing, SNP capture and mtDNA capture. All other Botocudos were sequenced with a shotgun approach (*i.e.*, no capture experiments for those individuals).

Basecalling

Sequencing runs generated from libraries for the shotgun experiments (Bot13-e1-lib1, Bot15-B3a, Bot15-B3b, Bot17-e2-lib1, Bot65-e1-lib1), the SNP capture experiment (Bot15-B3-SNPCapture) and the mtDNA capture experiment (Bot15-B3-mtDNACapture) were all basecalled using the Illumina software CASAVA v1.8.2 requiring an exact match between the 6bp index sequence used in the experiments and the one observed in the runs.

Shotgun sequence processing (Bot13, Bot15, Bot17, Bot65)

For the libraries Bot15-B3a, Bot15-B3b, Bot17-e2-lib1 that were all sequenced on an Illumina HiSeq 2000, a large fraction of the reads had undetermined indexes due to low complexity in the index and these went through an additional filter allowing 1 mismatch in the barcode. This procedure retrieved 149.9 and 15.0 million additional reads from the Bot15, and Bot17 samples.

For all shotgun libraries, the raw reads were trimmed using AdapterRemoval-1.1 (Lindgreen 2012) for adapter sequence and leading/trailing Ns to a minimum length of 25 nt (--minlength 25 --trimns). Statistics for each sample are given in Supplementary Table 3.1. Hereafter the reads were mapped to the human reference genome build37.1 using bwa-0.6.2 (H. Li and Durbin 2009) with seed disabled to allow for better sensitivity (Schubert et al. 2012). Alignments were filtered for reads with a mapping quality of at least 30, sorted and merged to libraries using Picard (http://picard.sourceforge.net). Library BAMs had duplicates removed using Picard MarkDuplicates, were merged to sample level and realigned using GATK (DePristo et al. 2011). Last, the md-tags were recalculated using samtools (Li et al. 2009). The level of endogenous DNA was determined as percentage of mapped reads after filtering for mapping quality of 30 compared to the raw amount of reads produced. The Bot13 and Bot65 individuals had low endogenous DNA content and were therefore not used for further analysis (see Supplementary Table 3.2). Coverage and average read depth were estimated using BEDtools (Quinlan and Hall 2010) and pysam (http://code.google.com/p/pysam/). Statistics related to coverage (*i.e.*, the number of positions in the genome covered by at least one read) and depth (*i.e.*, the average number of reads covering each position) for the Bot15 and Bot17 individuals are given in Supplementary Table 3.3, and Supplementary Figure 3.1. In total the Bot15 and Bot17 individuals had an average mapped depth of 1.5X and 1.2X, respectively.

Shotgun sequence processing of present day human genomes

We processed Illumina data from 5 high coverage modern humans produced by Meyer et al. (2012). The data (SRX103808) was first deplexed allowing 1 mismatch in the indexes and hereafter both low (ERR033731- ERR033734 and ERR019686-ERR019689) and high coverage data was mapped to the human reference genome build 37.1 by bwa-0.6.2 (H. Li and Durbin 2009) using –q 15 for trimming low quality ends of the reads. Then the resulting BAM-files were processed similarly to the Botocudo samples and statistics are shown in Supplementary Table 3.4.

mtDNA sequence processing (Bot15 and Bot17)

The mitochondrial DNA sequences (consensus) of the two Botocudo individuals were determined using the shotgun data for Bot17, while for Bot15 we used primarily the mtDNA capture data and we confirm the result with the shotgun data. As described above, the raw reads were trimmed using AdapterRemoval-1.1 (Lindgreen 2012) for adapter sequence and leading/trailing Ns to a minimum length of 25 nt (--minlength 25 --trimns). We used two different mapping strategies; one to call the consensus sequence and one to determine the contamination fraction.

Consensus sequence

To determine the consensus sequence, we proceeded by a two-step “manual iteration” to avoid biasing the result by the reference sequenced used. It has been shown by Li et al. (2012) that reads from nuclear inserts are at a low enough frequency in capture experiments such that they do not influence the determination of the authentic mtDNA sequence. We assume here that this is also true for shotgun experiments. We therefore mapped the reads to mtDNA sequences to determine the consensus (as opposed to the whole genome). For each step, we visualized the results using tablet (Milne et al. 2013).

For the first iteration, we mapped the reads to the revised Cambridge reference sequence (rCRS, Andrews et al. 1999) as described above. The depth after mapping to the rCRS is 179.9X for the capture data (and 84.3X for shotgun data) for the Bot15 individual and 62.5X (shotgun only) for the Bot17 individual.

To call the consensus we discarded bases with a base quality lower than 20. All indels were then checked manually. Several approaches were applied to evaluate the quality of the intermediate and final consensus. First, we checked that all positions had a major allele frequency greater than 70% and additionally found that only 7 positions had major allele frequency lower than 90% for the Bot15 capture enriched library and 8 positions for Bot15 shotgun library. Meanwhile for the Bot17 shotgun library, there were 7 positions with a major allele frequency ranging from 70% to 90%, most of which were located in C-stretch regions and the two ends of the mtDNA genome. Second, all positions were covered by reads that did not contain any mismatches relative to the consensus, and the number of such reads was greater or equivalent to that of the read with 1 mismatch (relative to the consensus) at all positions, which is expected under the scenario that the right consensus is called.

In the next step, we use a worldwide set of complete mtDNA genomes enriched for Native American and Oceanic sequences (S6) to determine the closest sequence. For Bot15 the closest sequence was from an individual from the Cook Islands (GenBank: AY289068) and for Bot17 the closest sequence was from an individual from the Philippines (GenBank: GQ119029).

We then mapped the original reads against these new sequences for each case and called a new consensus for each sample as described above. For Bot15, the consensus was identical to the Cook Islands (GenBank: AY289068) and we stopped the iteration. For Bot17, the consensus was different from sequence from the Philippines (GenBank: GQ119029) and we realigned the reads to the consensus. A new consensus was called, only to find that the procedure had converged already.

To assess the effect of the stringent threshold of 30 on the mapping quality, in each case, we also mapped the original reads against the final consensus with gem (-e 0.05, Marco-Sola et al. 2012), bwa aln (-n 10 –e3 –o 1, Li and Durbin 2009) and bwa bwasw (-z 2, Li and Durbin 2010) to verify that we did not produce any biases.

In summary, after mapping to the final consensus: for Bot15, the mtDNA capture library had an average depth of 180X and ranged between 22X and 188X and all positions had a depth of at least 60X except the first 10 bp and the last 10 bp on the genome (the circularity of the genome is not taken into account in the mapping step). The average depth for the shotgun library for Bot15 was 97X, ranging between 4X and 140X, with all positions having a depth of at least 20X except the first 10 bp and the last 10 bp on the genome. Both the shotgun and the capture data gave rise to the same consensus, which is identical to the Cook mtDNA. The enrichment, defined as the ratio of the depths before and after capture normalized by the number of trimmed reads, was around 25.2. A summary of the mtDNA assembly for both mtDNA capture and shotgun experiments can be found in Supplementary Table 3.5. From the Bot17 shotgun library, we got an average depth of 89X, ranging between 1X to 117X, where all positions had a depth greater than 30X except the first and last 10 bp, see Supplementary Figure 3.2.

In addition to the mapping consensus-building strategy a *de novo* assembly approach was taken to attempt to reduce the bias introduced by mapping to existing mtDNA sequences and also to somehow overcome the limitation of the assumed linear molecule at the mapping step above. The reads used for the *de novo* assemblies were the Bot15 mtDNA capture data and the Bot17 shotgun reads that mapped by bwa-mem (bwa-0.7.2) to the Bot17-mtDNA consensus sequence. The assemblies were performed using Velvet-1.2.08 (Zerbino and Birney 2008) and SOAPdenovo-1.05 (R. Li et al. 2010) using different *k*-mers and selecting the best assemblies resulting in the mtDNA assembled in one contig. Because the mtDNA is circular and the assemblers are also unaware of this, we removed artificial overlapping sequence from the ends of the contig by alignment to the respective consensus sequences. The *de novo* assembly produced the same consensus as above for all cases.

Mapping the data for contamination estimates

For contamination estimates, once the consensus sequence was determined, we mapped the reads against the entire build37.1 augmented with the respective consensus, to remove the possibility of nuclear inserts that would look like contaminants (see S6 for more details).

SNP-capture experiment sequence processing (Bot15)

The SNP capture experiment targeting 5744 SNPs (see S7 for details on SNPs selection) resulted in 151 million reads that were processed as described above for the shotgun data. In total 99.7% of the reads passed trimming with an average length of 66bp and 33.8 million reads (22.4%) mapped to the human genome, see Supplementary Table 3.6.

References

Andrews, Richard M., Iwona Kubacka, Patrick F. Chinnery, Robert N. Lightowlers, Douglass M. Turnbull, and Neil Howell. 1999. “Reanalysis and Revision of the Cambridge Reference Sequence for Human Mitochondrial DNA.” *Nature Genetics* 23 (2): 147–147. doi:10.1038/13779.

DePristo, Mark A, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data.” *Nature Genetics* 43 (5): 491–98. doi:10.1038/ng.806.

Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

———. 2010. “Fast and Accurate Long-Read Alignment with Burrows–Wheeler Transform.” *Bioinformatics* 26 (5): 589–95. doi:10.1093/bioinformatics/btp698.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Li, Mingkun, Roland Schroeder, Albert Ko, and Mark Stoneking. 2012. “Fidelity of Capture-Enrichment for mtDNA Genome Sequencing: Influence of NUMTs.” *Nucleic Acids Research* 40 (18): e137. doi:10.1093/nar/gks499.

Li, Ruiqiang, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, et al. 2010. “De Novo Assembly of Human Genomes with Massively

Parallel Short Read Sequencing.” *Genome Research* 20 (2): 265–72. doi:10.1101/gr.097261.109.

Lindgreen, Stinus. 2012. “AdapterRemoval: Easy Cleaning of next-Generation Sequencing Reads.” *BMC Research Notes* 5: 337. doi:10.1186/1756-0500-5-337.

Marco-Sola, Santiago, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. 2012. “The GEM Mapper: Fast, Accurate and Versatile Alignment by Filtration.” *Nature Methods* 9 (12): 1185–88. doi:10.1038/nmeth.2221.

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, et al. 2012. “A High-Coverage Genome Sequence from an Archaic Denisovan Individual.” *Science* 338 (6104): 222–26. doi:10.1126/science.1224344.

Milne, Iain, Gordon Stephen, Micha Bayer, Peter J. A. Cock, Leighton Pritchard, Linda Cardle, Paul D. Shaw, and David Marshall. 2013. “Using Tablet for Visual Exploration of Second-Generation Sequencing Data.” *Briefings in Bioinformatics* 14 (2): 193–202. doi:10.1093/bib/bbs012.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42. doi:10.1093/bioinformatics/btq033.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, et al. 2010. “Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia.” *Nature* 468 (7327): 1053–60. doi:10.1038/nature09710.

Reich, David, Nick Patterson, Martin Kircher, Frederick Delfin, Madhusudan R. Nandineni, Irina Pugach, Albert Min-Shan Ko, et al. 2011. “Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania.” *The American Journal of Human Genetics* 89 (4): 516–28. doi:10.1016/j.ajhg.2011.09.005.

Schubert, Mikkel, Aurelien Ginolhac, Stinus Lindgreen, John F Thompson, Khaled AS AL-Rasheid, Eske Willerslev, Anders Krogh, and Ludovic Orlando. 2012. “Improving Ancient DNA Read Mapping against Modern Reference Genomes.” *BMC Genomics* 13: 178. doi:10.1186/1471-2164-13-178.

Zerbino, Daniel R., and Ewan Birney. 2008. “Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs.” *Genome Research* 18 (5): 821–29. doi:10.1101/gr.074492.107.

Supplementary Table 3.1 Summary of the number of reads of the shotgun libraries before and after trimming and after mapping against build 37.1.

Sample	Library	Raw reads	Trimmed reads	Average read length	Gbases (after trim)
Bot13	e1_lib1	2,642,371	2,623,335	50.4	5.4
Bot15	B3a	362,032,870	361,466,599	57.6	20.8
Bot15	B3b	972,657,984	971,628,393	58.1	56.4
Bot15	All	1,334,690,854	1,333,094,992	58.0	77.3
Bot17	e2_lib1	662,984,192	647,045,202	49.8	32.2
Bot65	e1_lib1	4,197,865	3,967,339	50.2	0.2

Supplementary Table 3.2 Number of reads, endogenous content and proportion of duplicates after trimming and mapping the shotgun libraries against build 37.1.

Sample	Library	Trimmed reads	Mapped q30	Final Bam	% endogenous	% duplicates
Bot13	e1_lib1	2,623,335	1,236	1,219	0.05	0.0
Bot15	B3a	361,466,599	20,025,002	19,358,543	5.5	3.3
Bot15	B3b	971,628,393	53,976,237	49,948,813	5.6	7.5
Bot15	All	1,333,094,992	74,001,239	69,307,356	5.6	6.3
Bot17	e2_lib1	647,045,202	80,561,886	70,941,275	12.5	11.9
Bot65	e1_lib1	3,967,339	15,497	14,482	0.37	6.55

Supplementary Table 3.3 Summary of the average depth and percent of the genome covered by at least one read after trimming and mapping the shotgun libraries against build 37.1.

Sample	Library	Average Depth (X)	Mapped Gbases	Covered >= 1X (%)
Bot13	e1_lib1	0.0	0.0	0.0
Bot15	B3a	0.4	1.3	30.4
Bot15	B3b	1.1	3.4	53.7
Bot15	All	1.5	4.7	62.3
Bot17	e2_lib1	1.2	4.3	62.6
Bot65	e1_lib2	0.0	0.0	0.0

Supplementary Table 3.4 Summary of the present day human samples used in the analysis.

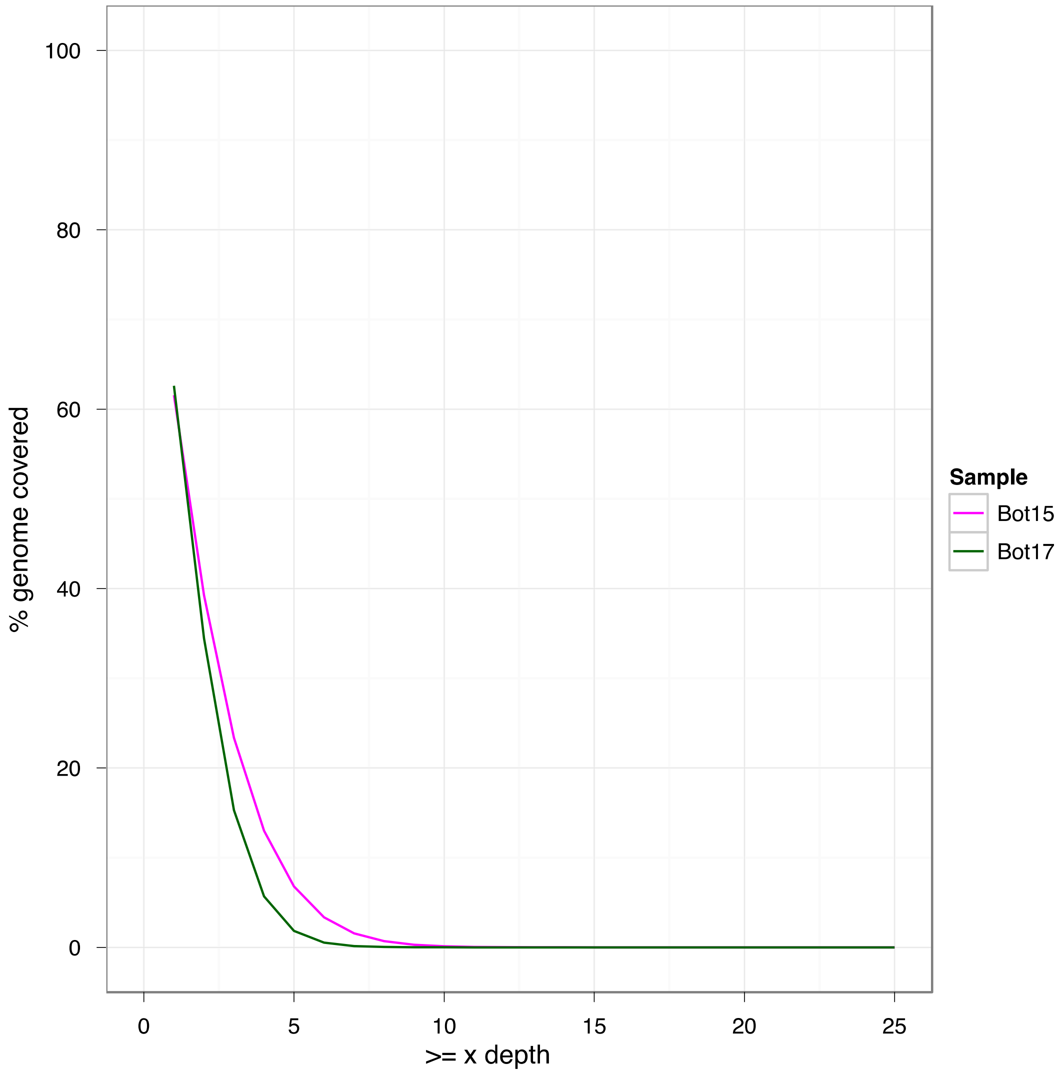
Sample	Population	Reference	Average Depth (X)	Covered > 1X (%)
HGDP00521	French	Meyer et al. 2012	22.6	90
HGDP00542	Papuan	Meyer et al. 2012	21.6	90
HGDP00778	Han	Meyer et al. 2012	22.3	90
HGDP00927	Yoruba	Meyer et al. 2012	26.7	90
HGDP00998	Karitiana	Meyer et al. 2012	21.3	90

Supplementary Table 3.5 Average depth and percent of the genome covered by at least one read after trimming and mapping to the mtDNA consensus for each experiment.

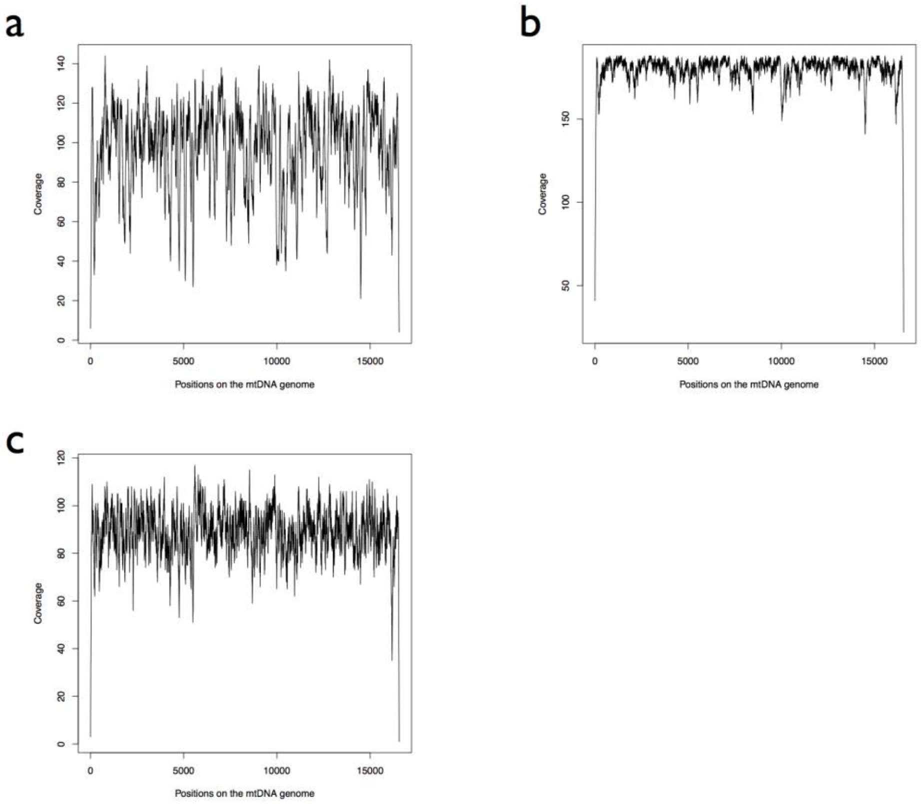
Sample	Experiment	Trimmed reads	Mapped q30	Final Bam	%duplicates
Bot15	mtDNA Capture	97,901,620	9,566,138	32,476	99.66
Bot15	Shotgun	1,333,094,992	65,611	21,391	67.40
Bot17	Shotgun	647,045,202	91,188	25,634	71.89
Sample	**Experiment**	**Average Depth**	**Mapped Gbases**	**Enrichment**	**Covered >= 1X**
Bot15	mtDNA Capture	179.81	2,977,524	25.2	100%
Bot15	Shotgun	97.17	1,609,080	NA	100%
Bot17	Shotgun	91.11	1,508,498	NA	100%

Supplementary Table 3.6 Summary statistics for the assembly after mapping against build 37.1 for the Bot15 SNP capture experiment. See S8 for more details.

Sample	Experiment	Trimmed reads	Mapped q30	Final Bam	%duplicates
Bot15	SNP Capture	151,024,433	33,877,154	1,831,895	94.59



Supplementary Figure 3.1 Coverage distributions for Bot15 and Bot17 individuals mapped to build37.1 showing percentage of genome covered at a depth of X reads or more.



Supplementary Figure 3.2 Read depth across the mtDNA for the Bot15 shotgun (a), Bot15 mtDNA capture (b) and the Bot17 shotgun (c) experiments. Libraries when mapped against the consensus sequences.

Supplementary 4
Error, damage and contamination rates for Bot15 and Bot17 (shotgun data)

Anders Albrechtsen*, Ida Moltke, Yong Wang , Anna-Sapfo Malaspinas*

*to whom correspondence should be addressed (albrecht@binf.ku.dk, annasapfo@gmail.com)

In this section we describe how we evaluated the quality and the authenticity of the sequence data for the ancient samples by measuring patterns specific to ancient DNA, error rates and the amount of contamination for each sample for the nuclear data.

Ancient DNA damage for Bot15 and Bot17

It has been shown that DNA in museum specimens is fragmented and chemically modified (Briggs et al. 2007; Sawyer et al. 2012). Both patterns can therefore be used to assess the authenticity of ancient DNA data. We measured the fragment length distribution and the substitutions at each position of the sequenced reads compared to the reference genome for all genetic experiments (SNP array capture, mtDNA capture and shotgun experiments). Since we obtained similar results for all experiments, we only discuss the shotgun experiment results in what follows.

We first looked at the read length distribution. Indeed, the average read length has been found to be correlated with age (Allentoft et al. 2012), but this trend seems to depend on several environmental parameters, and is therefore hard to demonstrate in practice when comparing remains from different sites (Sawyer et al. 2012). We call "read length" the length of the reads after trimming and mapping to the human genome. The shotgun data was produced on an Illumina HiSeq 2000 that was run for 92-94 cycles not including the six cycles for the index. The average read length we report here is therefore biased downwards since reads longer than 92-94 can only be sequenced for the first 92-94 base pairs. We observed that, as expected, the average read length is rather short; 67.1bp (Bot15) and 52.7bp (Bot17) (see Supplementary Figure 4.1). Moreover, as can be seen in the figure the distribution has a single mode which is compatible, assuming the potential contaminant has a different read length distribution, with a case without a large fraction of contamination.

A common type of chemical feature observed in ancient DNA is an increased frequency of apparent cytosine (C) to thymine (T) substitutions close to the ends of the DNA fragments (see, *e.g.*, (Briggs et al. 2007)). This has been explained by a potential increase of deamination of C residues at single stranded overhangs. For ancient DNA fragments, we therefore expect an increased C->T at the 5' end and an increased G->A at the 3' end for double stranded libraries. We calculated the frequencies of observing a given nucleotide (*e.g.*, T) in Bot15 and Bot17 conditioning on the reference allele (*e.g.*, C) and the position along the read (from both 5' and 3'). Comparing to other types of mismatches, we observed an increase rate of C->T mismatches near the 5' end, and an increase rate of G->A mismatches near the 3' end (Supplementary Figure 4.2). The rate of C->T and G->A is similar to what has been observed for samples around 400-600 years old (Sawyer et al. 2012).

Error estimation for Bot15 and Bot17

Data and data filtering

To get a rough estimate of the error rates in the sequenced individuals, we compared Bot15 and Bot17 (shotgun data) to a high quality genome data from the 1000 genomes project individual with ID NA12778[1] and used the chimpanzee genome as an outgroup (pantro2 from the hg19 multiz46) to determine the ancestral allele. The high quality genome was filtered with minimum base quality 35 and minimum mapping quality 35. For Bot15 and Bot17 we removed all reads with a mapping quality below 30 (S3) and all bases with a quality score below 20.

Estimating overall error rates

We estimated the overall error rates with a method similar to the method used by (Reich et al. 2010). The estimation is based on the idea that any given human sample should have the same expected number of derived alleles compared to the chimp sequence. We estimate the expected number of derived alleles from the high quality genome and assume that if we observe a higher number of derived alleles in the sample of interest this excess is due to errors. If the high quality genome has no errors, this leads to an error rate estimate, which is equal to the true error rate. If the high quality genome does have errors, the estimated error rate can roughly be understood as the excess error rate relative to the error rate of the high quality genome. Details of the method can be found in (Orlando et al. (2013) supplementary 4).

Results for the error rates

The estimated error rates are shown in Supplementary Figure 4.3 for Bot15, Bot17 and several modern genomes (described in S3)**.** The average error rate is ~10 times higher for the ancient samples compared to the modern genomes. This is mostly due to the expected increase in error rate for C->T and G->A for **Bot15** and **Bot17** as discussed above. Indeed, when excluding C->T and G->A errors, the average error rate of Bot15 and Bot17 is comparable to the other genomes. The transition and overall error rates are 1.2% and 0.59% for Bot15 and 0.9% and 0.46% for Bot17, respectively. Those values are similar to previously reported error rates for other ancient genomes. For example, the Siberian Mal'ta (Raghavan et al. 2013) had a transition error rate of about 0.4% (0.27% overall) while the ancient Anzick Native American had a transition error rate of about 1.7% (0.84% overall, Rasmussen et al. (2014)).

[1] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data/NA12778/alignment/

Contamination Analysis for Bot15 and Bot17 based on the X chromosome

To measure contamination we used two estimation methods which are described in detail in previous work (Rasmussen et al. (2011) supplementary 12). In what follows, we will briefly recapitulate the idea behind these methods and then describe the results.

Idea behind: X chromosome analysis

Since both Bot15 and Bot17 are males, they carry only one X chromosome and thus only a single allele at each site on that chromosome (if one disregards the small part where the X chromosome is homologous with the Y chromosome). Reads that cover the same position but do not contain the same base must therefore either be due to errors (sequencing or mapping) or contamination, i.e., reads that derive from other individuals. The frequency with which such mismatches are observed can therefore be used for estimating the amount of contamination in the two samples. To disentangle to what extent the observed mismatches observed in each sample are caused by error and to what extent they are caused by contamination, we exploit the fact that contamination will have no detectable effect at sites at which the sample and the contamination source(s) share the same allele, and hence it will never have an effect at sites that are monomorphic in humans. We exploit this by - in broad terms - estimating the contamination fraction as the excess mismatch rate at polymorphic sites compared to the rate at monomorphic sites.

Since we cannot list all polymorphic sites in humans, we restricted ourselves to polymorphic sites in the 60 unrelated European HapMap CEPH individuals Frazer et al. (2007). Similarly, to represent the monomorphic sites we restricted ourselves to the sites adjacent to the polymorphic sites, because these sites are less likely to be polymorphic and their error structure will be more similar to the error structure at the selected polymorphic sites than the error structure at random sites. Hence we first compared the mismatch rates for the ancient samples at the sites that are polymorphic in the 60 CEPH individuals to the mismatch rates at their adjacent sites. Then we computed the contamination fraction based on these mismatch rates and allele frequencies from Europeans (*i.e.*, we assume the contaminant is European).

Data

Using HapMap phase II (release 27) data, we identified all sites that are polymorphic in the 60 unrelated CEPH individuals. The sites were lifted to HG19 using liftover[2]. Subsequently, we pruned this set of polymorphic sites such that no sites were less than 10 bases apart. We also estimated the allele frequency in the European population based on these 60 individuals.

The contamination estimates were based on the read data from Bot15 and Bot17 that map to the X chromosome. However, this data was first filtered as follows:

[2] http://genome.ucsc.edu/cgi-bin/hgLiftOver

- the ends of the X chromosome were trimmed to remove the regions that are homologous with the Y chromosome,
- the sites were filtered based on mapability (100mer), so that no region will map to another region of the genome with an identity above 98%,
- reads with a mapping quality score of less than 30 and bases with a base quality score less than 20 were removed,
- all sites with a read depth below 2 or above 40 were removed.

Model

We used two different ML based estimation methods. The first method "test 1" uses all reads available and is more powerful, but assumes independent error rates both within and between sites. The second method "test 2" uses only a single randomly sampled read per site and is less powerful, but makes no assumptions about the independence of errors. For both the resulting estimates the standard error was estimated using a jackknife procedure.

As explained above the contamination estimates are based on mismatch rates. To be able to calculate these rates, we need to know which allele is the true allele. Assuming low error and low contamination rates, the true allele for each sample can be inferred by choosing the allele most frequently seen in the reads covering the site. For example, if the depth is five and the error rate plus the contamination rate is below 2% then the probability of inferring the wrong allele by choosing the most frequent allele is below 0.0008. Hence when calculating the mismatch rates we assumed that the allele with the highest frequency is the true one and thus that the lower frequency reads, "minor reads", are mismatches. Note that as was shown in (Rasmussen et al. 2011), the contamination estimation methods will still easily be able to detect contamination even if the contamination rate plus the error rate is higher than 2% and thus the methods are fairly robust to violation of this assumption.

Results

The mismatch counts around the polymorphic sites are shown in Supplementary Tables 4.1 and 4.2 while the estimates of the contamination rates achieved based on these counts are shown on Supplementary Table 4.3. For both ancient samples the contamination fraction is below 3% and nuclear estimates are in good agreement with the estimates based on the mtDNA (see S6 and Supplementary Table 4.3). The contamination rates are similar to the ancient Siberian Mal'ta individual (Raghavan et al. 2013) and the Anzick Native American (Rasmussen et al. 2014), which had an estimated contamination rate of 2% and 1.2%, respectively.

References

Allentoft, Morten E., Matthew Collins, David Harker, James Haile, Charlotte L. Oskam, Marie L. Hale, Paula F. Campos, et al. 2012. "The Half-Life of DNA in Bone: Measuring Decay Kinetics in 158 Dated Fossils." *Proceedings of the Royal Society B: Biological Sciences*, October. doi:10.1098/rspb.2012.1745.

Briggs, Adrian W., Udo Stenzel, Philip L. F. Johnson, Richard E. Green, Janet Kelso, Kay Prufer, Matthias Meyer, et al. 2007. "Patterns of Damage in Genomic DNA Sequences from a Neandertal." *Proceedings of the National Academy of Sciences* 104 (37): 14616–21. doi:10.1073/pnas.0704665104.

Frazer, Kelly A., Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, John W. Belmont, et al. 2007. "A Second Generation Human Haplotype Map of over 3.1 Million SNPs." *Nature* 449 (7164): 851–61. doi:10.1038/nature06258.

Orlando, Ludovic, Aurélien Ginolhac, Guojie Zhang, Duane Froese, Anders Albrechtsen, Mathias Stiller, Mikkel Schubert, et al. 2013. "Recalibrating Equus Evolution Using the Genome Sequence of an Early Middle Pleistocene Horse." *Nature* 499 (7456): 74–78. doi:10.1038/nature12323.

Raghavan, Maanasa, Pontus Skoglund, Kelly E. Graf, Mait Metspalu, Anders Albrechtsen, Ida Moltke, Simon Rasmussen, et al. 2013. "Upper Palaeolithic Siberian Genome Reveals Dual Ancestry of Native Americans." *Nature* advance online publication (November). doi:10.1038/nature12736.

Rasmussen et al. "The Genome of a Late Pleistocene Human from a Clovis Burial Site in Western Montana."

Rasmussen, Morten, Sarah L. Anzick, Michael R. Waters, Pontus Skoglund, Michael DeGiorgio, Thomas W. Stafford Jr, Simon Rasmussen, et al. 2014. "The Genome of a Late Pleistocene Human from a Clovis Burial Site in Western Montana." *Nature* 506 (7487): 225–29. doi:10.1038/nature13025.

Rasmussen, Morten, Xiaosen Guo, Yong Wang, Kirk E. Lohmueller, Simon Rasmussen, Anders Albrechtsen, Line Skotte, et al. 2011. "An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia." *Science* 334 (6052): 94–98. doi:10.1126/science.1211177.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327): 1053–60. doi:10.1038/nature09710.

Sawyer, Susanna, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and Svante Paabo. 2012. "Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA." *PLoS ONE* 7 (3): e34131. doi:10.1371/journal.pone.0034131.

position	-4	-3	-2	-1	0	1	2	3	4
minor reads	49	66	78	91	199	81	72	74	77
all reads	13455	13470	13456	13474	13408	13519	13544	13492	13472

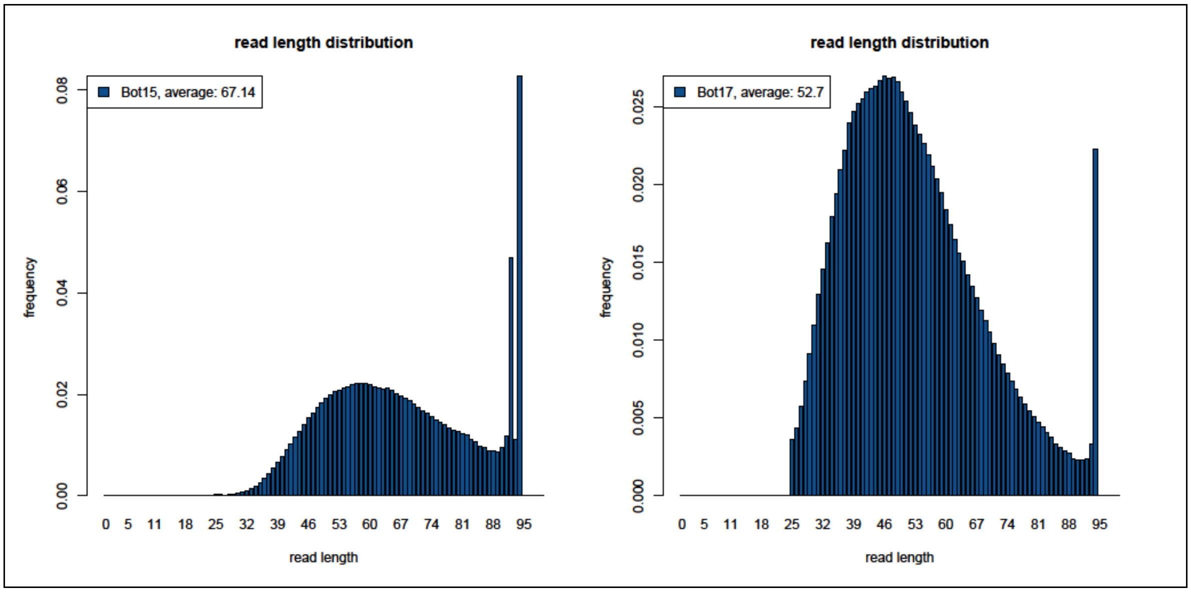
Supplementary Table 4.1 Counts of minor reads for Bot15 around known polymorphic sites on the X chromosome. A minor read is a read that is less or equally common to a major read. The top row is the relative position respective to the polymorphic site.

position	-4	-3	-2	-1	0	1	2	3	4
Minor reads	65	76	81	79	124	80	69	69	77
All reads	20361	20395	20411	20417	20405	20439	20444	20395	20376

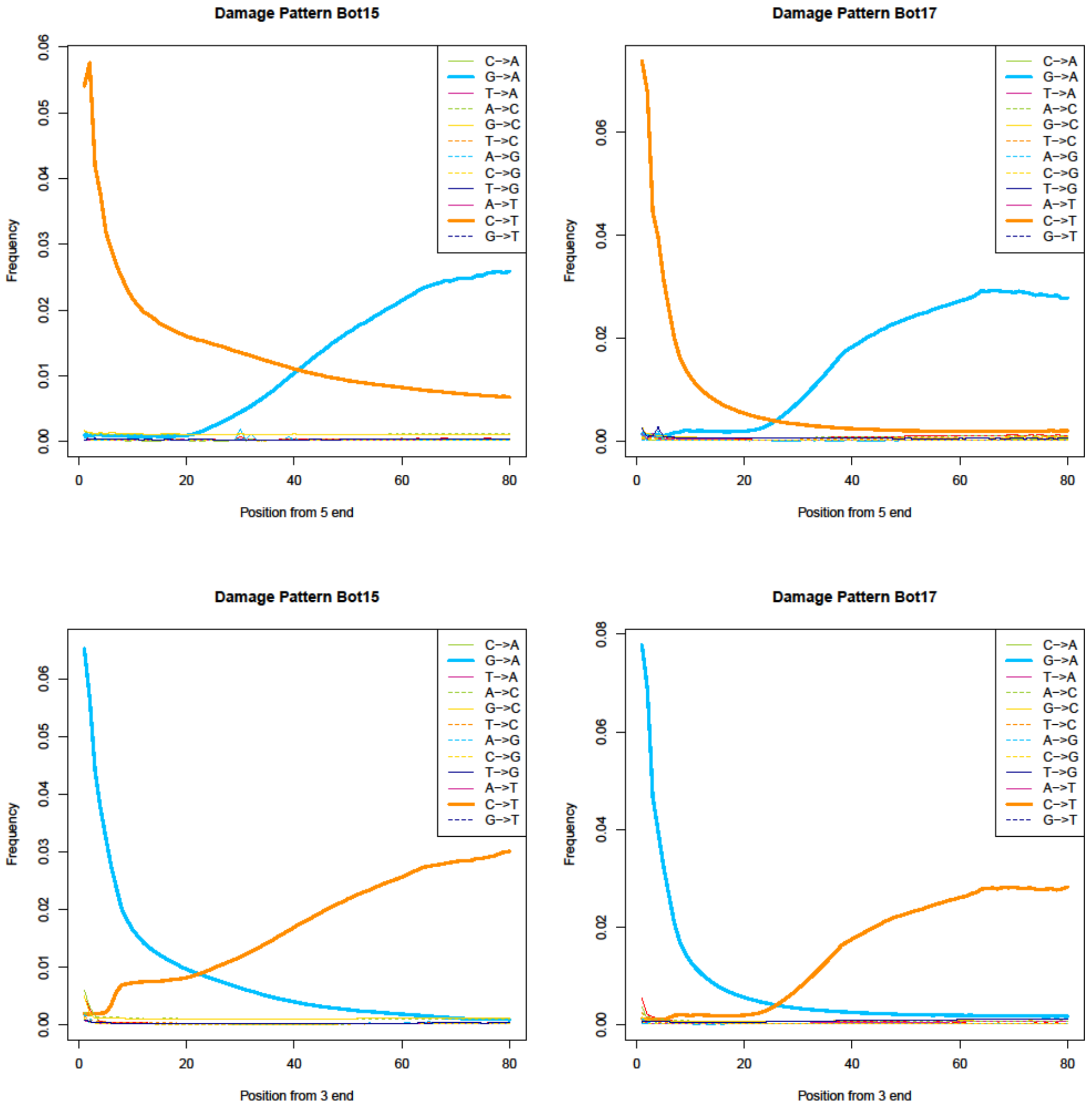
Supplementary Table 4.2 Counts of minor reads for Bot17 around known polymorphic sites on the X chromosome. A minor read is a read that is less or equally common to a major read. The top row is the relative position respective to the polymorphic site.

Contamination estimate (%)	X chromosome		mtDNA, shotgun experiment (credible intervals)
	Test1	Test2	
Bot15	2.5±0.2	2.2±0.3	1.5-3.6
Bot17	0.57±0.14	0.33±0.20	0.0-0.8

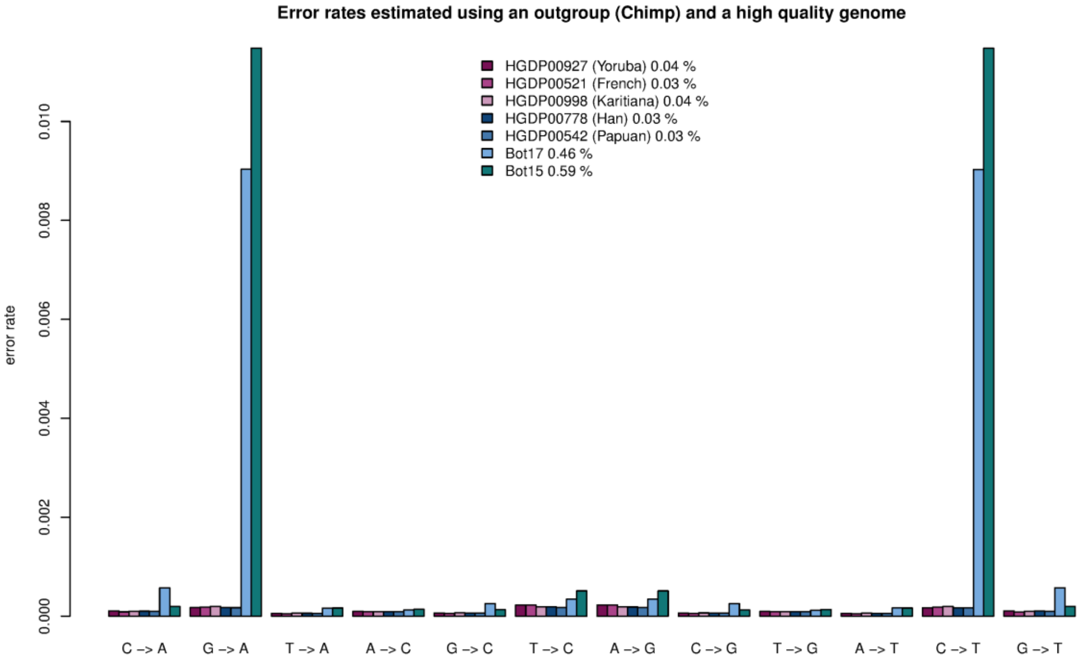
Supplementary Table 4.3 We report the contamination fraction in % with standard error for Bot15 and Bot17. For comparison, we also reproduce here the 95% credible interval of the contamination fraction for the mtDNA.



Supplementary Figure 4.1 Read length distribution for the shotgun experiments for Bot15 and Bot17. For Bot15 the data was merged from different sequencing runs with different number of cycles (92 and 94), while for Bot17 the number of cycles was 94 for all runs.



Supplementary Figure 4.2 Frequencies of observing a given nucleotide (e.g. T) in Bot15 (left) and Bot17 (right) conditioning on the reference allele (e.g. C) and the position along the read (from both 5' and 3').



Supplementary Figure 4.3 Type specific error rates for the whole genome data. The overall error rates are shown in the legend.

Supplementary 5
Ancestry of Bot15 and Bot17 based on shotgun nuclear data

Ida Moltke*, Anders Albrechtsen, Michael DeGiorgio, J. Víctor Moreno-Mayar , Anna-Sapfo Malaspinas*

*to whom correspondence should be addressed (ida@binf.ku.dk, annasapfo@gmail.com)

Publicly available datasets

To determine the ancestry of Bot15 and Bot17 we compiled several relevant datasets including whole genome data and SNP chip data.

Whole genome data

The whole genome data comprises data from five individuals from France, China (Han), Brazil (Karitiana), Papua New Guinea and Nigeria (Yoruba, see S3 for assembly details) all sequenced with at least 20X depth (Supplementary Table 5.1).

SNP chip data

The SNP chip data was compiled from three different studies (Wollstein et al. 2010; Xing et al. 2010; Reich et al. 2012) and has be chosen to include worldwide populations including Native American and Polynesian populations.

We merged the data into two main datasets: (1) "Wollstein_Xing" with data from Wollstein et al. and Xing et al. This data contained 46 Polynesian individuals from seven islands (Cook Islands, Futuna, Niue, Samoa, Tonga, Tokelau and Tuvalu) and 48 Native Americans (Totonacs and Bolivians). (2) "Reich_Wollstein" with data from Reich at al. and Wollstein et al. These data includes 19 Polynesian samples and samples from 52 Native American populations (from North and South America).

Wollstein_Xing The Wollstein et al. and Xing et al. genotypes were obtained on Affymetrix SNP arrays. Wollstein et al. shared their data in plink format with a total of 294 individuals (selected to have low European admixture) and 869,019 SNPs. This dataset included individuals from the HapMap project (Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT), Yoruba from Ibadan, Nigeria (YRI), and U.S. European Americans from Utah with Northern and Western European ancestry (CEU, Frazer et al. 2007)) as well as data generated for their own study. We excluded individuals NA18193 and NA19238 that were identified as being close relatives using *KING* (Manichaikul et al. 2010) as has been reported before (Roberson and Pevsner 2009). We lifted over the coordinates to HG19 using liftover[1].

[1] http://genome.ucsc.edu/cgi-bin/hgLiftOver

Xing at al. shared the raw data from their study (in Birdseed format[2]). We excluded from this dataset (1) the HapMap CEU individuals that were not included in their publication, (2) individual F089339 that proved to be a mislabeling as mentioned in their study, and (3) second-degree relatives as determined by *KING* (Manichaikul et al. 2010). The final dataset included 291 individuals and 868,265 SNPs.

Before merging, we chose the alleles to match the Database of Single Nucleotide Polymorphisms (dbSNP Build ID: 132, Sherry et al. (2001)) and only kept bi-allelic sites. We then filtered out SNPs with more than 10% missingness for the merged dataset. The merged dataset included 583 individuals and over 820,000 SNPs (over 250,000 excluding transitions) (Supplementary Table 5.2).

Reich_Wollstein Reich et al. shared the exact datasets used in their study (with the filtering they performed). These data include samples from several studies, as for example HapMap3 (The International HapMap 3 Consortium 2010) and CEPH-HGDP (Li et al. 2008). Genotyping was performed on Illumina arrays for all merged datasets (HapMap3 data was also genotyped on Affymetrix). The data was carefully curated by Reich et al. and included 52 Native American populations (the HapMap MEX population not being one of them). Since this data did not contain any Polynesian populations, we merged it with the individuals from Borneo, Fiji, NewGuineaHighlands and Polynesian populations of Wollstein et al. retaining only overlapping positions, bi-allelic sites and checking for matching strands. This merging decreased substantially the number of sites because the two datasets were produced on different platforms; only about 1/3 of the SNPs from the Reich et al. data remained after merging (*i.e.*, 108,662 SNPs). The remaining SNPs were mostly transitions (only 1,376 transversions) and we therefore did not exclude transitions from this dataset in what follows.

Average gene tree

To get a rough idea of the ancestry of our two samples we first built a phylogenetic tree based on the whole genome data. Although phylogenetics is not *a priori* the appropriate tool to describe the history of populations (potential migration or admixture events are not modeled for example), it approximates the average gene tree, and can thus be a good starting point to assess population affinities (see *e.g.*, Pickrell and Pritchard (2012)).

Since Bot15 and Bot17 were sequenced at low depth, we did not attempt to call genotypes. We first filtered the data with a procedure similar to the one used in ((Reich et al. 2011) see also *e.g.*, Orlando et al. (2013)): all reads with mapping quality below 30 were removed in the assembly step (S3). Subsequently, we removed low quality bases by dividing them into eight base categories (A, C, G, T on the plus strand and A, C, G, T on the minus strand) and then discarding the 50% of the bases with the lowest quality score from each of the eight categories. More specifically within each base category we:

1. identified the highest base quality score, Q, for which less than half of the bases in the base category had a quality score smaller than Q,
2. removed all bases with quality score smaller than Q,

[2] http://www.broadinstitute.org/mpg/birdsuite/birdseed.html

3. and randomly sampled and removed bases with quality score equal to Q until 50% of the bases from the base category had been removed in total.
We then sampled one read at every position as has been done before (e.g., Green et al. (2010)) for each site and each individual. The result is a depth either 0 or 1X per site for each individual.

All transitions were then removed given the increased C to T and G to A transitions due to ancient DNA damage (S4). Identical filtering was performed for the ABBA-BABA (D-statistic) results described below.

For the alignment of the phylogenetic analysis, we then only retained sites 1) where all seven individuals have a depth of 1X, 2) with exactly two observed alleles (i.e. assuming no recurrent mutations), 3) that are reported as SNPs in the dbSNP Build ID: 138[3], which includes over 51 million SNPs (Sherry et al. 2001). We performed the latter filtering because the average error rates were higher for the ancient low-depth genomes when compared to the high-depth modern genomes, even when excluding transitions (S4), as the difference in sequenced depth makes the variance in error larger in the lower depth genomes. This procedure led to an alignment of 459,902 sites for seven individuals.

We built a maximum likelihood tree using RAxML (Stamatakis 2006) with a GTR model of substitution, allowing rates to vary across sites (-m GTRGAMMA) and using the Yoruba as an outgroup (result shown in Supplementary Figure 5.1).

We found that on average, the two Native American Botocudo individuals cluster with each other and form a monophyletic group with the Papuan genome and not the Brazilian (Karitiana) genome. We then tested for departure from this tree using the D-statistic or ABBA-BABA test.

ABBA-BABA test: principle and notation

To investigate the potential admixture events between our two ancient samples and the modern human genomes we performed an ABBA-BABA test also referred to as the D-statistic (Durand et al. 2011). In this test sequencing data from Bot15, Bot17 and other individuals are compared using their allelic differences to sequencing data from an outgroup, in this case a chimpanzee. More specifically we test for departure from the tree topology (((H1, H2), H3), outgroup).

The test is performed using one randomly sampled allele from each of the individuals at each site and is only based on sites that are consistent with incomplete lineage sorting assuming no recurrent mutations. If we denote the allele present in the outgroup, which in this case is the chimpanzee, as A and the other observed allele B, the sites used must all have one of two polymorphic patterns: the “ABBA” pattern and the “BABA” pattern. In the ABBA pattern, the individual from H1 has the A allele, while H2 and H3 have the B allele and in the BABA pattern, the individual from H2 has the A allele, while H1 and H3 have the B allele. Assuming only one mutation event has occurred, both patterns are inconsistent with the specified tree. The null

[3] ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/ 00-All.vcf

hypothesis that is tested is that the tree is correct and that there has been no gene flow between H3 and H1 or H2 or structure in the ancestral population of H1, H2 and H3 that persists until the H1-H2 split. Under this hypothesis, the two patterns, ABBA and BABA must both be due to incomplete lineage sorting and are equally likely to occur. Hence, we expect there to be equally many sites with the ABBA pattern as there are sites with the BABA pattern. On the other hand, if the tree is incorrect or there has been gene flow between H3 and H1 or H2 or ancestral population structure then we expect to observe a departure from this equilibrium. We can thus test the null hypothesis using the following test statistic introduced in Green et al. (2010):

D = (nABBA-nBABA)/ (nABBA+nBABA)

where *nABBA* is the number of ABBA sites and *nBABA* is the number of BABA sites.

A positive value of D can be interpreted as H2 being closer to H3 than H1 is to H3. A negative value of D can be interpreted as H1 being closer to H3 than H2 is. Assuming the tree is correct, a test statistic that differs significantly from 0 is evidence of gene flow or ancestral structure (assuming no contamination or differential error rates).

As in Green et al. (2010) and Reich et al. (2010), we estimated the standard error of the tests statistic using "delete-m Jackknife for unequal m" (Busing, Meijer, and Leeden 1999): we divided the genome into 5 Mb blocks, calculated the test statistic leaving each of these blocks out in turn and estimated the standard error using the resulting test statistics each weighted according to the number of ABBA and BABA sites used to calculate the test statistic.

We then computed a Z-score assuming that the D-statistic is normally distributed with a standard deviation equal to the jackknife standard deviation estimate. Following (Green et al. 2010; Reich et al. 2011) absolute D-values higher than 3.0 were considered to be significant deviations from the null hypothesis.

ABBA-BABA test: results

We analyzed the samples shown in Supplementary Table 5.1 and used the same filtering and sampling approach as was used in the average gene tree analysis. The results of the ABBA-BABA tests are shown in Supplementary Table 5.3 with either Bot15 or Bot17 as H3. We only include these results because we have shown in (Rasmussen et al. 2011) that results from tests with ancient genomes, or other genomes with different error rates, as H1 or H2 can lead to wrong conclusions.

Two aspects complicate the interpretation of the result: (1) the lack of genomes close to Bot15 and Bot17 such as a Polynesian genome. (2) The ancient admixture of the Denisovan into the Papuan (Meyer et al. 2012) and an early migration into the region (Rasmussen et al. 2011). The latter point complicates the interpretation when including the Papuan in the analysis.

None of the results show evidence of admixture between the Botocudos and the Native Americans. The tests ((Han, Karitiana),Bot15) and ((Han, Karitiana), Bot17) are both rejected with D-stat>0. This suggests that Han is closer than Karitiana to Bot15 and Bot17. If there had been a significant amount of admixture we would expect Bot15 and Bot17 to be closer to Native Americans than to the Han and not the other way around.

We conclude from the D-stastistic results that we have not found evidence of Native American/Botocudo admixture, but that a Polynesian genome is needed to draw more solid conclusions. In what follows, we compare the whole genome data from Bot15 and Bot17 to the genotype data that includes Polynesian individuals.

Multidimensional scaling analysis (MDS): methods

We visualize the similarity between the Botocudos and worldwide populations using a multidimensional scaling (MDS) analysis approach. To compute a pairwise genetic distance between the whole genome data (low depth) and the genotype data, we use a similar sampling approach as described above (see Malaspinas et al. (2014) for details).

MDS analysis: results We used the two datasets described in Supplementary Table 5.2 as reference datasets and present results for Wollstein_Xing for the case without transitions in Supplementary Figure 5.2 and results for Reich_Wollstein for all SNPs in Supplementary Figure 5.3. For Wollstein_Xing, we obtain almost identical results with (not shown) or without transitions. For each of the two datasets we consider: (1) all the data and (2) subsets of the data: East Asians , Native American, Oceanian, Siberian (only in Reich_Wollstein) and Greenlandic (only in Reich_Wollstein) populations. For all cases, we find that the Botocudos cluster with the Polynesian samples. We also considered the case with only Polynesians (Supplementary Figure 5.3), which suggests that the Botocudos are closer to the Cook Islands rather than the other islands. This remains inconclusive though, as only a few individuals are included per island.

Admixture analysis

Methods

Since the two ancient genomes have only been sequenced at low depth, most of their genotypes can only be called with very high uncertainty. We therefore used the software program NGSadmix (Skotte, Korneliussen, and Albrechtsen 2013) to perform the admixture analyses. NGSadmix is a maximum likelihood method that is based on a model similar to the model that underlies other maximum likelihood-based admixture inference methods such as Frappe and Admixture (Tang et al. 2005; Alexander, Novembre, and Lange 2009). The crucial difference is, that whereas all other admixture methods base their inference on called genotypes and implicitly assume that the genotypes are called without error, NGSadmix bases its inference on genotype likelihoods and in this way takes into account the uncertainty of the genotypes that is inherently present in next generation sequencing data, especially in low depth data.

Datasets

We performed analyses of Bot15 and Bot17 in subsets of the two datasets described above.

The first dataset is based on raw read data from Bot15, Bot17 and the five modern genomes (Supplementary Table 5.1) combined with a subset of the Wollstein_Xing SNP chip dataset (Supplementary Table 5.2). The sequencing data from the modern genomes were included to show that analyses that include both SNP data and sequencing data indeed lead to sensible results.

The second dataset is based on raw read data from Bot15 and Bot17 combined with a subset of Reich_Wollstein (Supplementary Table 5.2). In this dataset we included all the Polynesians from Wollstein et al. and the Human Genome Diversity Panel (HGDP-CEPH) Han population (that we abbreviate "Han" in what follows). Moreover, we included a subset of 5 Native American and 3 Siberian populations from Reich et al. which we selected as follows: we only included "nonadmixed" individuals, i.e. individuals for which the combined %European and %African admixture is below 0.025% (as estimated and defined in Reich et al.). For the Siberian populations, we were thereby left with a total of 56 samples from three populations (Naukan, Koryak and Chukchis) and from two different linguistic families (Eskimo-Aleut, and Chukchi-Kamchatkan). Among the Native American populations, we considered only populations with nine or more individuals left and chose five populations (totaling 84 individuals) representative of several major linguistic families, including two populations from Brazil: Mixe (Central Amerind), Karitiana and Surui (Brazil, Equatorial–Tucanoan), Pima (Nortern Amerind) and Cabecar (Chibchan–Paezan).

Data filtering

Before performing the admixture analyses both SNP chip datasets were filtered by removing

- all non-autosomal SNPs
- all SNPs with more than 0.05 missingness
- all SNPs with a minor allele frequency lower than 0.05

From the Wollstein_Xing dataset we also removed all transitions. After this filtering the two SNP datasets contained 182,723 and 79,580 SNPs, respectively.

The sequence data were filtered by removing all reads with a mapping quality below 30 and all bases with a base quality score below 20.

Data processing

NGSadmix analyses are as explained above based on genotype likelihoods. For the sequenced genomes these likelihoods were generated using the software package ANGSD (Korneliussen 2013), which implements the method from samtools (Li et al. 2009). We only generated likelihoods for the SNP sites that were in the SNP chip datasets and only used the likelihoods for the three genotypes observed in the SNP datasets. For the SNP chip data we assumed no errors and thus the genotype likelihoods were set to 1 for the observed genotype and 0 otherwise. For sites not covered by any reads, the genotype likelihoods for all three possible genotype were set to 1/3.

Admixture Results

Wollstein_Xing

We ran NGSadmix with the number of population components, K, set to 2-6. Six was picked as the highest K because it was the lowest K value for which the Polynesians were estimated to have their own cluster. For each K value, we ran NGSadmix 25 times with different starting

values and in all cases the difference in likelihood units between the highest likelihood and the 10th highest likelihood was at most 1 for any K, suggesting convergence was achieved.
The results can be seen in Supplementary Figure 5.4. Notably, for K=6 both Bot15 and Bot17 is estimated to have >0.99 of their ancestry from the Polynesian cluster. Also, all five modern sequenced individuals were estimated to belong to the same clusters as the SNP chip genotyped individuals from the same population, as expected.

Reich_Wollstein
We ran NGSadmix with the number of population components, K, set to 2-7. Seven was picked as the highest K value, because it was the lowest K value for which the Polynesians were assigned their own cluster. For each K value we ran NGSadmix 25 times with different starting values. The difference in likelihood units between the highest likelihood and the 10th highest likelihood was at most 0.5 for any K, suggesting the convergence was achieved.
The results can be seen in Supplementary Figure 5.5. Notably, for K=7 both Bot15 and Bot17 is estimated to respectively have >0.9999 and 0.9976 of their ancestry from the cluster, which all the Polynesian have most of their ancestry from.

Simulating admixture
The admixture results suggest that the ancestors of the Botocudo individuals did not admix with Native Americans. However, it is possible that we cannot detect low levels of admixture with our analysis because of the low depth in the two individuals or because the admixture proportion is too low. To investigate the effect of low depth on the results, we simulated low levels (1-10%) of Native American admixture in Bot17 (the individual with the lowest depth) and ran NGSadmix.

Methods
Our simulations correspond to a very simplified admixture event scenario that is meant to match the admixture model implied by NGSadmix. We downsampled the Karitiana genome such that its depth of coverage is similar to Bot17 and obtained genotype likelihoods as described above. We then replaced the genotypes likelihoods from Bot17 with the genotype likelihoods obtained for the downsampled Karitiana genome at 1%, 2%, …, 10% randomly selected set of the sites included in the Wollstein_Xing dataset. We ran NGSadmix with K=6 (the minimum K value for which there are separate Polynesian and Native American components) for the Wollstein_Xing dataset including the artificially admixed Bot17 individual.

Results

The results of the admixture proportions corresponding with the Native American cluster are shown in Supplementary Figure 5.6. We found a strong correlation (r=0.99) between the simulated and the measured admixture proportions, while the Native American component can be detected even with admixture as low as 1%. Note that our simulations are unrealistic in many ways: for example, the Karitiana genome is not necessarily representative of the ancestral Native American population that would have admixed with the Botocudos, and we would expect some variance in the admixture proportions in each individual today given a pulse of a certain magnitude. We are also ignoring the fact that the sites are not independent. Keeping in mind those caveats, our results suggest that we have enough data to detect a potential admixture event.

References

Alexander, David H., John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research*, July. doi:10.1101/gr.094052.109.

Busing, Frank M. T. A., Erik Meijer, and Rien Van Der Leeden. 1999. “Delete-M Jackknife for Unequal M.” *Statistics and Computing* 9 (1): 3–8. doi:10.1023/A:1008800423698.

Durand, Eric Y., Nick Patterson, David Reich, and Montgomery Slatkin. 2011. “Testing for Ancient Admixture between Closely Related Populations.” *Molecular Biology and Evolution* 28 (8): 2239–52. doi:10.1093/molbev/msr048.

Frazer, Kelly A., Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, John W. Belmont, et al. 2007. “A Second Generation Human Haplotype Map of over 3.1 Million SNPs.” *Nature* 449 (7164): 851–61. doi:10.1038/nature06258.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. “A Draft Sequence of the Neandertal Genome.” *Science* 328 (5979): 710–22. doi:10.1126/science.1188021.

Korneliussen, Thorfinn. 2013. *Angsd*. University of Copenhagen, Denmark. http://www.popgen.dk/angsd.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Li, Jun Z., Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, et al. 2008. “Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation.” *Science* 319 (5866): 1100–1104. doi:10.1126/science.1153717.

Malaspinas, Anna-Sapfo, Ole Tange, José Víctor Moreno-Mayar, Morten Rasmussen, Michael DeGiorgio, Yong Wang, Cristina E. Valdiosera, Gustavo Politis, Eske Willerslev, and Rasmus Nielsen. 2014. “Bammds: A Tool for Assessing the Ancestry of Low-Depth Whole-Genome Data Using Multidimensional Scaling (MDS).” *Bioinformatics*, June, btu410. doi:10.1093/bioinformatics/btu410.

Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michele Sale, and Wei-Min Chen. 2010. “Robust Relationship Inference in Genome-Wide Association Studies.” *Bioinformatics* 26 (22): 2867–73. doi:10.1093/bioinformatics/btq559.

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, et al. 2012. “A High-Coverage Genome Sequence from an Archaic Denisovan Individual.” *Science* 338 (6104): 222–26. doi:10.1126/science.1224344.

Orlando, Ludovic, Aurélien Ginolhac, Guojie Zhang, Duane Froese, Anders Albrechtsen, Mathias Stiller, Mikkel Schubert, et al. 2013. “Recalibrating Equus Evolution Using the Genome

Sequence of an Early Middle Pleistocene Horse." *Nature* 499 (7456): 74–78. doi:10.1038/nature12323.

Pickrell, Joseph K., and Jonathan K. Pritchard. 2012. "Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data." *PLoS Genet* 8 (11): e1002967. doi:10.1371/journal.pgen.1002967.

Rasmussen, Morten, Xiaosen Guo, Yong Wang, Kirk E. Lohmueller, Simon Rasmussen, Anders Albrechtsen, Line Skotte, et al. 2011. "An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia." *Science* 334 (6052): 94–98. doi:10.1126/science.1211177.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327): 1053–60. doi:10.1038/nature09710.

Reich, David, Nick Patterson, Desmond Campbell, Arti Tandon, Stephane Mazieres, Nicolas Ray, Maria V. Parra, et al. 2012. "Reconstructing Native American Population History." *Nature* 488 (7411): 370–74. doi:10.1038/nature11258.

Reich, David, Nick Patterson, Martin Kircher, Frederick Delfin, Madhusudan R. Nandineni, Irina Pugach, Albert Min-Shan Ko, et al. 2011. "Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania." *The American Journal of Human Genetics* 89 (4): 516–28. doi:10.1016/j.ajhg.2011.09.005.

Roberson, Elisha D. O., and Jonathan Pevsner. 2009. "Visualization of Shared Genomic Regions and Meiotic Recombination in High-Density SNP Data." *PLoS ONE* 4 (8): e6711. doi:10.1371/journal.pone.0006711.

Sherry, S. T., M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. "dbSNP: The NCBI Database of Genetic Variation." *Nucleic Acids Research* 29 (1): 308–11. doi:10.1093/nar/29.1.308.

Skotte, Line, Thorfinn Sand Korneliussen, and Anders Albrechtsen. 2013. "Estimating Individual Admixture Proportions from Next Generation Sequencing Data." *Genetics* 195 (3): 693–702. doi:10.1534/genetics.113.154138.

Stamatakis, Alexandros. 2006. "RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models." *Bioinformatics* 22 (21): 2688–90. doi:10.1093/bioinformatics/btl446.

Tang, Hua, Jie Peng, Pei Wang, and Neil J. Risch. 2005. "Estimation of Individual Admixture: Analytical and Study Design Considerations." *Genetic Epidemiology* 28 (4): 289–301. doi:10.1002/gepi.20064.

The International HapMap 3 Consortium. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58. doi:10.1038/nature09298.

Wollstein, Andreas, Oscar Lao, Christian Becker, Silke Brauer, Ronald J. Trent, Peter Nurnberg, Mark Stoneking, and Manfred Kayser. 2010. "Demographic History of Oceania Inferred from Genome-Wide Data." *Current Biology* 20 (22): 1983–92. doi:10.1016/j.cub.2010.10.040.

Xing, Jinchuan, W. Scott Watkins, Adam Shlien, Erin Walker, Chad D. Huff, David J. Witherspoon, Yuhua Zhang, et al. 2010. "Toward a More Uniform Sampling of Human Genetic

Diversity: A Survey of Worldwide Populations by High-Density Genotyping." *Genomics* 96 (4): 199–210. doi:10.1016/j.ygeno.2010.07.004.

Sample	Population	Reference	Average Depth (X)	Covered > 1X (%)
HGDP00521	French	(Meyer et al. 2012)	22.6	90
HGDP00542	Papuan	(Meyer et al. 2012)	21.6	90
HGDP00778	Han	(Meyer et al. 2012)	22.3	90
HGDP00927	Yoruba	(Meyer et al. 2012)	26.7	90
HGDP00998	Karitiana	(Meyer et al. 2012)	21.3	90

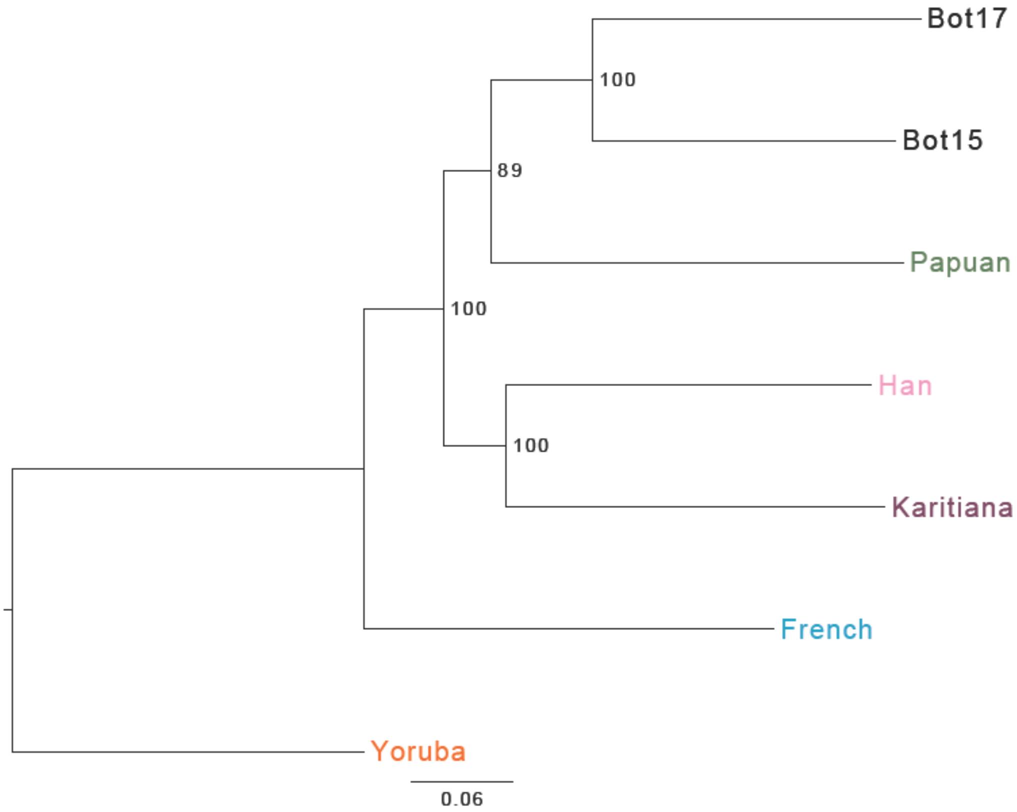
Supplementary Table 5.1 Assembly details of the five human genomes used for comparison.

	#number of SNPs	#number of SNPs (excl. transitions)	#populations	#individuals	#Polynesian individuals	#Native American individuals
Wollstein_Xing	823,805	256,261	20	583	45	48
Reich_Wollstein	108,662	1,376	130	2,436	19	493

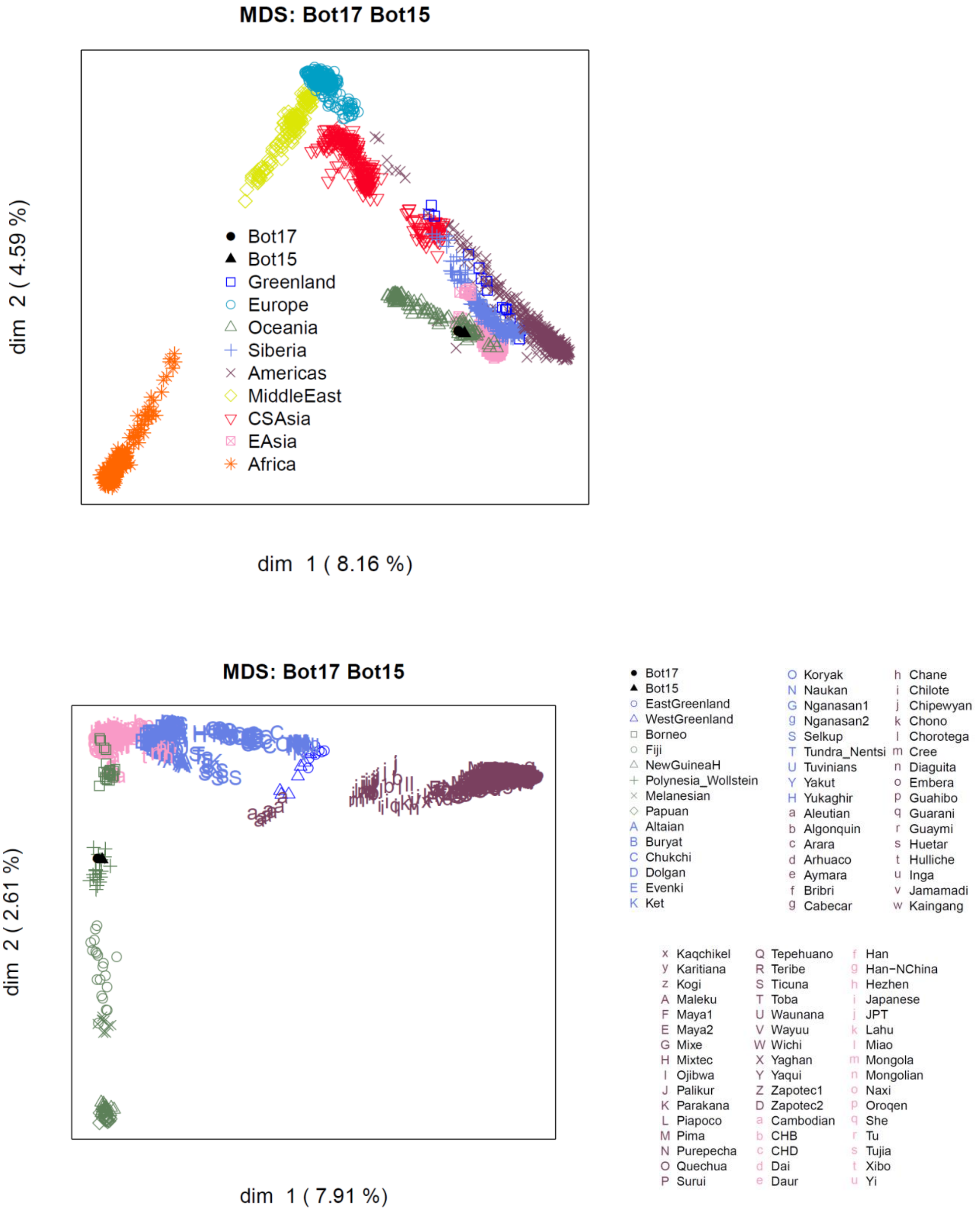
Supplementary Table 5.2 Summary of the genotype data used for analysis.

		H3 = Bot17				**H3 =Bot15**			
H1	H2	nABBA-nBABA	nABBA+nBABA	D	Z	nABBA-nBABA	nABBA+nBABA	D	Z
French	**Papuan***	**10066**	**120822**	**0.083**	**12.1**	**10937**	**127043**	**0.086**	**12.4**
French	**Han***	**15769**	**118719**	**0.133**	**21.2**	**17216**	**124768**	**0.138**	**22.7**
Papuan	**Han***	**5661**	**119379**	**0.047**	**6.4**	**6264**	**125452**	**0.05**	**6.7**
French*	**Yoruban**	**-38650**	**132502**	**-0.292**	**-55.3**	**-39880**	**139192**	**-0.287**	**-55.3**
Papuan*	**Yoruban**	**-48739**	**140215**	**-0.348**	**-58.9**	**-50817**	**147363**	**-0.345**	**-57.6**
Han*	**Yoruban**	**-54462**	**140074**	**-0.389**	**-77.5**	**-57147**	**147489**	**-0.387**	**-71.7**
French	**Karitiana***	**11394**	**115564**	**0.099**	**14.9**	**11198**	**120944**	**0.093**	**15.3**
Papuan	Karitiana	1258	118464	0.011	1.4	294	124804	0.002	0.3
Han*	**Karitiana**	**-4505**	**110325**	**-0.041**	**-6.5**	**-6025**	**115665**	**-0.052**	**-8.4**

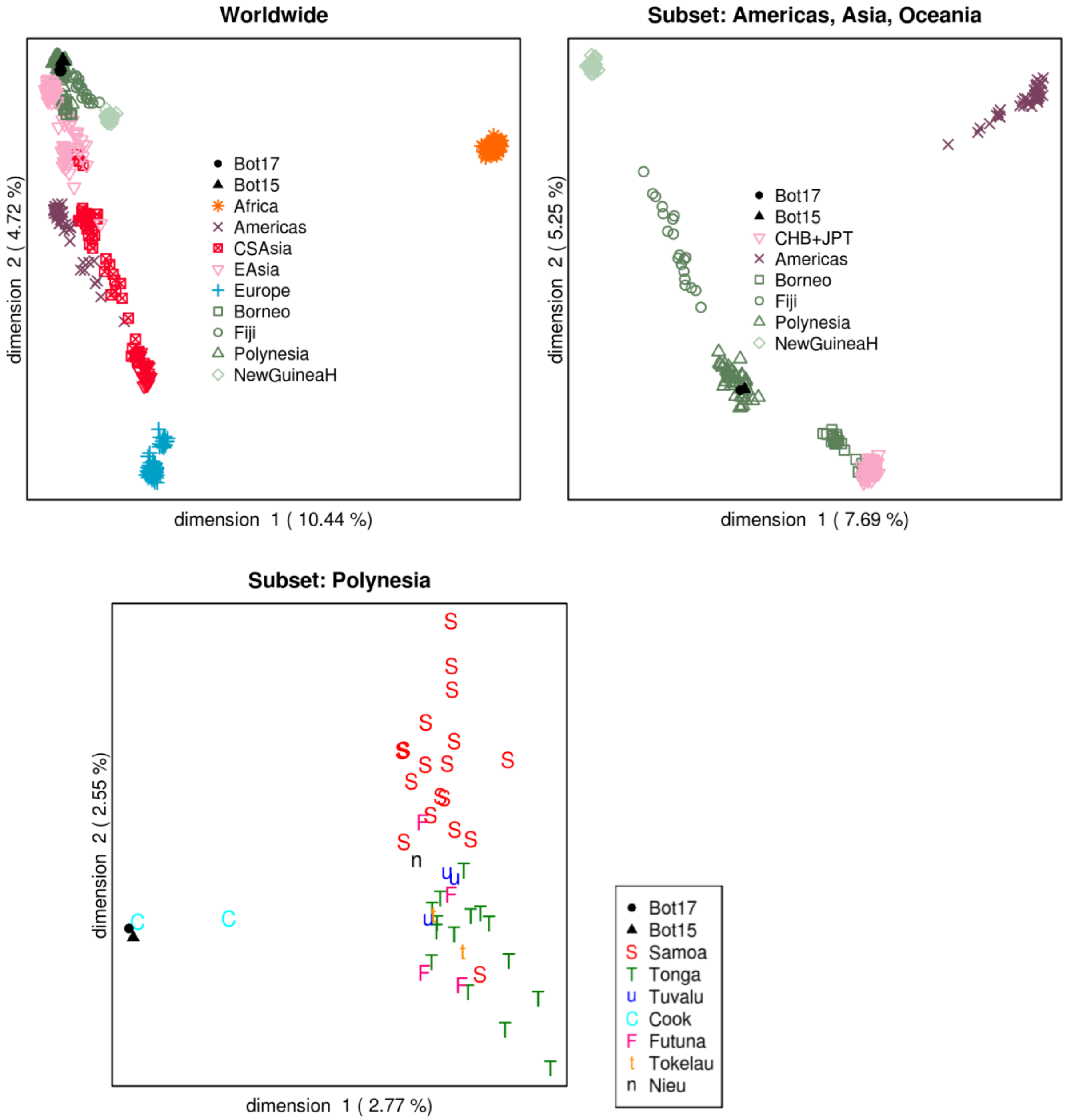
Supplementary Table 5.3 Results for the ABBA-BABA tests. The table shows ABBA-BABA *D*-statistics and Z-values based on jackknife estimation. The tests in bold are the tests that show significant deviation from the null hypothesis. * shows the population that the Bot15 and Bot17 samples is closest to.



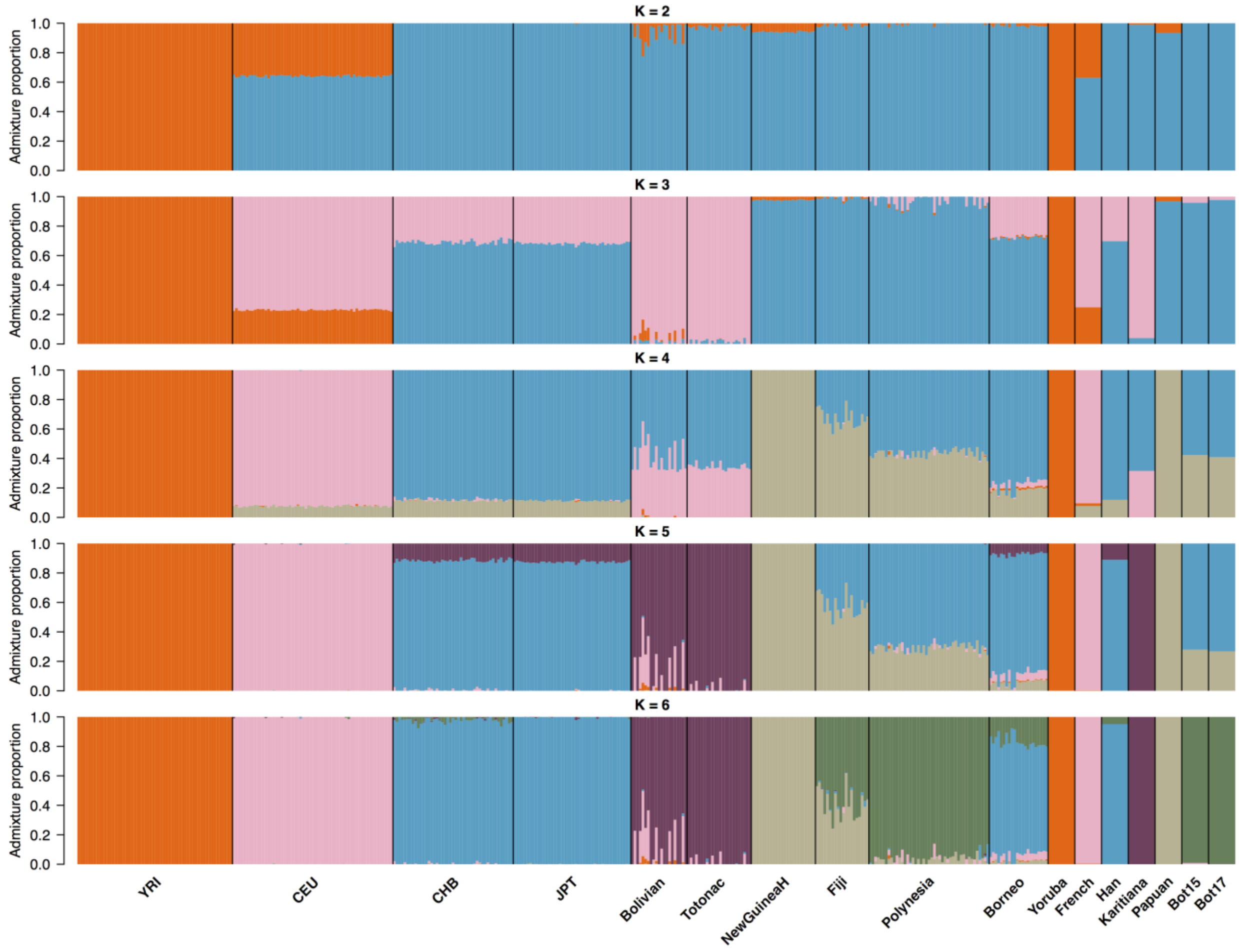
Supplementary Figure 5.1 Phylogenetic tree of the whole genome data for Bot15, Bot17 and the five modern genomes described in Supplementary Table 5.1 raxmlHPC-PTHREADS was used to build the tree with 10,000 bootstrap replicates.



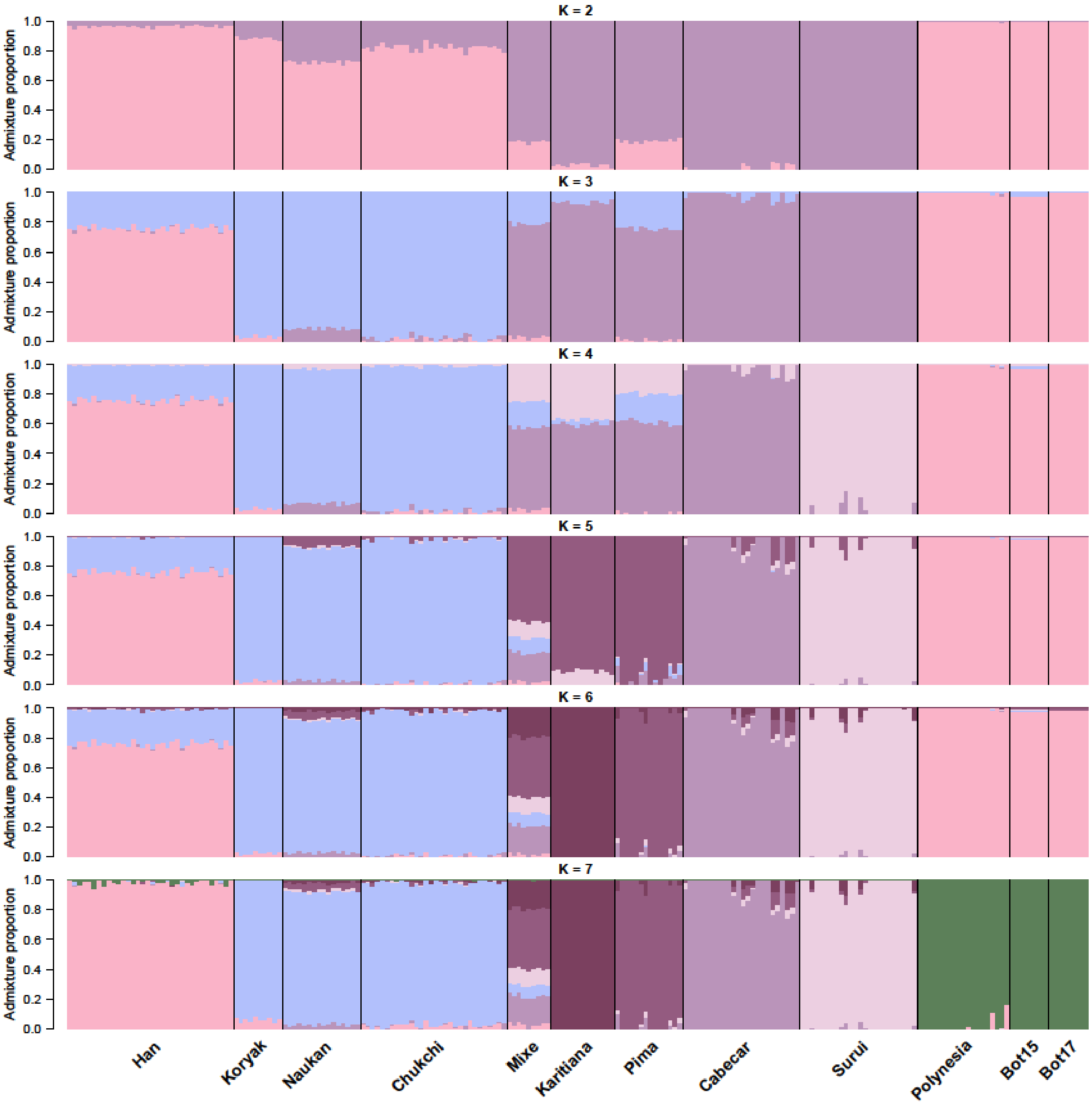
Supplementary Figure 5.2 MDS results for the Reich_Wollstein dataset. The first dimensions of the MDS analysis for all the populations in Reich_Wollstein (top) and a subset of the populations including all Native American, Oceanian, Siberian and Greenlandic populations (bottom). The Botocudos are shown in filled black triangle and circle and clusters with the Polynesians in both plots.



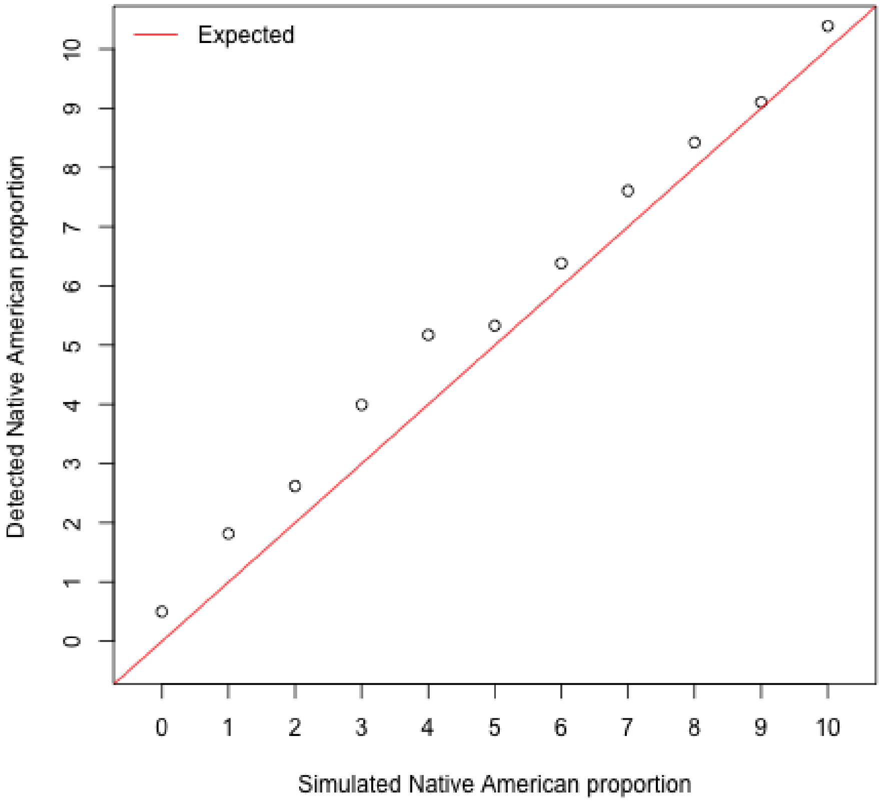
Supplementary Figure 5.3 MDS results for the Wollstein_Xing dataset. Two first dimensions of the MDS analysis for all the populations (labeled by geographic regions) in Wollstein_Xing excluding transitions (top left), a subset of the populations including all Native American, Oceanian, and two East Asian populations (top right), only the Polynesian labeled by Island (bottom). The Botocudos are shown in filled black triangle and circle and clusters with the Polynesians in both top plots.



Supplementary Figure 5.4 Estimated admixture proportions for a subset of Wollstein_Xing. The SNP chip data includes African (YRI), Asian (CHB, JPT, Borneo), Oceanian (New Guinea Highlands, Fiji, Polynesia, Borneo) and Native American (Bolivian, Totonac) samples. Note the bars for the seven sequenced individuals (Yoruba, French, Han, Karitiana, Papuan, Bot15, Bot17) to the right have been made wider to make them more visible.



Supplementary Figure 5.5 Estimated admixture proportions for Reich_Wollstein. The SNP chip data includes Asian (Han), Siberian (Koryak, Naukan, Chukchi) and Native American (Mixe, Karitiana, Pima, Cabecar, Surui) samples. Note the bars for Bot15 and Bot17 to the right have been made wider to make them more visible.



Supplementary Figure 5.6 Simulated and measured admixture proportions in an artificially admixed Botocudo individual. We simulated different levels of Native American admixture by artificially substituting parts of the Bot17 with a subsampled version of the Karitiana (Brazil) genome. The simulated (X-axis) and the measured admixture proportions in the simulated individual as estimated by NGSadmix (Y-axis) are reported.

Supplementary 5
Ancestry of Bot15 and Bot17 based on shotgun nuclear data

Ida Moltke*, Anders Albrechtsen, Michael DeGiorgio, J. Víctor Moreno-Mayar , Anna-Sapfo Malaspinas*

*to whom correspondence should be addressed (ida@binf.ku.dk, annasapfo@gmail.com)

Publicly available datasets

To determine the ancestry of Bot15 and Bot17 we compiled several relevant datasets including whole genome data and SNP chip data.

Whole genome data

The whole genome data comprises data from five individuals from France, China (Han), Brazil (Karitiana), Papua New Guinea and Nigeria (Yoruba, see S3 for assembly details) all sequenced with at least 20X depth (Supplementary Table 5.1).

SNP chip data

The SNP chip data was compiled from three different studies (Wollstein et al. 2010; Xing et al. 2010; Reich et al. 2012) and has be chosen to include worldwide populations including Native American and Polynesian populations.

We merged the data into two main datasets: (1) "Wollstein_Xing" with data from Wollstein et al. and Xing et al. This data contained 46 Polynesian individuals from seven islands (Cook Islands, Futuna, Niue, Samoa, Tonga, Tokelau and Tuvalu) and 48 Native Americans (Totonacs and Bolivians). (2) "Reich_Wollstein" with data from Reich at al. and Wollstein et al. These data includes 19 Polynesian samples and samples from 52 Native American populations (from North and South America).

Wollstein_Xing The Wollstein et al. and Xing et al. genotypes were obtained on Affymetrix SNP arrays. Wollstein et al. shared their data in plink format with a total of 294 individuals (selected to have low European admixture) and 869,019 SNPs. This dataset included individuals from the HapMap project (Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT), Yoruba from Ibadan, Nigeria (YRI), and U.S. European Americans from Utah with Northern and Western European ancestry (CEU, Frazer et al. 2007)) as well as data generated for their own study. We excluded individuals NA18193 and NA19238 that were identified as being close relatives using *KING* (Manichaikul et al. 2010) as has been reported before (Roberson and Pevsner 2009). We lifted over the coordinates to HG19 using liftover[1].

[1] http://genome.ucsc.edu/cgi-bin/hgLiftOver

Xing at al. shared the raw data from their study (in Birdseed format[2]). We excluded from this dataset (1) the HapMap CEU individuals that were not included in their publication, (2) individual F089339 that proved to be a mislabeling as mentioned in their study, and (3) second-degree relatives as determined by *KING* (Manichaikul et al. 2010). The final dataset included 291 individuals and 868,265 SNPs.

Before merging, we chose the alleles to match the Database of Single Nucleotide Polymorphisms (dbSNP Build ID: 132, Sherry et al. (2001)) and only kept bi-allelic sites. We then filtered out SNPs with more than 10% missingness for the merged dataset. The merged dataset included 583 individuals and over 820,000 SNPs (over 250,000 excluding transitions) (Supplementary Table 5.2).

Reich_Wollstein Reich et al. shared the exact datasets used in their study (with the filtering they performed). These data include samples from several studies, as for example HapMap3 (The International HapMap 3 Consortium 2010) and CEPH-HGDP (Li et al. 2008). Genotyping was performed on Illumina arrays for all merged datasets (HapMap3 data was also genotyped on Affymetrix). The data was carefully curated by Reich et al. and included 52 Native American populations (the HapMap MEX population not being one of them). Since this data did not contain any Polynesian populations, we merged it with the individuals from Borneo, Fiji, NewGuineaHighlands and Polynesian populations of Wollstein et al. retaining only overlapping positions, bi-allelic sites and checking for matching strands. This merging decreased substantially the number of sites because the two datasets were produced on different platforms; only about 1/3 of the SNPs from the Reich et al. data remained after merging (*i.e.*, 108,662 SNPs). The remaining SNPs were mostly transitions (only 1,376 transversions) and we therefore did not exclude transitions from this dataset in what follows.

Average gene tree

To get a rough idea of the ancestry of our two samples we first built a phylogenetic tree based on the whole genome data. Although phylogenetics is not *a priori* the appropriate tool to describe the history of populations (potential migration or admixture events are not modeled for example), it approximates the average gene tree, and can thus be a good starting point to assess population affinities (see *e.g.*, Pickrell and Pritchard (2012)).

Since Bot15 and Bot17 were sequenced at low depth, we did not attempt to call genotypes. We first filtered the data with a procedure similar to the one used in ((Reich et al. 2011) see also *e.g.*, Orlando et al. (2013)): all reads with mapping quality below 30 were removed in the assembly step (S3). Subsequently, we removed low quality bases by dividing them into eight base categories (A, C, G, T on the plus strand and A, C, G, T on the minus strand) and then discarding the 50% of the bases with the lowest quality score from each of the eight categories. More specifically within each base category we:

1. identified the highest base quality score, Q, for which less than half of the bases in the base category had a quality score smaller than Q,
2. removed all bases with quality score smaller than Q,

[2] http://www.broadinstitute.org/mpg/birdsuite/birdseed.html

3. and randomly sampled and removed bases with quality score equal to Q until 50% of the bases from the base category had been removed in total.
We then sampled one read at every position as has been done before (e.g., Green et al. (2010)) for each site and each individual. The result is a depth either 0 or 1X per site for each individual.

All transitions were then removed given the increased C to T and G to A transitions due to ancient DNA damage (S4). Identical filtering was performed for the ABBA-BABA (D-statistic) results described below.

For the alignment of the phylogenetic analysis, we then only retained sites 1) where all seven individuals have a depth of 1X, 2) with exactly two observed alleles (i.e. assuming no recurrent mutations), 3) that are reported as SNPs in the dbSNP Build ID: 138[3], which includes over 51 million SNPs (Sherry et al. 2001). We performed the latter filtering because the average error rates were higher for the ancient low-depth genomes when compared to the high-depth modern genomes, even when excluding transitions (S4), as the difference in sequenced depth makes the variance in error larger in the lower depth genomes. This procedure led to an alignment of 459,902 sites for seven individuals.

We built a maximum likelihood tree using RAxML (Stamatakis 2006) with a GTR model of substitution, allowing rates to vary across sites (-m GTRGAMMA) and using the Yoruba as an outgroup (result shown in Supplementary Figure 5.1).

We found that on average, the two Native American Botocudo individuals cluster with each other and form a monophyletic group with the Papuan genome and not the Brazilian (Karitiana) genome. We then tested for departure from this tree using the D-statistic or ABBA-BABA test.

ABBA-BABA test: principle and notation

To investigate the potential admixture events between our two ancient samples and the modern human genomes we performed an ABBA-BABA test also referred to as the D-statistic (Durand et al. 2011). In this test sequencing data from Bot15, Bot17 and other individuals are compared using their allelic differences to sequencing data from an outgroup, in this case a chimpanzee. More specifically we test for departure from the tree topology (((H1, H2), H3), outgroup).

The test is performed using one randomly sampled allele from each of the individuals at each site and is only based on sites that are consistent with incomplete lineage sorting assuming no recurrent mutations. If we denote the allele present in the outgroup, which in this case is the chimpanzee, as A and the other observed allele B, the sites used must all have one of two polymorphic patterns: the “ABBA” pattern and the “BABA” pattern. In the ABBA pattern, the individual from H1 has the A allele, while H2 and H3 have the B allele and in the BABA pattern, the individual from H2 has the A allele, while H1 and H3 have the B allele. Assuming only one mutation event has occurred, both patterns are inconsistent with the specified tree. The null

[3] ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/VCF/ 00-All.vcf

hypothesis that is tested is that the tree is correct and that there has been no gene flow between H3 and H1 or H2 or structure in the ancestral population of H1, H2 and H3 that persists until the H1-H2 split. Under this hypothesis, the two patterns, ABBA and BABA must both be due to incomplete lineage sorting and are equally likely to occur. Hence, we expect there to be equally many sites with the ABBA pattern as there are sites with the BABA pattern. On the other hand, if the tree is incorrect or there has been gene flow between H3 and H1 or H2 or ancestral population structure then we expect to observe a departure from this equilibrium. We can thus test the null hypothesis using the following test statistic introduced in Green et al. (2010):

D = (nABBA-nBABA)/ (nABBA+nBABA)

where *nABBA* is the number of ABBA sites and *nBABA* is the number of BABA sites.

A positive value of D can be interpreted as H2 being closer to H3 than H1 is to H3. A negative value of D can be interpreted as H1 being closer to H3 than H2 is. Assuming the tree is correct, a test statistic that differs significantly from 0 is evidence of gene flow or ancestral structure (assuming no contamination or differential error rates).

As in Green et al. (2010) and Reich et al. (2010), we estimated the standard error of the tests statistic using "delete-m Jackknife for unequal m" (Busing, Meijer, and Leeden 1999): we divided the genome into 5 Mb blocks, calculated the test statistic leaving each of these blocks out in turn and estimated the standard error using the resulting test statistics each weighted according to the number of ABBA and BABA sites used to calculate the test statistic.

We then computed a Z-score assuming that the D-statistic is normally distributed with a standard deviation equal to the jackknife standard deviation estimate. Following (Green et al. 2010; Reich et al. 2011) absolute D-values higher than 3.0 were considered to be significant deviations from the null hypothesis.

ABBA-BABA test: results

We analyzed the samples shown in Supplementary Table 5.1 and used the same filtering and sampling approach as was used in the average gene tree analysis. The results of the ABBA-BABA tests are shown in Supplementary Table 5.3 with either Bot15 or Bot17 as H3. We only include these results because we have shown in (Rasmussen et al. 2011) that results from tests with ancient genomes, or other genomes with different error rates, as H1 or H2 can lead to wrong conclusions.

Two aspects complicate the interpretation of the result: (1) the lack of genomes close to Bot15 and Bot17 such as a Polynesian genome. (2) The ancient admixture of the Denisovan into the Papuan (Meyer et al. 2012) and an early migration into the region (Rasmussen et al. 2011). The latter point complicates the interpretation when including the Papuan in the analysis.

None of the results show evidence of admixture between the Botocudos and the Native Americans. The tests ((Han, Karitiana),Bot15) and ((Han, Karitiana), Bot17) are both rejected with D-stat>0. This suggests that Han is closer than Karitiana to Bot15 and Bot17. If there had been a significant amount of admixture we would expect Bot15 and Bot17 to be closer to Native Americans than to the Han and not the other way around.

We conclude from the D-stastistic results that we have not found evidence of Native American/Botocudo admixture, but that a Polynesian genome is needed to draw more solid conclusions. In what follows, we compare the whole genome data from Bot15 and Bot17 to the genotype data that includes Polynesian individuals.

Multidimensional scaling analysis (MDS): methods

We visualize the similarity between the Botocudos and worldwide populations using a multidimensional scaling (MDS) analysis approach. To compute a pairwise genetic distance between the whole genome data (low depth) and the genotype data, we use a similar sampling approach as described above (see Malaspinas et al. (2014) for details).

MDS analysis: results We used the two datasets described in Supplementary Table 5.2 as reference datasets and present results for Wollstein_Xing for the case without transitions in Supplementary Figure 5.2 and results for Reich_Wollstein for all SNPs in Supplementary Figure 5.3. For Wollstein_Xing, we obtain almost identical results with (not shown) or without transitions. For each of the two datasets we consider: (1) all the data and (2) subsets of the data: East Asians , Native American, Oceanian, Siberian (only in Reich_Wollstein) and Greenlandic (only in Reich_Wollstein) populations. For all cases, we find that the Botocudos cluster with the Polynesian samples. We also considered the case with only Polynesians (Supplementary Figure 5.3), which suggests that the Botocudos are closer to the Cook Islands rather than the other islands. This remains inconclusive though, as only a few individuals are included per island.

Admixture analysis

Methods

Since the two ancient genomes have only been sequenced at low depth, most of their genotypes can only be called with very high uncertainty. We therefore used the software program NGSadmix (Skotte, Korneliussen, and Albrechtsen 2013) to perform the admixture analyses. NGSadmix is a maximum likelihood method that is based on a model similar to the model that underlies other maximum likelihood-based admixture inference methods such as Frappe and Admixture (Tang et al. 2005; Alexander, Novembre, and Lange 2009). The crucial difference is, that whereas all other admixture methods base their inference on called genotypes and implicitly assume that the genotypes are called without error, NGSadmix bases its inference on genotype likelihoods and in this way takes into account the uncertainty of the genotypes that is inherently present in next generation sequencing data, especially in low depth data.

Datasets

We performed analyses of Bot15 and Bot17 in subsets of the two datasets described above.

The first dataset is based on raw read data from Bot15, Bot17 and the five modern genomes (Supplementary Table 5.1) combined with a subset of the Wollstein_Xing SNP chip dataset (Supplementary Table 5.2). The sequencing data from the modern genomes were included to show that analyses that include both SNP data and sequencing data indeed lead to sensible results.

The second dataset is based on raw read data from Bot15 and Bot17 combined with a subset of Reich_Wollstein (Supplementary Table 5.2). In this dataset we included all the Polynesians from Wollstein et al. and the Human Genome Diversity Panel (HGDP-CEPH) Han population (that we abbreviate "Han" in what follows). Moreover, we included a subset of 5 Native American and 3 Siberian populations from Reich et al. which we selected as follows: we only included "nonadmixed" individuals, i.e. individuals for which the combined %European and %African admixture is below 0.025% (as estimated and defined in Reich et al.). For the Siberian populations, we were thereby left with a total of 56 samples from three populations (Naukan, Koryak and Chukchis) and from two different linguistic families (Eskimo-Aleut, and Chukchi-Kamchatkan). Among the Native American populations, we considered only populations with nine or more individuals left and chose five populations (totaling 84 individuals) representative of several major linguistic families, including two populations from Brazil: Mixe (Central Amerind), Karitiana and Surui (Brazil, Equatorial–Tucanoan), Pima (Nortern Amerind) and Cabecar (Chibchan–Paezan).

Data filtering

Before performing the admixture analyses both SNP chip datasets were filtered by removing

- all non-autosomal SNPs
- all SNPs with more than 0.05 missingness
- all SNPs with a minor allele frequency lower than 0.05

From the Wollstein_Xing dataset we also removed all transitions. After this filtering the two SNP datasets contained 182,723 and 79,580 SNPs, respectively.

The sequence data were filtered by removing all reads with a mapping quality below 30 and all bases with a base quality score below 20.

Data processing

NGSadmix analyses are as explained above based on genotype likelihoods. For the sequenced genomes these likelihoods were generated using the software package ANGSD (Korneliussen 2013), which implements the method from samtools (Li et al. 2009). We only generated likelihoods for the SNP sites that were in the SNP chip datasets and only used the likelihoods for the three genotypes observed in the SNP datasets. For the SNP chip data we assumed no errors and thus the genotype likelihoods were set to 1 for the observed genotype and 0 otherwise. For sites not covered by any reads, the genotype likelihoods for all three possible genotype were set to 1/3.

Admixture Results

Wollstein_Xing

We ran NGSadmix with the number of population components, K, set to 2-6. Six was picked as the highest K because it was the lowest K value for which the Polynesians were estimated to have their own cluster. For each K value, we ran NGSadmix 25 times with different starting values and in all cases the difference in likelihood units between the highest likelihood and the 10th highest likelihood was at most 1 for any K, suggesting convergence was achieved*.*

The results can be seen in Supplementary Figure 5.4. Notably, for K=6 both Bot15 and Bot17 is estimated to have >0.99 of their ancestry from the Polynesian cluster. Also, all five modern sequenced individuals were estimated to belong to the same clusters as the SNP chip genotyped individuals from the same population, as expected.

Reich_Wollstein

We ran NGSadmix with the number of population components, K, set to 2-7. Seven was picked as the highest K value, because it was the lowest K value for which the Polynesians were assigned their own cluster. For each K value we ran NGSadmix 25 times with different starting values. The difference in likelihood units between the highest likelihood and the 10th highest likelihood was at most 0.5 for any K, suggesting the convergence was achieved.

The results can be seen in Supplementary Figure 5.5. Notably, for K=7 both Bot15 and Bot17 is estimated to respectively have >0.9999 and 0.9976 of their ancestry from the cluster, which all the Polynesian have most of their ancestry from.

Simulating admixture

The admixture results suggest that the ancestors of the Botocudo individuals did not admix with Native Americans. However, it is possible that we cannot detect low levels of admixture with our analysis because of the low depth in the two individuals or because the admixture proportion is too low. To investigate the effect of low depth on the results, we simulated low levels (1-10%) of Native American admixture in Bot17 (the individual with the lowest depth) and ran NGSadmix.

Methods

Our simulations correspond to a very simplified admixture event scenario that is meant to match the admixture model implied by NGSadmix. We downsampled the Karitiana genome such that its depth of coverage is similar to Bot17 and obtained genotype likelihoods as described above. We then replaced the genotypes likelihoods from Bot17 with the genotype likelihoods obtained for the downsampled Karitiana genome at 1%, 2%, …, 10% randomly selected set of the sites included in the Wollstein_Xing dataset. We ran NGSadmix with K=6 (the minimum K value for which there are separate Polynesian and Native American components) for the Wollstein_Xing dataset including the artificially admixed Bot17 individual.

Results

The results of the admixture proportions corresponding with the Native American cluster are shown in Supplementary Figure 5.6. We found a strong correlation (r^2=0.99) between the simulated and the measured admixture proportions, while the Native American component can be detected even with admixture as low as 1%. Note that our simulations are unrealistic in many ways: for example, the Karitiana genome is not necessarily representative of the ancestral Native American population that would have admixed with the Botocudos, and we would expect some variance in the admixture proportions in each individual today given a pulse of a certain magnitude. We are also ignoring the fact that the sites are not independent. Keeping in mind those caveats, our results suggest that we have enough data to detect a potential admixture event.

References

Alexander, David H., John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research*, July. doi:10.1101/gr.094052.109.

Busing, Frank M. T. A., Erik Meijer, and Rien Van Der Leeden. 1999. "Delete-M Jackknife for Unequal M." *Statistics and Computing* 9 (1): 3–8. doi:10.1023/A:1008800423698.

Durand, Eric Y., Nick Patterson, David Reich, and Montgomery Slatkin. 2011. "Testing for Ancient Admixture between Closely Related Populations." *Molecular Biology and Evolution* 28 (8): 2239–52. doi:10.1093/molbev/msr048.

Frazer, Kelly A., Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, John W. Belmont, et al. 2007. "A Second Generation Human Haplotype Map of over 3.1 Million SNPs." *Nature* 449 (7164): 851–61. doi:10.1038/nature06258.

Green, Richard E., Johannes Krause, Adrian W. Briggs, Tomislav Maricic, Udo Stenzel, Martin Kircher, Nick Patterson, et al. 2010. "A Draft Sequence of the Neandertal Genome." *Science* 328 (5979): 710–22. doi:10.1126/science.1188021.

Korneliussen, Thorfinn. 2013. *Angsd*. University of Copenhagen, Denmark. http://www.popgen.dk/angsd.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Li, Jun Z., Devin M. Absher, Hua Tang, Audrey M. Southwick, Amanda M. Casto, Sohini Ramachandran, Howard M. Cann, et al. 2008. "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation." *Science* 319 (5866): 1100–1104. doi:10.1126/science.1153717.

Malaspinas, Anna-Sapfo, Ole Tange, José Víctor Moreno-Mayar, Morten Rasmussen, Michael DeGiorgio, Yong Wang, Cristina E. Valdiosera, Gustavo Politis, Eske Willerslev, and Rasmus Nielsen. 2014. "Bammds: A Tool for Assessing the Ancestry of Low-Depth Whole-Genome Data Using Multidimensional Scaling (MDS)." *Bioinformatics*, June, btu410. doi:10.1093/bioinformatics/btu410.

Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michele Sale, and Wei-Min Chen. 2010. "Robust Relationship Inference in Genome-Wide Association Studies." *Bioinformatics* 26 (22): 2867–73. doi:10.1093/bioinformatics/btq559.

Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, et al. 2012. "A High-Coverage Genome Sequence from an Archaic Denisovan Individual." *Science* 338 (6104): 222–26. doi:10.1126/science.1224344.

Orlando, Ludovic, Aurélien Ginolhac, Guojie Zhang, Duane Froese, Anders Albrechtsen, Mathias Stiller, Mikkel Schubert, et al. 2013. "Recalibrating Equus Evolution Using the Genome

Sequence of an Early Middle Pleistocene Horse." *Nature* 499 (7456): 74–78. doi:10.1038/nature12323.

Pickrell, Joseph K., and Jonathan K. Pritchard. 2012. "Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data." *PLoS Genet* 8 (11): e1002967. doi:10.1371/journal.pgen.1002967.

Rasmussen, Morten, Xiaosen Guo, Yong Wang, Kirk E. Lohmueller, Simon Rasmussen, Anders Albrechtsen, Line Skotte, et al. 2011. "An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia." *Science* 334 (6052): 94–98. doi:10.1126/science.1211177.

Reich, David, Richard E. Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y. Durand, Bence Viola, et al. 2010. "Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia." *Nature* 468 (7327): 1053–60. doi:10.1038/nature09710.

Reich, David, Nick Patterson, Desmond Campbell, Arti Tandon, Stephane Mazieres, Nicolas Ray, Maria V. Parra, et al. 2012. "Reconstructing Native American Population History." *Nature* 488 (7411): 370–74. doi:10.1038/nature11258.

Reich, David, Nick Patterson, Martin Kircher, Frederick Delfin, Madhusudan R. Nandineni, Irina Pugach, Albert Min-Shan Ko, et al. 2011. "Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania." *The American Journal of Human Genetics* 89 (4): 516–28. doi:10.1016/j.ajhg.2011.09.005.

Roberson, Elisha D. O., and Jonathan Pevsner. 2009. "Visualization of Shared Genomic Regions and Meiotic Recombination in High-Density SNP Data." *PLoS ONE* 4 (8): e6711. doi:10.1371/journal.pone.0006711.

Sherry, S. T., M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. "dbSNP: The NCBI Database of Genetic Variation." *Nucleic Acids Research* 29 (1): 308–11. doi:10.1093/nar/29.1.308.

Skotte, Line, Thorfinn Sand Korneliussen, and Anders Albrechtsen. 2013. "Estimating Individual Admixture Proportions from Next Generation Sequencing Data." *Genetics* 195 (3): 693–702. doi:10.1534/genetics.113.154138.

Stamatakis, Alexandros. 2006. "RAxML-VI-HPC: Maximum Likelihood-Based Phylogenetic Analyses with Thousands of Taxa and Mixed Models." *Bioinformatics* 22 (21): 2688–90. doi:10.1093/bioinformatics/btl446.

Tang, Hua, Jie Peng, Pei Wang, and Neil J. Risch. 2005. "Estimation of Individual Admixture: Analytical and Study Design Considerations." *Genetic Epidemiology* 28 (4): 289–301. doi:10.1002/gepi.20064.

The International HapMap 3 Consortium. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58. doi:10.1038/nature09298.

Wollstein, Andreas, Oscar Lao, Christian Becker, Silke Brauer, Ronald J. Trent, Peter Nurnberg, Mark Stoneking, and Manfred Kayser. 2010. "Demographic History of Oceania Inferred from Genome-Wide Data." *Current Biology* 20 (22): 1983–92. doi:10.1016/j.cub.2010.10.040.

Xing, Jinchuan, W. Scott Watkins, Adam Shlien, Erin Walker, Chad D. Huff, David J. Witherspoon, Yuhua Zhang, et al. 2010. "Toward a More Uniform Sampling of Human Genetic

Diversity: A Survey of Worldwide Populations by High-Density Genotyping." *Genomics* 96 (4): 199–210. doi:10.1016/j.ygeno.2010.07.004.

Sample	Population	Reference	Average Depth (X)	Covered > 1X (%)
HGDP00521	French	(Meyer et al. 2012)	22.6	90
HGDP00542	Papuan	(Meyer et al. 2012)	21.6	90
HGDP00778	Han	(Meyer et al. 2012)	22.3	90
HGDP00927	Yoruba	(Meyer et al. 2012)	26.7	90
HGDP00998	Karitiana	(Meyer et al. 2012)	21.3	90

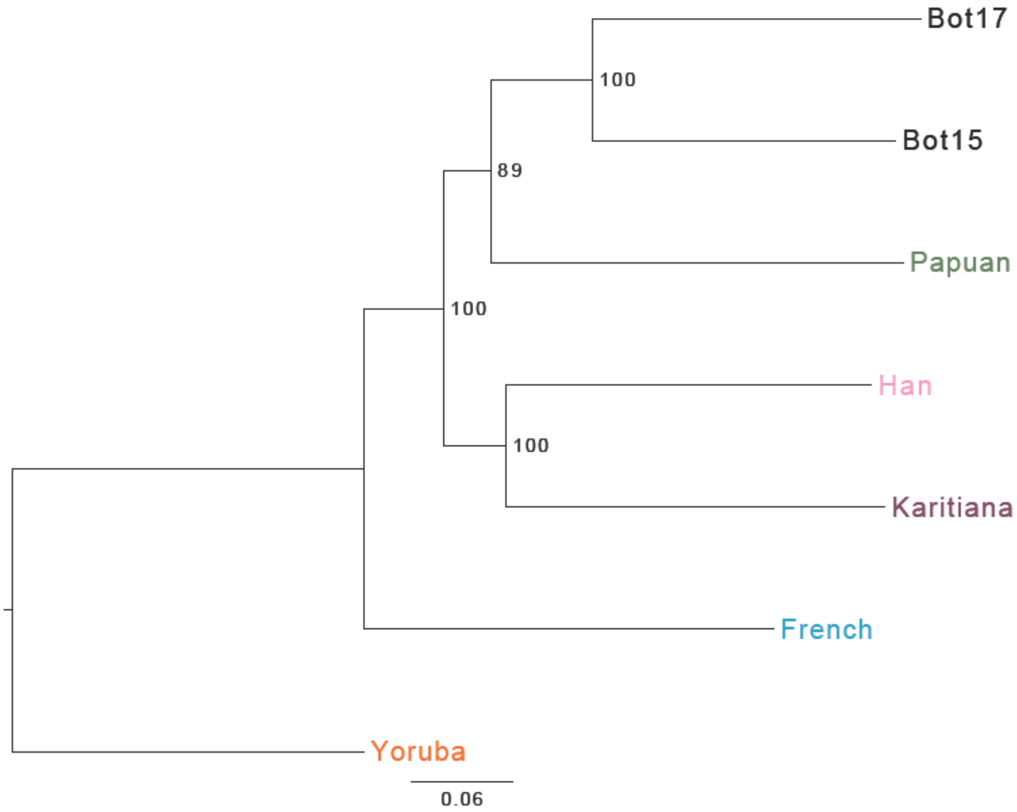
Supplementary Table 5.1 Assembly details of the five human genomes used for comparison.

	#number of SNPs	#number of SNPs (excl. transitions)	#populations	#individuals	#Polynesian individuals	#Native American individuals
Wollstein_Xing	823,805	256,261	20	583	45	48
Reich_Wollstein	108,662	1,376	130	2,436	19	493

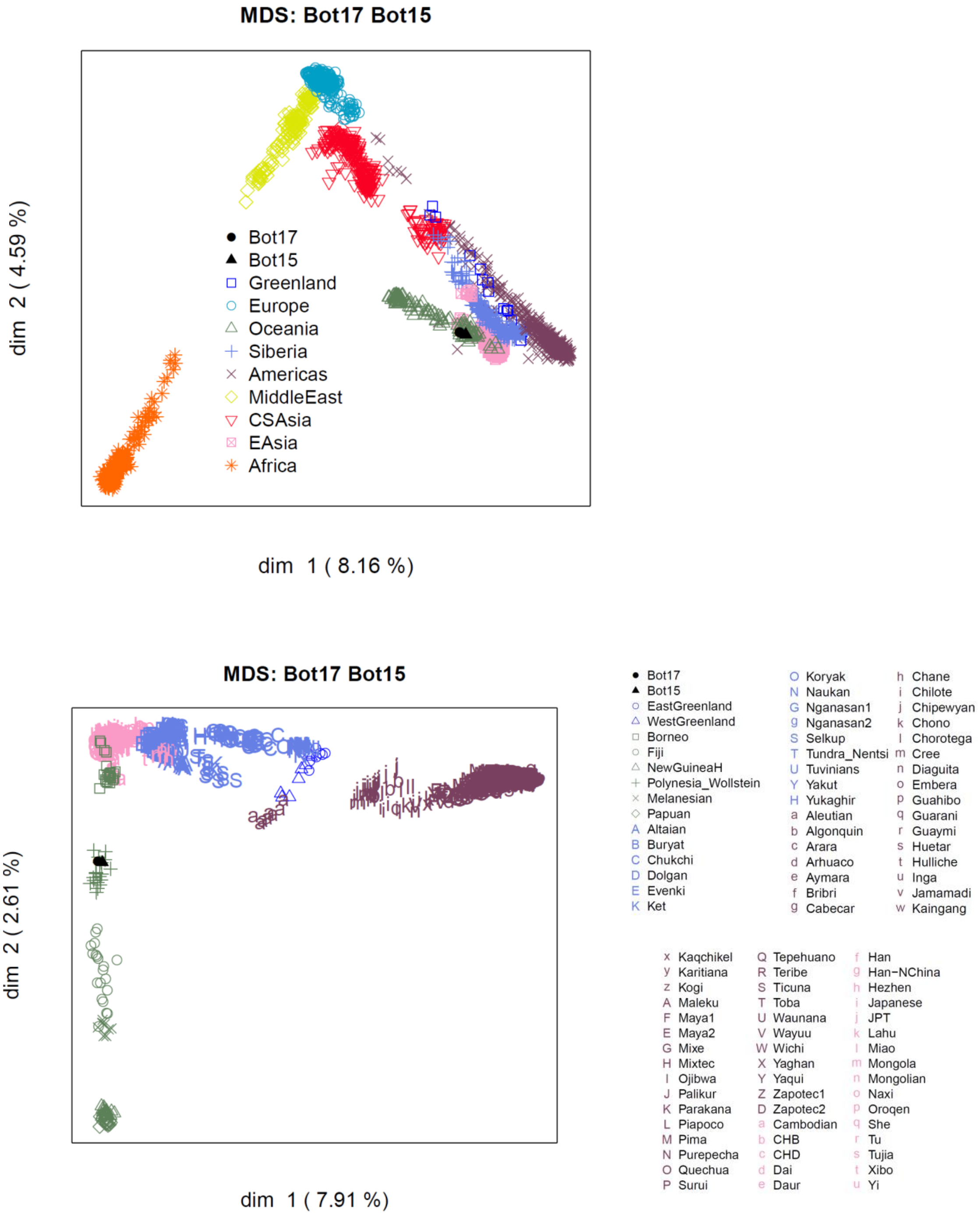
Supplementary Table 5.2 Summary of the genotype data used for analysis.

		H3 = Bot17				H3 =Bot15			
H1	H2	nABBA-nBABA	nABBA+nBABA	D	Z	nABBA-nBABA	nABBA+nBABA	D	Z
French	**Papuan***	**10066**	**120822**	**0.083**	**12.1**	**10937**	**127043**	**0.086**	**12.4**
French	**Han***	**15769**	**118719**	**0.133**	**21.2**	**17216**	**124768**	**0.138**	**22.7**
Papuan	**Han***	**5661**	**119379**	**0.047**	**6.4**	**6264**	**125452**	**0.05**	**6.7**
French*	**Yoruban**	**-38650**	**132502**	**-0.292**	**-55.3**	**-39880**	**139192**	**-0.287**	**-55.3**
Papuan*	**Yoruban**	**-48739**	**140215**	**-0.348**	**-58.9**	**-50817**	**147363**	**-0.345**	**-57.6**
Han*	**Yoruban**	**-54462**	**140074**	**-0.389**	**-77.5**	**-57147**	**147489**	**-0.387**	**-71.7**
French	**Karitiana***	**11394**	**115564**	**0.099**	**14.9**	**11198**	**120944**	**0.093**	**15.3**
Papuan	Karitiana	1258	118464	0.011	1.4	294	124804	0.002	0.3
Han*	**Karitiana**	**-4505**	**110325**	**-0.041**	**-6.5**	**-6025**	**115665**	**-0.052**	**-8.4**

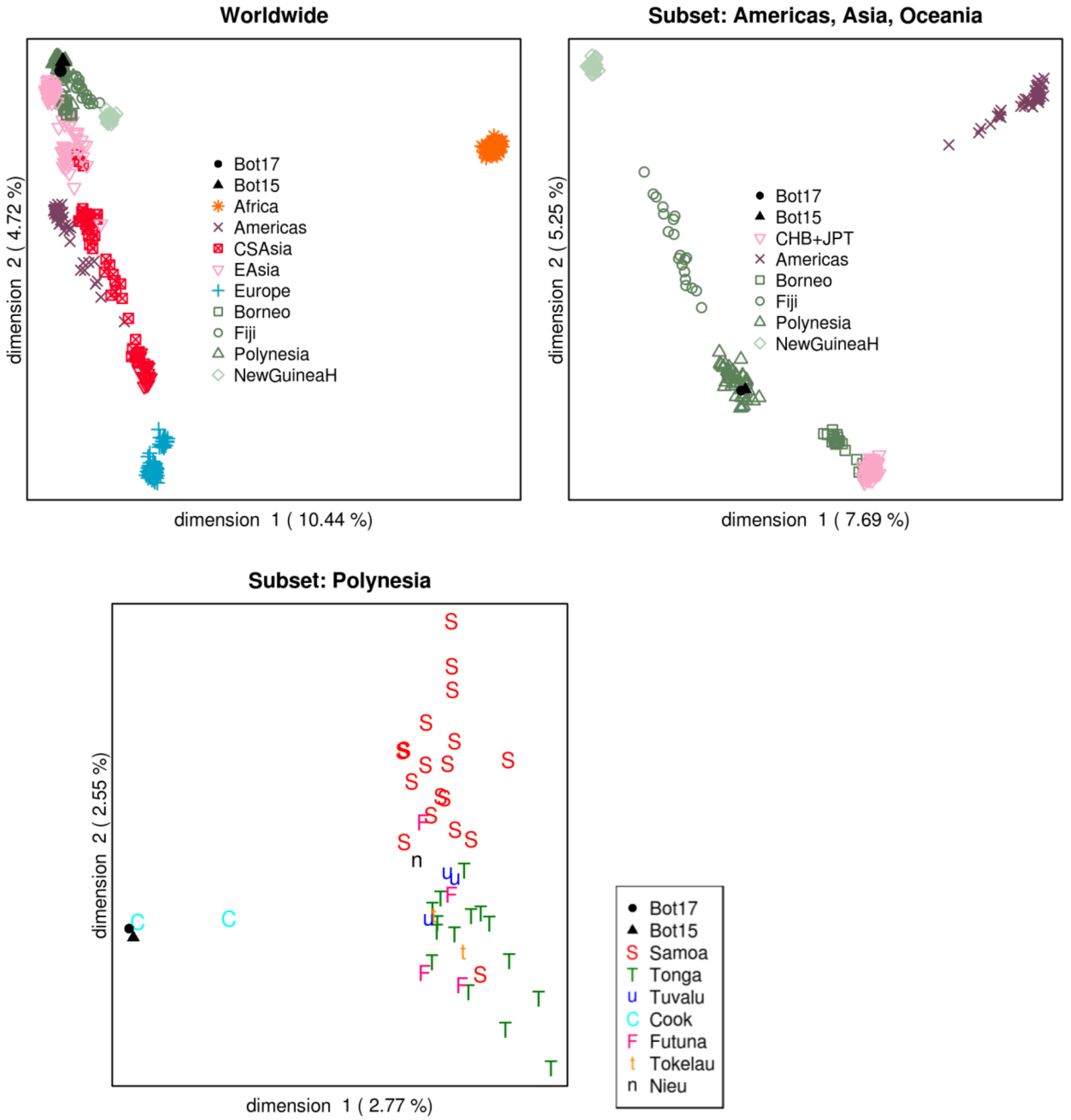
Supplementary Table 5.3 Results for the ABBA-BABA tests. The table shows ABBA-BABA *D*-statistics and Z-values based on jackknife estimation. The tests in bold are the tests that show significant deviation from the null hypothesis. * shows the population that the Bot15 and Bot17 samples is closest to.



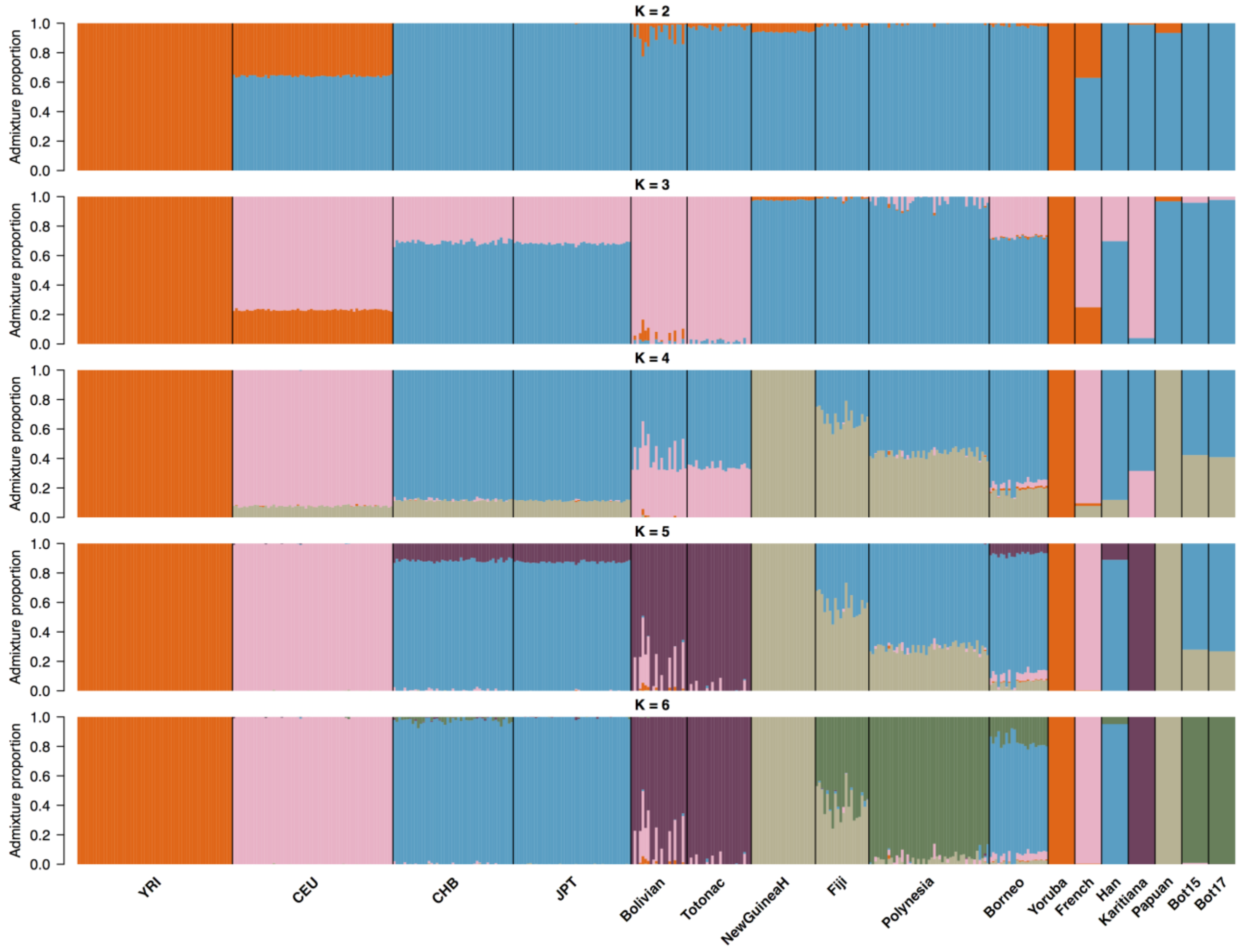
Supplementary Figure 5.1 Phylogenetic tree of the whole genome data for Bot15, Bot17 and the five modern genomes described in Supplementary Table 5.1 raxmlHPC-PTHREADS was used to build the tree with 10,000 bootstrap replicates.



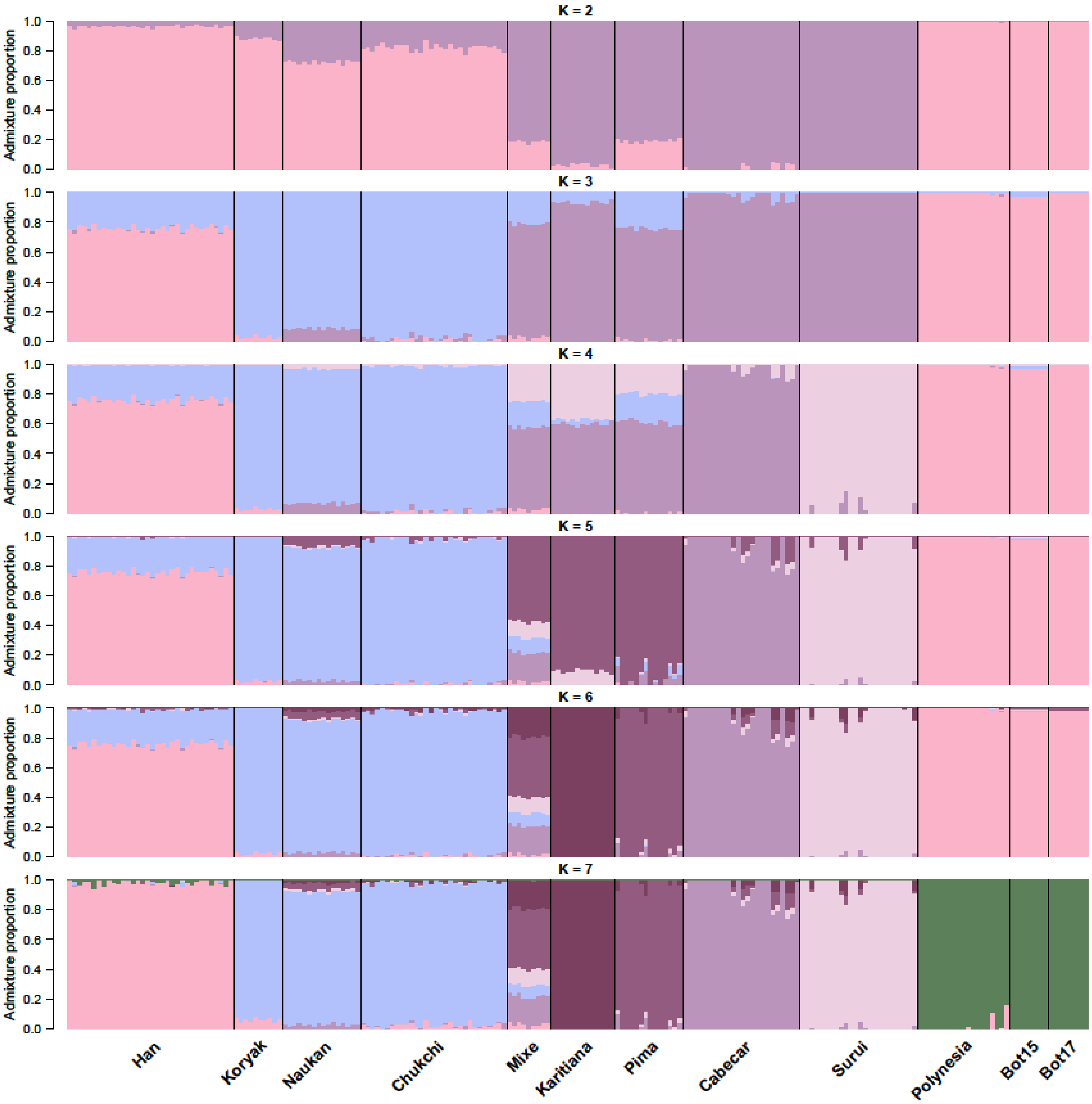
Supplementary Figure 5.2 MDS results for the Reich_Wollstein dataset. The first dimensions of the MDS analysis for all the populations in Reich_Wollstein (top) and a subset of the populations including all Native American, Oceanian, Siberian and Greenlandic populations (bottom). The Botocudos are shown in filled black triangle and circle and clusters with the Polynesians in both plots.



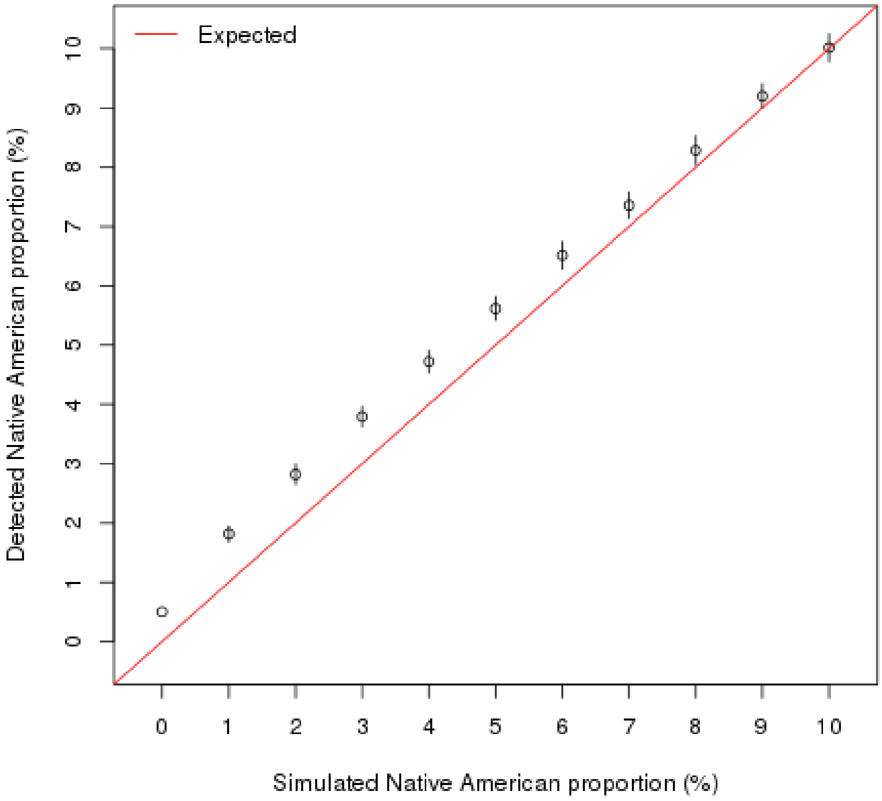
Supplementary Figure 5.3 MDS results for the Wollstein_Xing dataset. Two first dimensions of the MDS analysis for all the populations (labeled by geographic regions) in Wollstein_Xing excluding transitions (top left), a subset of the populations including all Native American, Oceanian, and two East Asian populations (top right), only the Polynesian labeled by Island (bottom). The Botocudos are shown in filled black triangle and circle and clusters with the Polynesians in both top plots.



Supplementary Figure 5.4 Estimated admixture proportions for a subset of Wollstein_Xing. The SNP chip data includes African (YRI), Asian (CHB, JPT, Borneo), Oceanian (New Guinea Highlands, Fiji, Polynesia, Borneo) and Native American (Bolivian, Totonac) samples. Note the bars for the seven sequenced individuals (Yoruba, French, Han, Karitiana, Papuan, Bot15, Bot17) to the right have been made wider to make them more visible.



Supplementary Figure 5.5 Estimated admixture proportions for Reich_Wollstein. The SNP chip data includes Asian (Han), Siberian (Koryak, Naukan, Chukchi) and Native American (Mixe, Karitiana, Pima, Cabecar, Surui) samples. Note the bars for Bot15 and Bot17 to the right have been made wider to make them more visible.



Supplementary Figure 5.6 Simulated and measured admixture proportions in an artificially admixed Botocudo individual. We simulated different levels of Native American admixture by artificially substituting parts of the Bot17 with a subsampled version of the Karitiana (Brazil) genome. The simulated (X-axis) and the measured admixture proportions in the simulated individual as estimated by NGSadmix (Y-axis) are reported. Standard deviations for each estimate were obtained through 100 replicates.

Supplementary 6
Analysis of the Bot15 and Bot17 mtDNA data

Philip L. F. Johnson*, Ana T. Duggan, Maanasa Raghavan, Simon Rasmussen, Anna-Sapfo Malaspinas *

to whom correspondence should be addressed (plfjohnson@emory.edu, annasapfo@gmail.com)

Contamination estimates

We estimated the contamination fraction for the mtDNA for Bot15 (mtDNA capture experiment and shotgun experiment) and Bot17 (shotgun experiment). As a control, we also estimated the contamination fraction for the bait (a mitogenome from a contemporary individual with African ancestry) that was used to capture the mtDNA for Bot15 (see S5).

As detailed in S6, we mapped the reads from each experiment to the whole nuclear genome (genome build 37.1) as well as to the consensus mtDNA from that experiment. For our contamination analysis, we only retained those reads that mapped best to the consensus mtDNA, which has the effect of eliminating most nuclear copies of mitochondrial genes. The total number of reads and the average depth at each position covered by at least one read is given in Supplementary Table 6.1.

To estimate contamination, we used a method detailed in the Supplemental Information section 5 of Fu et al. (2013) that generates a moment-based estimate of the error rate and a Bayesian-based estimate of the posterior probability of the contamination fraction. Both estimates are given in Supplementary Table 6.1. As expected, the bait has a contamination fraction close to 0 (95% credible interval of 0.0-0.2%). The individual Bot15 is more contaminated (1.5-3.7%) than Bot17. For Bot17, the credible interval includes 0 (0.0-0.8%). These estimates are on the same order of magnitude as contamination fractions for other ancient genomes (see, *e.g.*, Fu et al. (2013)).

Haplogroup determination

We determined the haplogroup for Bot15 and Bot17 by using mthap v15.0 by uploading the consensus sequences on James Lick's website: http://dna.jameslick.com/mthap/. As reported in the main text, Bot15 and Bot17 belong to the haplogroups B4a1a1a and B4a1a1, respectively, confirming an earlier result that was not based on the whole genome (Gonçalves et al. 2013), while the bait belongs to haplogroup L2a1a1.

Phylogenetic analysis

In order to place our consensus sequences on a phylogenetic tree, we compiled a reference dataset of relevant mtDNA whole genomes. We used 164 sequences from Oceania (Soares et al. 2011), 85 sequences from the Americas (Fagundes et al. 2008), 3 Malagasy sequences (Razafindrazaka et al. 2010), 16 African mtDNA genomes publicly available (Genbank accession numbers: AF346968, AF346969, AF346976, AF346977, AF346985, AF346986, AF346987, AF346992, AF346995, AF347008, AF347009, AY195777, AY195783, AY195788, AY195789, AY963585), and, a gorilla, a chimp and a bonobo sequence as outgroups (D38114, D38113, D38116).

We aligned the reference set with the Bot15 and Bot17 consensus sequences using MAFFT v7.023b (Katoh and Standley 2013). We then built a phylogenetic tree using MrBayes v3.2.1 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) setting the number of substitution types to 6 (nst = 6, GTR model) and a gamma distribution for the rates across sites (rates = gamma). We ran the Markov chain Monte Carlo (mcmc) for 20 million replicates (ngen = 20000000) with default parameters.

The Potential Scale Reduction Factor (PSRF) was between 1.000 and 1.001 suggesting the mcmc has converged. The average split frequency was around 0.02, i.e. below 0.05, which is adequate in our case since we are interested only in well-supported nodes.

The resulting consensus tree, plotted with Mesquite v.2.75 (Maddison, W.P. and Maddison, D.R.), is shown in Supplementary Figure 6.1. Given the mtDNA is only one locus, or one gene tree, it is hard to draw conclusions on the history of the populations from this result. Nevertheless, we find that the Bot15 and Bot17 individuals cluster within a well-supported clade with a posterior probability of > 0.8 that includes no Native American sequences.

The haplogroups tend to cluster on the tree, but for several cases they do not form a monophyletic group, which can be expected given the small number of variable sites that defines them.

Comparison with modern Oceanian populations

To assess the relationship with present day Oceanian diversity, we compared Bot15 and Bot17 to 1,001 samples belonging to the B4 mtDNA lineage from 33 populations across Near and Remote Oceania (Duggan and Stoneking 2013; Duggan, Ana T. et al. 2014). After removing the two poly-C regions (pos 303-315 and 16182-16193) pairwise alignments between each of Bot15 and Bot17 and the 1001 samples were conducted with MUSCLE v3.8.31 (Edgar 2004) and mismatches were calculated with an in-house Perl script.

We find that Bot15 shares a haplotype with a cluster of 40 individuals from Polynesian and Polynesian Outlier populations. Polynesian Outliers are populations located in the Solomon Islands and believed to be descended from individuals who back-migrated from Polynesia to Near Oceania (Kirch 1984; Green 1995). We find no exact matches to Bot17 within the comparison data set, but do find several samples which differ by only one or two positions.

Notably, Bot17 falls within one mutational step of samples belonging to both haplogroups B4a1a1 and B4a1a1a and samples originating from Polynesian, Solomon Islands and Fijian populations. Overall, these findings are consistent with the conclusion that Bot15 and Bot17 are of Polynesian origin. Tables summarizing the number of mismatches between Bot15, Bot17 and the 1001 samples mentioned above can be obtained upon request.

To visualize the relationships of 994 samples belonging to mtDNA lineage B4a (excluding seven samples from the B4b lineage) and Bot15 and Bot17, we constructed a network, first with a reduced median algorithm (Bandelt et al. 1995) and then with a median joining algorithm (Bandelt, Forster, and Röhl 1999) and with transversions weighted three times greater than transitions, with the program Network and post-processed with Network Publisher (http:fluxus-engineering.com, Supplementary Figure 6.2). Network analysis shows that both Bot15 and Bot17 are within the diversity of modern Oceanian populations. The position of Bot17 in the network, as a point of reticulation between haplogroups B4a1a1 and B4a1a1a, suggests two possibilities: firstly, that Bot17 shares a T16126C mutation with four B4a1a1a samples (from the Polynesian populations of Cook Islands and Niue) but has subsequently undergone a back mutation to the ancestral A allele at position 16247. The 16247G allele defines haplogroup B4a1a1a, and the instability of the derived allele at this position and its tendency to undergo multiple independent back-mutations has been described previously (Duggan and Stoneking 2013). Hence, we favor this explanation. The less-likely possibility is that there has been a parallel mutation of T16126C on both the B4a1a1 and B4a1a1a lineages and that Bot17 is more closely related to the B4a1a1 samples from the Solomon Islands, Fiji and Tonga.

References

Bandelt, H J, P Forster, B C Sykes, and M B Richards. 1995. “Mitochondrial Portraits of Human Populations Using Median Networks.” *Genetics* 141 (2): 743–53.

Bandelt, H. J., P. Forster, and A. Röhl. 1999. “Median-Joining Networks for Inferring Intraspecific Phylogenies.” *Molecular Biology and Evolution* 16 (1): 37–48.

Duggan, Ana T., and Mark Stoneking. 2013. “A Highly Unstable Recent Mutation in Human mtDNA.” *The American Journal of Human Genetics* 92 (2): 279–84. doi:10.1016/j.ajhg.2012.12.004.

Duggan, Ana T., Bethwyn Evans, Françoise R. Friedlaender, Jonathan S. Friedlaender, George Koki, D. Andrew Merriwether, Manfred Kayser, and Mark Stoneking. 2014. “Maternal History of Oceania from Complete mtDNA Genome Sequences: Contrasting Ancient Diversity with Recent Homogenization due to the Austronesian Expansion.”

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97. doi:10.1093/nar/gkh340.

Fagundes, Nelson J.R., Ricardo Kanitz, Roberta Eckert, Ana C.S. Valls, Mauricio R. Bogo, Francisco M. Salzano, David Glenn Smith, et al. 2008. “Mitochondrial Population Genomics Supports a Single Pre-Clovis Origin with a Coastal Route for the Peopling of the Americas.” *American Journal of Human Genetics* 82 (3): 583–92. doi:10.1016/j.ajhg.2007.11.013.

Fu, Qiaomei, Alissa Mittnik, Philip L. F. Johnson, Kirsten Bos, Martina Lari, Ruth Bollongino, Chengkai Sun, et al. 2013. “A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes.” *Current Biology* 23 (7): 553–59. doi:10.1016/j.cub.2013.02.044.

Gonçalves, V. F., J. Stenderup, C. Rodrigues-Carvalho, Hilton P. Silva, H. Gonçalves-Dornelas, A. Líryo, T. Kivisild, et al. 2013. “Identification of Polynesian mtDNA Haplogroups in Remains of Botocudo Amerindians from Brazil.” *Proceedings of the National Academy of Sciences*. doi:10.1073/pnas.1217905110. http://www.pnas.org/content/early/2013/03/28/1217905110.

Green, Roger C. 1995. “Linguistic, Biological and Cultural Origins of the Initial Inhabitants of Remote Oceania.” 17: 5–27.

Huelsenbeck, John P., and Fredrik Ronquist. 2001. “MRBAYES: Bayesian Inference of Phylogenetic Trees.” *Bioinformatics* 17 (8): 754–55. doi:10.1093/bioinformatics/17.8.754.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80. doi:10.1093/molbev/mst010.

Kirch, P. V. 1984. “The Polynesian Outliers: Continuity, Change, and Replacement ∗.” *The Journal of Pacific History* 19 (4): 224–38. doi:10.1080/00223348408572496.

Maddison, W.P., and Maddison, D.R. *Mesquite: A Modular System for Evolutionary Analysis.* (version 2.75). http://mesquiteproject.org.

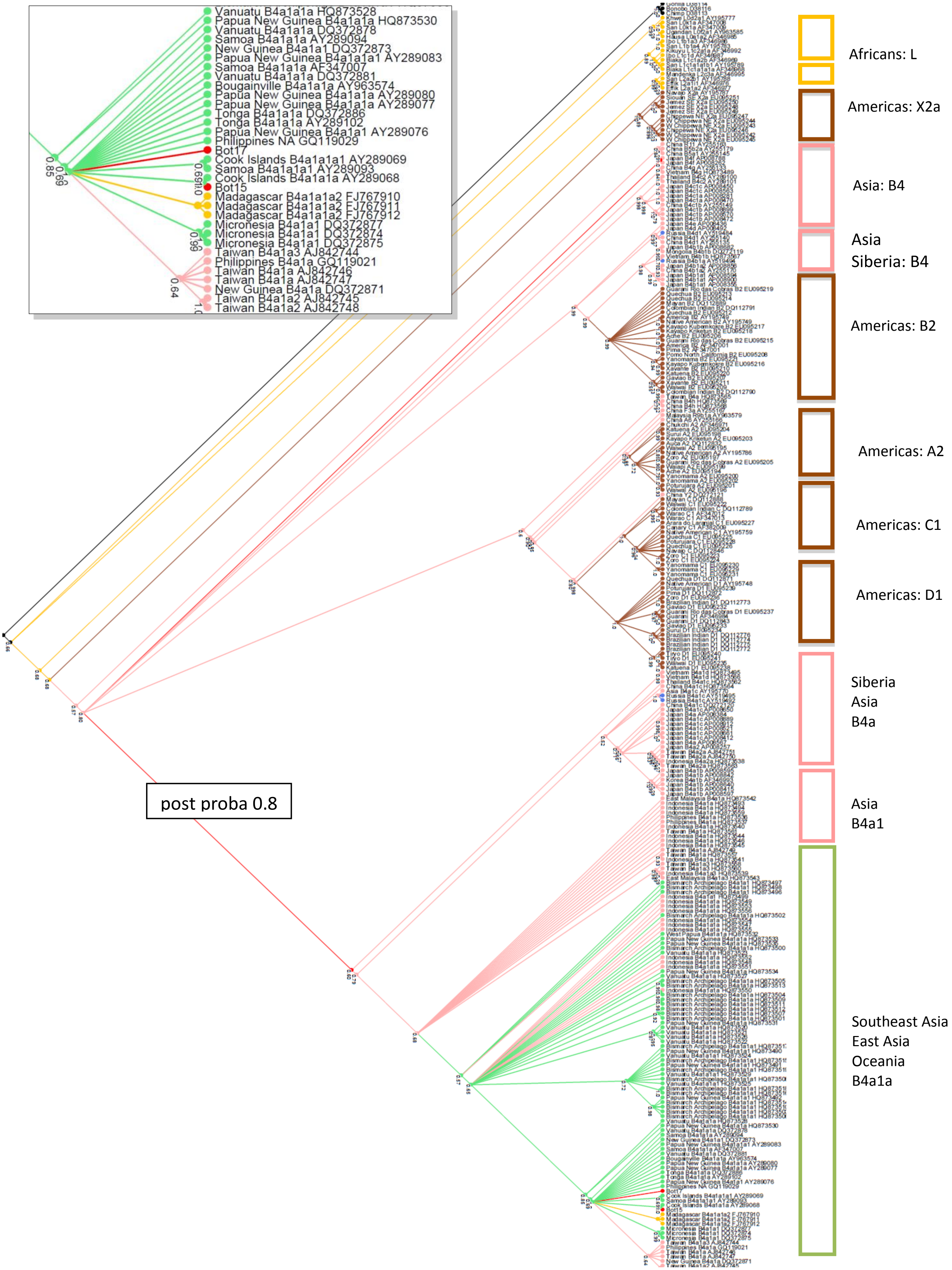
Razafindrazaka, Harilanto, Francois-X Ricaut, Murray P Cox, Maru Mormina, Jean-Michel Dugoujon, Louis P Randriamarolaza, Evelyne Guitard, Laure Tonasso, Bertrand Ludes, and Eric Crubezy. 2010. “Complete Mitochondrial DNA Sequences Provide New Insights into the Polynesian Motif and the Peopling of Madagascar.” *European Journal of Human Genetics* 18 (5): 575–81. doi:10.1038/ejhg.2009.222.

Ronquist, Fredrik, and John P. Huelsenbeck. 2003. “MrBayes 3: Bayesian Phylogenetic Inference under Mixed Models.” *Bioinformatics* 19 (12): 1572–74. doi:10.1093/bioinformatics/btg180.

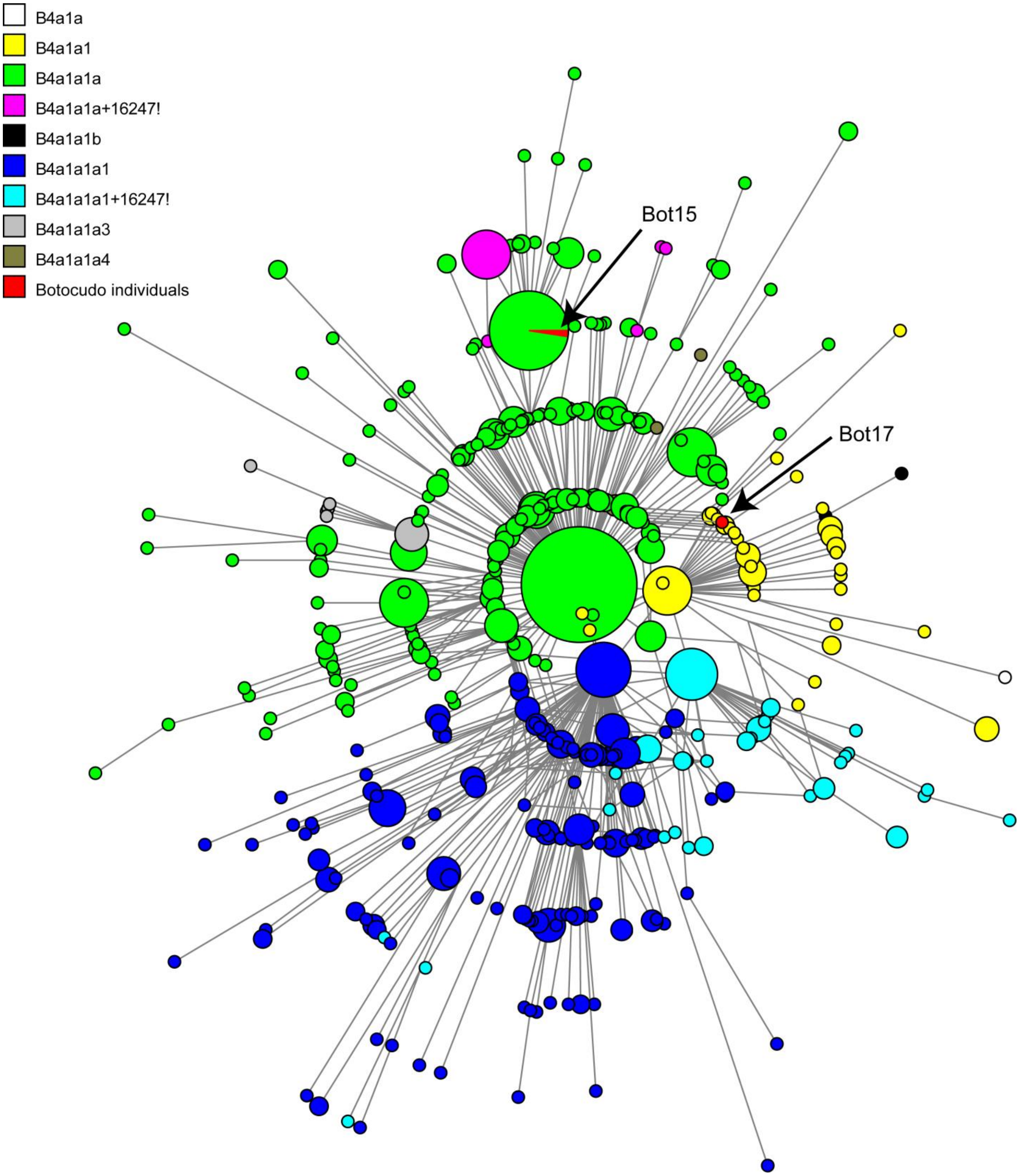
Soares, Pedro, Teresa Rito, Jean Trejaut, Maru Mormina, Catherine Hill, Emma Tinkler-Hundal, Michelle Braid, et al. 2011. “Ancient Voyaging and Polynesian Origins.” *American Journal of Human Genetics* 88 (2): 239–47. doi:10.1016/j.ajhg.2011.01.009.

	#reads	read length	covered >1 read	average depth	error rate %	contamination %
Bot15 mtDNA capture	24,397	91	15,478	144.1	1.1	2.1-3.7
Bot15 shotgun	14,348	77	14,686	75.1	0.8	1.5-3.6
Bot17 shotgun	14,913	61	14,421	63.3	0.8	0.0-0.8
Bait mtDNA capture	17,804	51	12,683	71.2	0.2	0.0-0.2

Supplementary Table 6.1 We report here the number of reads that map to the respective consensus, the average length of those reads, the number of positions covered by at least one read (note that the length of the consensus sequences is around 16,560 bp), the average depth for each position covered by at least one read, the average error rate, and the 95% credible interval (2.5 and 97.5% quantiles) of the posterior probability of contamination for each relevant experiment. We define “read length” as the length of the reads after adaptor trimming and mapping to the human genome. All the sequencing was done with an index read of 6bp on an Illumina HiSeq with 98-100 cycles, except for the bait that was sequenced on an Illumina MiSeq with 58 cycles.



Supplementary Figure 6.1 MrBayes consensus tree for 271 human worldwide sequences and a gorilla, a chimp and a bonobo used as outgroups (on top). We colored in green Oceanian, in pink Asian, East and Southeast Asian, in brown Native American, in light blue Russian (Siberia) and in orange African populations. The side panel indicates some of the common haplogroups that tend to cluster together, the color corresponding to the continent represented at higher frequency for those. In red are Bot15 and Bot17 (on lower part of the main tree and on the enlarged subtree).



Supplementary Figure 6.2 Network for 996 samples belonging to mtDNA lineage B4a with positions of Bot15 and Bot17 indicated. Nodes are scaled proportional to frequency and colored according to assigned haplogroup. Botocudo samples are indicated in red.

Supplementary 7

Analysis of the Bot15 SNP array capture experiment data

Oscar Lao*, Morten Rasmussen, Anna-Sapfo Malaspinas*

*to whom correspondence should be addressed (o.laogrueso@erasmusmc.nl, annasapfo@gmail.com)

As detailed in the main text and in S2, the SNP capture experiment was only performed for Bot15.

Selection of ancestry informative markers

Reference datasets

We selected around 6,000 Ancestry Informative Markers (AIMs) to detect European, East Asian, Polynesian and Native American genetic ancestry using allele frequencies in each of these populations. The SNP dataset for estimating the allele frequencies comprised: HapMap II ((Frazer et al. 2007), worldwide populations (Xing et al. 2010), Polynesian individuals without European admixture (labeled POL, Wollstein et al. (2010)) and HGDP-CEPH populations (López Herráez et al. 2009). Some of these datasets shared some populations, such as CHB or YRI. All datasets were genotyped on the Affymetrix platform using different microarrays. The merged dataset comprised 214,819 single nucleotide polymorphisms (SNPs). From those, we excluded the SNPs that either involved A-T and C-G transversions or had an Informativeness of Ancestry (*In*; as defined in Rosenberg et al. (2003)) value >0.05 between shared populations in different datasets after allelic and strand SNP matching. In the case of the POL dataset we observed that, after allelic and strand SNP matching, there were few SNPs highly differentiated (that is, *In* ≈log(2), corresponding to the maximum *In* value for comparison of two populations) to other populations. Because we were particularly interested in genetically distinguishing POL from other populations, these SNPs were clear AIMs candidates. Alternatively, given such extreme degree of genetic differentiation, these markers could just indicate genotyping errors in the POL dataset. However, if the genetic differentiation between POL and other populations were real, we would expect to see the same *In* signal in surrounding markers. Therefore, we ascertained from the (Wollstein et al. 2010) dataset the neighbor (<80 kb) markers to the ones showing extreme patterns of POL population differentiation; we computed for each SNP the *In* statistic between POL and CHB. We then compared the amount of information for differentiating these two populations between the outlier SNP and the SNPs from the surrounding genomic region. The last step was performed by means of computing the conditional amount of information for inferring population ancestry of one marker given the information already provided by another one (*In(A|B)*) following the basic definition of conditional Information (Cover and Thomas 1991):

$$In(A|B) = In(A;B)-In(B)$$

Where *In(A|B)* is the Informativeness of ancestry of SNP A given the information of SNP B, *In(A;B)* is the Informativeness of ancestry of SNPs *A* and *B* considered together (taking all the observed allelic combinations between SNP *A* and *B*) and *In(B)* is the Informativeness of ancestry of SNP *B* (considering only the allelic combinations of the SNP *B*). A marker *A* with a conditional *In* value given *B* close to 0 implies that the amount of information for distinguishing POL from CHB provided by the marker *A* is already explained by the marker *B*, and therefore it is not different from what is observed in the overall genomic region.

The final cleaned dataset comprised 179,943 SNPs.

AIMs ascertainment algorithm

After data cleaning, the samples were divided in two datasets. *Dataset 1* was used for AIMs ascertainment and comprised populations from each of the ancestries of interest: CEU for European ancestry (from HapMap II), CHB for East Asian ancestry (from HapMap II), POL for Polynesian ancestry (Wollstein et al. 2010) and Bolivian for Amerindian (Xing et al. 2010).

Dataset 2 comprised all the remaining populations (except African Sub-Saharan and Central Asian populations) and was used to validate the ascertained markers.

The *In* statistic was computed for each SNP using the allelic frequencies observed in the populations of Dataset 1, and sorted from the largest *In* to the smallest one. In order to minimize the effects of LD on the final AIMs dataset, further SNP exclusion criteria were applied. First, the SNP with the largest *In* was included in the list of candidate AIMs. The following SNPs from the list were iteratively included in the final SNP dataset if they were at a distance >100kb from the ones that had been already included or were at a distance<100kb but had a conditional *In* greater than 0.2, which is likely to suggest independent ancestry information from what had been already included. After that step, 33,443 candidate SNPs remained.

A final list of 6,000 AIMs was obtained by iteratively ascertaining without replacement sets of 10 markers (comprising a total of 600 iterations) as follows:

In each iteration, a greedy algorithm (Cormen 2001) to ascertain the 10 markers maximizing the combined *In* statistic was applied. The combined *In* statistic was computed as described in (Rosenberg et al. 2003), calculating for each population the frequencies of all the possible allelic combinations (*i.e.*, for 10 SNPs, there are 1024 allelic combinations). After obtaining the best set of 10 markers, these are removed from the list of candidate AIMs and included in the final list of AIMs. The combined *In* of each iteration was monitored, showing that the combined *In* of each set of 10 markers decreases with the number of iterations (see Supplementary Figure 7.1).

The ascertained markers were validated on *Dataset 2* by means of FRAPPE (Tang et al. 2005) and classical multidimensional scaling (MDS; (Cox and Cox 2000)) using an Identity By State (IBS) distance matrix between pairs of individuals (the number of alleles shared at each maker) implemented in the function cmdscale of the stats package of R software (R Development Core Team). On Supplementary Figure 7.2 and Supplementary Figure 7.3 we plot the estimated ancestry from MDS (for the first 3 dimensions) and FRAPPE (using K=4) on the samples pooled by continent. As can be seen in this figure, the proposed 6,000 AIMs are highly informative for detecting European, East Asian, Polynesian and Native American ancestry.

SNP selection for SNP capture array

For the array design, we applied a further filtering step to keep only markers in regions that have limited homology to elsewhere in the genome to maximize the efficiency of the capture by avoiding unspecific binding to homologue regions. Marker probes were tiled at 2 bp, for 100 bp upstream and downstream of the target position for the 6,000 makers. Probes were filtered according to Hodges et al. (2009). Briefly, markers not identified as 'single' in dbSNP were removed. Probes were compared against 15-mers found more than 100 times in the genome; if 25% (15bp) or more of the probe matched these sections it was removed. Lastly probes that had more than 85% homology with other regions of the genome in a BLAT search were removed. We were left with 5,744 markers.

Read distribution and Genotype calling

After array capture experiment and sequencing (S2), the Illumina reads were mapped as described in S3.

We further discarded bases with a base quality below 20. A total of 5,589 targeted sites were covered by at least one read (see Supplementary Table 7.1 for the distribution of markers per chromosome), thus covering 97.3% of the targeted sites by at least one read. In comparison, while we produced around 8.8 times more reads (after trimming, see S3) for the shotgun experiment for Bot15, 4,414 sites were covered by at least one read in this case, i.e., 78.9% of the targeted sites. Although the number of sites covered is similarly high, the goal of the capture is to increase the depth at targeted sites such that, in our case, we can confidently call genotypes. We therefore consider the depth as well; on average, each target is covered by 40.4 reads for the SNP capture (see Supplementary Figure 7.4 for depth distribution at each target). In comparison, the average number of reads covering the targets for the shotgun data is 1.9. We normalized the average depth at the targets by the number of trimmed reads for each experiment (around 1.3 billion for the shotgun experiments versus 150 million for the SNP capture) and found that the enrichment is ~187.0 fold. The average number of reads as well as the number of probes covering each position around the target for the SNP Capture is shown on Supplementary Figure 7.5.

We then called the genotypes for each SNP using the software *angsd* (Korneliussen 2013). We used the model that is a reimplementation of samtools 1.18 (H. Li et al. 2009; Y. Li et al. 2010), **-GL 1** in *angsd*, to compute the likelihood for each genotype. Note that this model does not take into account the increase in C->T and G->A observed for ancient DNA (e.g. (Briggs et al. 2007; Sawyer et al. 2012)), but it seems to be less sensitive to ancient DNA damage than the implementation in *angsd* of the GATK model ((DePristo et al. 2011), -**GL 2).** We then computed a posterior distribution by assuming an equal prior probability for each genotype. We finally called genotypes using a cutoff of 0.9 for the posterior probability (348 left out), keeping only SNPs with a depth of at least 5 (48 left out). We were left with a total 5,193 SNPs, *i.e.*, we filtered out a further 7%. As expected, from the 348 positions discarded by the cutoff on the posterior probability, genotypes affected by potential C->T or G->A are overrepresented (for 98% of those positions, the genotype with the highest posterior probability is either AA, AG, GG or CC, CT, TT).

Population genetic analysis

MDS

To investigate the genetic similarities of Bot15 with worldwide populations, we first performed a MDS using an IBS distance matrix considering all the individuals from *Dataset 1* and *Dataset 2* and Bot15 (comprising a total of 1,225 individuals). This analysis was performed twice. In the first analysis, we computed IBS distances between pairs of individuals using the genotype calls for all 5,193 AIMs. The second MDS analysis was performed with IBS distances computed with 947 AIMs after removing all markers with alleles C and T or G and A (to remove potentially damaged sites for Bot15). In both MDS analyses, a constant was added to the distance matrix to avoid negative eigenvalues (add parameter set to TRUE in the cmdscale function; see e.g. (Cox and Cox 2000) and citation therein). The resulting plots are shown on Supplementary Figure 7.6. We see that Bot15 falls within the Polynesians in both cases.

To compare the two MDS plots and see the effect of potentially damaged sites, we performed a symmetric Procrustes analysis (see e.g. (Wang et al. 2010)). The correlation between the first two dimensions of the MDS plots in a symmetric Procrustes was 0.995 (p-value = 0.001 based on 1000 permutations), indicating that the two plots are virtually indistinguishable.

Structure

We next inferred the proportions of genetic ancestry of Bot15 by means of the Bayesian approach implemented at the program *Structure* (Pritchard, Stephens, and Donnelly 2000; Falush, Stephens, and Pritchard 2003). *Structure* was run using unsupervised (that is, assuming no parental information) and supervised analyses. In all the cases, we considered six different geographical regions comprising 157 individuals:

Polynesian: POL, Tonga, Samoa

East Asian: Chinese, Japanese

Native American: Colombia, Karitian, Makrani, Maya, Pima, Surui, Totonac

Europe: Northern Europe

Melanesia: Melanesia (non-Austronesian population)

Papua New Guinea: Papua New Guinea (Papua NG, non-Austronesian population)

For the unsupervised analyses, we ran the MCMC for a total of 10,000 iterations discarding the first 5,000 as burnin, using default values for the rest of the parameters.

We ran 5 iterations for each number of assumed ancestries (K), ranging from 2 to 5. To get the consensus clustering for each K, we then performed a CLUMPP analysis (Jakobsson and Rosenberg 2007) using the greedy algorithm option. For K<5, the reproducibility of ancestry proportions among iterations estimated by means of H' (Jakobsson and Rosenberg 2007) is close to 1 (Supplementary Table 7.2). Moreover, the mean log-likelihood of the data is maximum at K=5 but the standard deviation of the 5 iterations is relatively high at K=5, which would suggest that the best K could be 4 or 3 (Evanno, Regnaut, and Goudet 2005).

The consensus clustering for each K is plotted on Supplementary Figure 7.7. For all Ks, the Bot15 individual tends to share the same ancestral components and ancestry proportions as POL, Tonga and Samoa populations. The most frequent ancestry component in Polynesians (>90% in each Polynesian samples) at K=4 ranges for Bot15 from ~90% to 98% depending on the run, while the second most frequent ancestry component is related to the Asian populations. The Native American component ranges between 0% and 0.2% for Bot15 with an average at 0.08%. Comparing these values to the other samples, we see that in the consensus clustering the Native American component in Bot15 is similar to those observed in all the non-Native American individuals. Indeed, while the average Native American component for Bot15 is 0.08%, the mean of the distribution of the average Native American component out of the 5 replicates in non-Native American individuals is 0.25% with a 95% CI of (0.04%, 1.32%).

For the supervised run, we split the populations into 6 groups: East Asian, Europeans, Native Americans, Polynesians, Melanesians, and Papuans guided by the results of (Wollstein et al. 2010). We performed another 5 runs as described above. We obtained a similar result to the unsupervised runs. For Bot15, the Polynesian component ranges between 92% and 95%. The second highest component is East Asians and ranges between 2% and 5% and the third one is Melanesian and Papuans (*i.e.*, non Austronesian populations) ranging between 0.5% and 3%. The Native American component ranges between 0.1% and 0.3%.

We replicated the analysis for the reduced set of markers that did not contain C->T or G->A. This lead to a similar result with an even higher Polynesian component (ranging from 97.5% to 98.7%).

Based on the MDS and the *Structure* analysis we conclude that Bot15 is most likely of Polynesian ancestry. Moreover, the fact that the percentage of Native American ancestry in the Bot15 is not different from the one observed in non-Native American samples suggests that this individual shows no evidence of Native American genetic admixture.

References

Briggs, Adrian W., Udo Stenzel, Philip L. F. Johnson, Richard E. Green, Janet Kelso, Kay Prüfer, Matthias Meyer, et al. 2007. "Patterns of Damage in Genomic DNA Sequences from a

Neandertal." *Proceedings of the National Academy of Sciences* 104 (37) (September 11): 14616–14621. doi:10.1073/pnas.0704665104.

Cormen, Thomas H. 2001. *Introduction To Algorithms*. MIT Press.

Cover, Thomas M., and Joy A. Thomas. 1991. *Elements of Information Theory*. 99th ed. Wiley-Interscience.

Cox, Trevor F., and M. A. A. Cox. 2000. *Multidimensional Scaling, Second Edition*. 2nd ed. Chapman and Hall/CRC.

Evanno, G., S. Regnaut, and J. Goudet. 2005. "Detecting the Number of Clusters of Individuals Using the Software Structure: a Simulation Study." *Molecular Ecology* 14 (8): 2611–2620. doi:10.1111/j.1365-294X.2005.02553.x.

Falush, Daniel, Matthew Stephens, and Jonathan K. Pritchard. 2003. "Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies." *Genetics* 164 (4) (August 1): 1567–1587.

Frazer, Kelly A., Dennis G. Ballinger, David R. Cox, David A. Hinds, Laura L. Stuve, Richard A. Gibbs, John W. Belmont, et al. 2007. "A Second Generation Human Haplotype Map of over 3.1 Million SNPs." *Nature* 449 (7164) (October 18): 851–861. doi:10.1038/nature06258.

Hodges, Emily, Michelle Rooks, Zhenyu Xuan, Arindam Bhattacharjee, D. Benjamin Gordon, Leonardo Brizuela, W. Richard McCombie, and Gregory J. Hannon. 2009. "Hybrid Selection of Discrete Genomic Intervals on Custom-designed Microarrays for Massively Parallel Sequencing." *Nature Protocols* 4 (6) (May): 960–974. doi:10.1038/nprot.2009.68.

Jakobsson, Mattias, and Noah A. Rosenberg. 2007. "CLUMPP: a Cluster Matching and Permutation Program for Dealing with Label Switching and Multimodality in Analysis of Population Structure." *Bioinformatics* 23 (14) (July 15): 1801–1806. doi:10.1093/bioinformatics/btm233.

Korneliussen, Thorfinn. 2013. *Angsd*. University of Copenhagen, Denmark. http://www.popgen.dk/angsd.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16) (August 15): 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, Yingrui, Nicolas Vinckenbosch, Geng Tian, Emilia Huerta-Sanchez, Tao Jiang, Hui Jiang, Anders Albrechtsen, et al. 2010. "Resequencing of 200 Human Exomes Identifies an Excess of Low-frequency Non-synonymous Coding Variants." *Nature Genetics* 42 (11) (November): 969–972. doi:10.1038/ng.680.

López Herráez, David, Marc Bauchet, Kun Tang, Christoph Theunert, Irina Pugach, Jing Li, Madhusudan R. Nandineni, Arnd Gross, Markus Scholz, and Mark Stoneking. 2009. "Genetic Variation and Recent Positive Selection in Worldwide Human Populations: Evidence from

Nearly 1 Million SNPs." *PLoS ONE* 4 (11) (November 18): e7888. doi:10.1371/journal.pone.0007888.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155 (2) (June 1): 945–959.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Rosenberg, Noah A., Lei M. Li, Ryk Ward, and Jonathan K. Pritchard. 2003. "Informativeness of Genetic Markers for Inference of Ancestry." *American Journal of Human Genetics* 73 (6) (December): 1402–1422.

Sawyer, Susanna, Johannes Krause, Katerina Guschanski, Vincent Savolainen, and Svante Pääbo. 2012. "Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA." *PLoS ONE* 7 (3) (March 30): e34131. doi:10.1371/journal.pone.0034131.

Tang, Hua, Jie Peng, Pei Wang, and Neil J. Risch. 2005. "Estimation of Individual Admixture: Analytical and Study Design Considerations." *Genetic Epidemiology* 28 (4): 289–301. doi:10.1002/gepi.20064.

Wang, Chaolong, Zachary A. Szpiech, James H. Degnan, Mattias Jakobsson, Trevor J. Pemberton, John A. Hardy, Andrew B. Singleton, and Noah A. Rosenberg. 2010. "Comparing Spatial Maps of Human Population-Genetic Variation Using Procrustes Analysis." *Statistical Applications in Genetics and Molecular Biology* 9 (1) (January 27). doi:10.2202/1544-6115.1493. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2861313/.

Wollstein, Andreas, Oscar Lao, Christian Becker, Silke Brauer, Ronald J. Trent, Peter Nürnberg, Mark Stoneking, and Manfred Kayser. 2010. "Demographic History of Oceania Inferred from Genome-wide Data." *Current Biology* 20 (22) (November): 1983–1992. doi:10.1016/j.cub.2010.10.040.

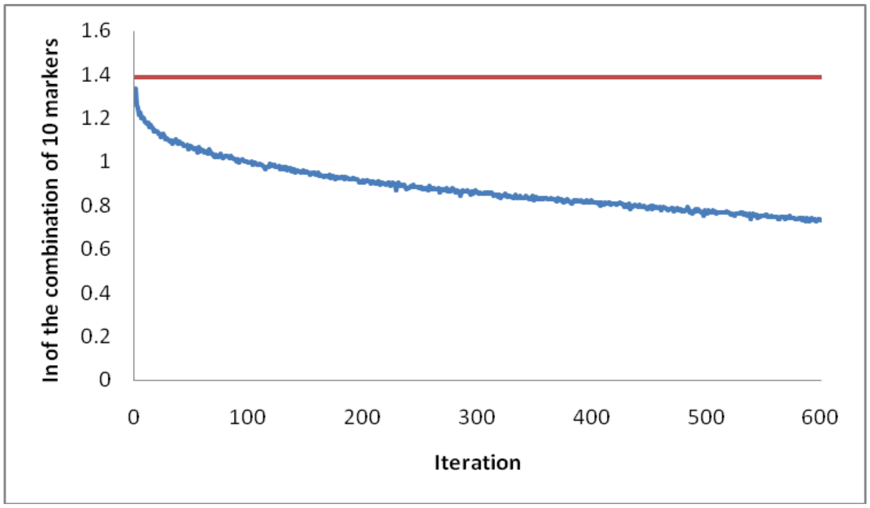
Xing, Jinchuan, W. Scott Watkins, Adam Shlien, Erin Walker, Chad D. Huff, David J. Witherspoon, Yuhua Zhang, et al. 2010. "Toward a More Uniform Sampling of Human Genetic Diversity: A Survey of Worldwide Populations by High-density Genotyping." *Genomics* 96 (4) (October): 199–210. doi:10.1016/j.ygeno.2010.07.004.

chrom	#SNPs
1	420
2	545
3	432
4	368
5	354
6	333
7	311
8	314
9	251
10	288
11	267
12	304
13	229
14	197
15	191
16	165
17	131
18	149
19	68
20	127
21	72
22	73
Total	5589

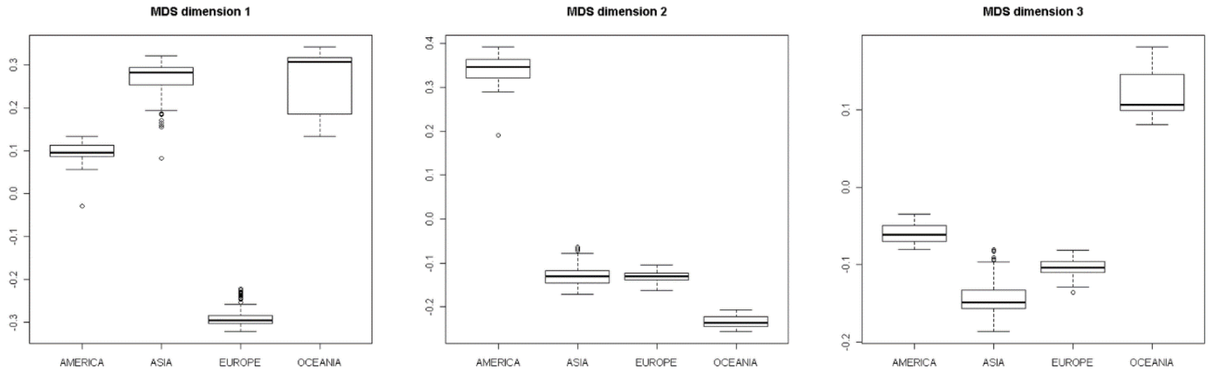
Supplementary Table 7.1 Number of SNPs per chromosome covered by at least one read.

K	mean (Lk)	sd(Lk)	H'
2	-853129	58.994	0.998
3	-793648	62.528	0.997
4	-781734	131.863	0.992
5	-780700	3141.445	0.908

Supplementary Table 7.2 Mean loglikelihood (Lk), standard deviation (sd), and reproducibility of ancestry proportions estimated by H' statistic among 5 replicates for each proposed number of ancestries (K)



Supplementary Figure 7.1 Amount of combined Informativeness of ancestry of the best 10 AIMs recovered each iteration. The red line depicts the maximum value that could be achieved considering four ancestral populations (*In* = ln(4); ref Rosenberg 2003). Notice the decay of *In* as the iteration increases, indicating that the best combinations of markers have been previously ascertained.



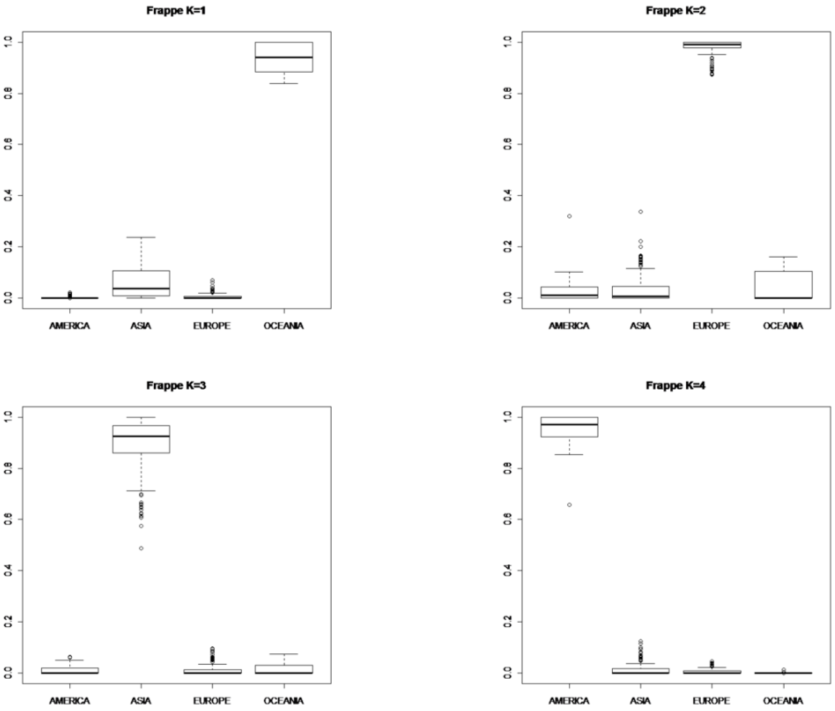
Supplementary Figure 7.2 Performance of the AIMs on the four ancestry groups. Boxplot of the ancestry estimates using the first 3 dimensions of a MDS analysis. Samples were pooled by continent of origin as follows:

America: Maya, Totonac, Colombians, Pima, Surui, Karitiana

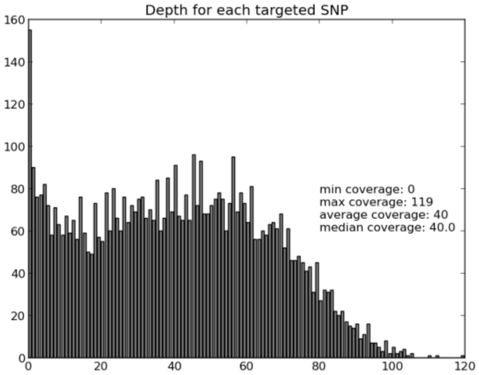
Asia: Thai, Yakut, Mongola, Cambodians, KhmerCambodian, Tu, Xibo, Oroqen, Daur, Hezhen, Han, Yizu, JPT, Vietnamese, Dai, Naxi, Japanese, Lahu, Chinese, Miaozu, Tujia, She

Europe: N.European, Slovenian, Tuscan, French, Orcadian, French_Basque, North_Italian, Sardinian, Druze, Stalskoe, Russian, Adygei, Mozabite

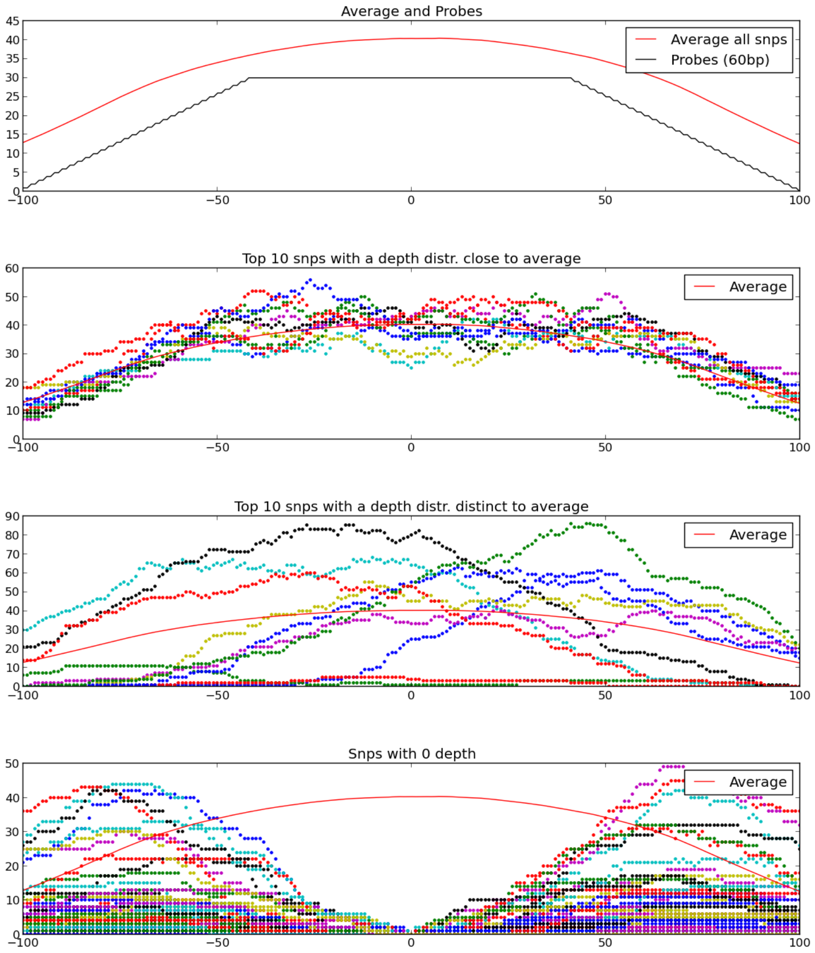
Oceania: Non-Austronesian Papuan and _Melanesian, Tongan, Samoan



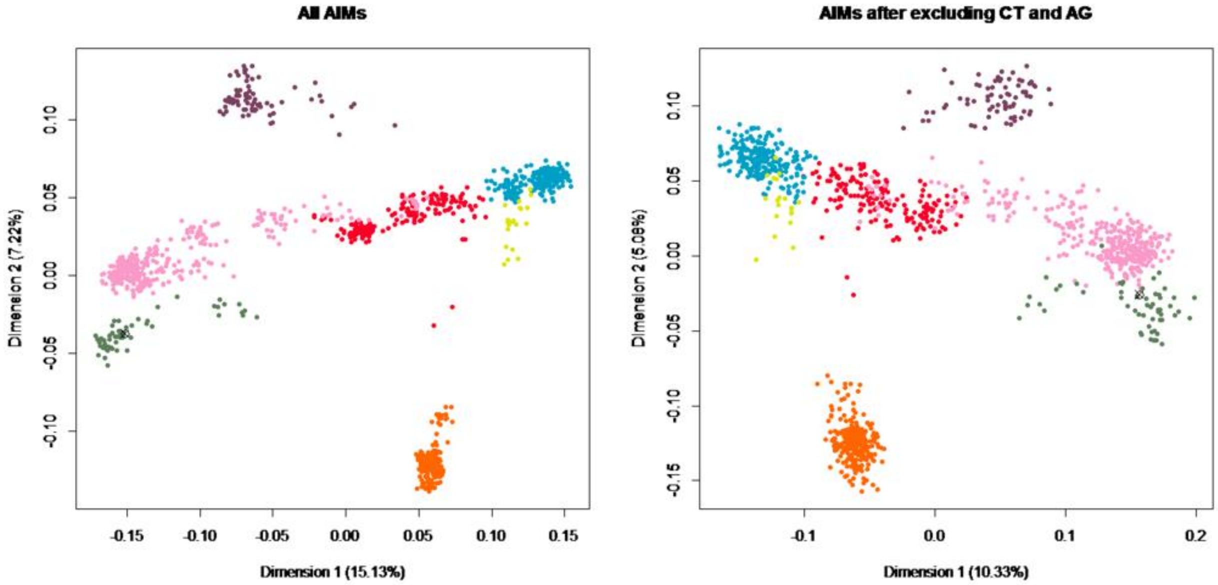
Supplementary Figure 7.3 Performance of the AIMs on the four parental ancestries. Boxplot of the ancestry estimates using FRAPPE (using K=4). See **Supplementary Figure 7.2** for sample pooling.



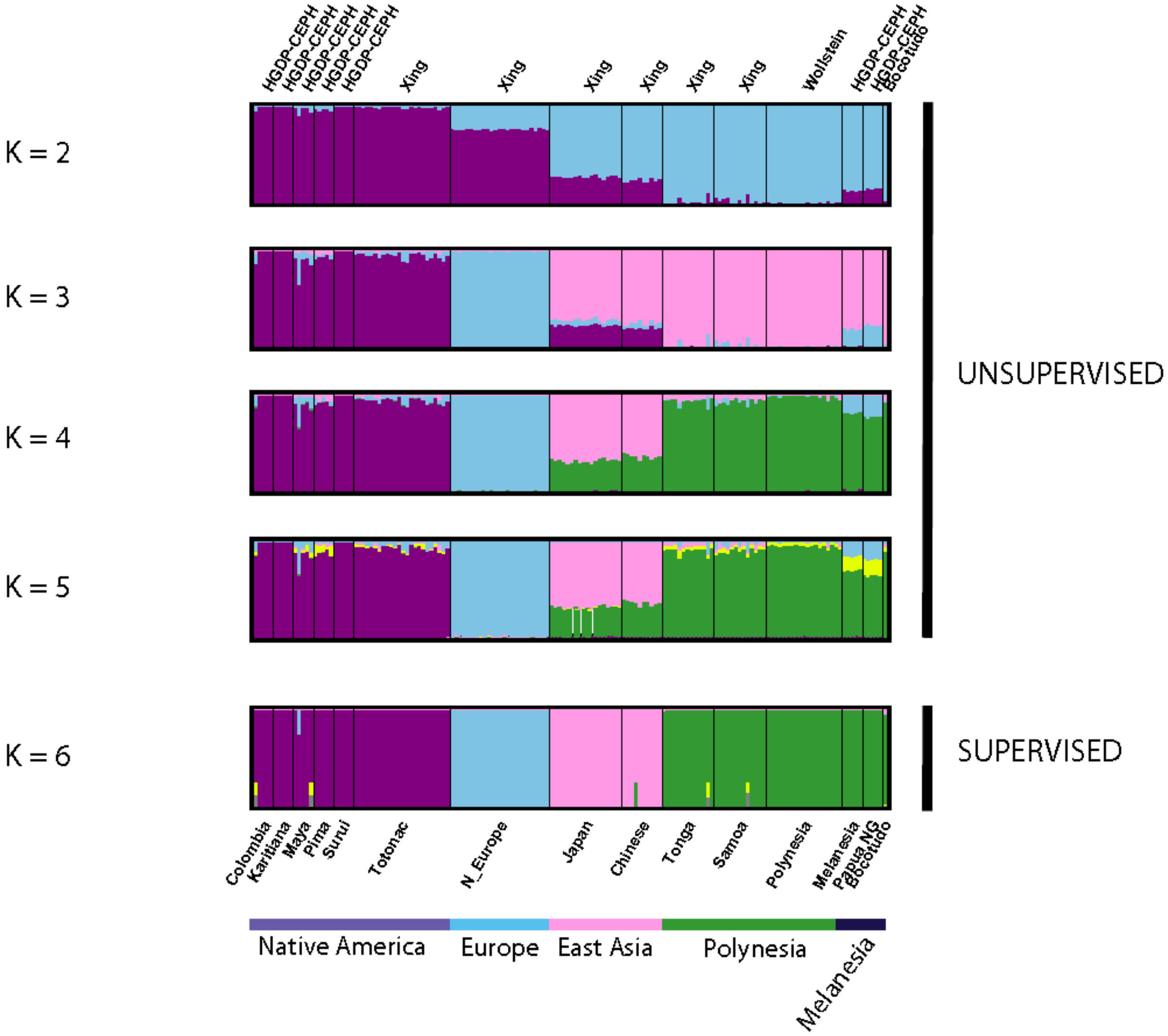
Supplementary Figure 7.4 Distribution of depth for the targeted markers after filtering bases with quality lower than 20.



Supplementary Figure 7.5 Number of reads covering each position around the target (position 0) on average and as expected by the number of probes. Since it is hard to visualize all 5,744 curves for the depth distribution, we then show three types of SNPs: 10 whose distribution reassemble the average, ten whose distribution is distinct and finally the SNPs with 0 depth (missed by the experiment).



Supplementary Figure 7.6 MDS plots for the first two dimensions for all the AIMs markers on the left and the markers excluding variants with alleles C and T and variants with alleles A and G (so that CT GA damaged sites are removed). The percentage of variation explained by each dimension is shown on the axis. For the legend: Oceania: green, East Asia: pink, Americas: purple, Central Asia: red, Africa: orange, Europe: light blue, Middle East: yellow. Bot15 is the black cross within the Oceanians.



Supplementary Figure 7.7 *Structure* analysis. On the 4 top panels un-supervised analysis for K=2…5, at the bottom panel the supervised analysis. The datasets are indicated on the very top and the broad geographical origin at the very bottom. The datasets are HGDP-CEPH (López Herráez et al. 2009), Xing (Xing et al. 2010), Wollstein (Wollstein et al. 2010). Bot15 is on the right of each panel.

Supplementary 8

Radiocarbon dating and isotopic analysis

Francisca Santana Sagredo*, Hannes Schroeder*, Murilo Bastos, Jan Heinemeier, Jesper Olsen, T. Douglas Price*, Thomas Higham*

*to whom correspondence should be addressed (francisca.santanasagredo@rlaha.ox.ac.uk, h.schroeder@snm.ku.dk, tdprice@wisc.edu, thomas.higham@rlaha.ox.ac.uk)

A. Radiocarbon dating and stable isotope analysis

Samples of tooth dentine from Botocudo human remains (Bot13, Bot15, Bot17, and Bot65) were dated at the Oxford Radiocarbon Accelerator Unit (ORAU) at the University of Oxford and the AMS ^{14}C Centre at Aarhus University (ACA). The dentine was sampled in Copenhagen in all cases. Bot15 was dated at ORAU and ACA. Standard collagen-extraction protocols were applied to the dating and preparation of the sample in Oxford (Brock, Ramsey, and Higham 2007). The dentine was treated in a 20mL precleaned glass vessel with sequential washes with 0.5 M HCl to decalcify the bone, 0.1 M NaOH for 30 min to remove humates, and 0.5 M HCl for 15 min at room temperature. Between each stage rinses with ultrapure MilliQ™ water were given. The 'collagen' was gelatinized in pH3 solution at 75°C for 20 hr then filtered with a pre-cleaned Ezee™-filter (Elkay, UK). The filtrate was then pipetted into a pre-cleaned Vivaspin 15™ 30 kD MWCO ultrafilter. This was centrifuged until 0.5-1.0 mL of the >30 kD fraction remained. This fraction was freeze-dried and weighed (see Supplementary Table 8.1). In Oxford, 4.89 mg of ultrafiltered collagen was combusted in an elemental analyser and the CO_2 graphitised prior to AMS dating. Samples for radiocarbon analysis in Aarhus were combusted in sealed evacuated tubes containing CuO and the resulting CO_2 was subsequently graphitised prior to AMS dating. The C/N atomic weight ratio and carbon and nitrogen stable isotope values of the collagen were measured using a Sercon Geo 20-20 mass spectrometer against alanine standards. In Aarhus the δ^{13}C, δ^{15}N and C/N atomic ratios were determined using an IsoPrime continuous flow IRMS coupled to an EuroVector elemental analyser. The pretreatment chemistry was identical to the Oxford method.

The results are shown in Supplementary Table 8.1. All samples yielded acceptable C/N atomic ratios and a high yield of collagen (between 1.4 and 13.7% by weight in both Oxford and Aarhus; modern collagen is ~20%). The other analytical parameters measured were also acceptable, and lead us to conclude that the material is quite well-preserved in terms of collagen.

The calibration of the radiocarbon results may be undertaken using the Southern Hemisphere calibration curve of McCormac et al. (2004) and the OxCal programme of (Ramsey 2009). Prior to calibration, however, it is important to check the dietary status of the individual to determine whether there are grounds to consider a reservoir effect, and whether this might influence the accuracy of the date. The stable isotope values for the individuals dated were compared with isotopic data from archaeological and modern fauna, as well as other archaeological human remains, from the regions of Minas Gerais, Sao Paulo, Rio de Janeiro and Santa Catarina. It is important to mention that stable isotope data from Brazilian archaeological contexts are currently scarce and therefore reliable comparisons are difficult. Other stable isotope values were also measured from Botocudo remains, although these were necessarily few in number (see below). We also obtained sulphur isotope values as a means to further support our dietary analysis of the humans in terms of a possible marine influence.

Finally, radiogenic strontium isotope analysis was performed on all four individuals with the aim of studying their mobility patterns and possible place of origin (see next section). It is important to note, however, that the dearth of biologically available strontium isotope values for South America limits the interpretation of the $^{87}Sr/^{86}Sr$ isotope results. In particular, it is difficult to associate the obtained values with a specific geographical-geological area.

Isotopic ecology from Minas Gerais, Sao Paulo, Rio de Janeiro and Santa Catarina

Isotopic data of archaeological fauna from Minas Gerais, Sao Paulo and Santa Catarina were compared with the values of the Botocudo individuals, but particularly Bot15 and Bot17 since we were interested in reliably calibrating their radiocarbon ages. Modern faunal values from Santa Catarina and Rio de Janeiro were also included.

The information from Minas Gerais (Hermenegildo 2009), including sites from Lagoa Santa and Vale do Peruaçu, embraces mainly archaeological terrestrial fauna, including small to medium sized animals from the Initial, Middle and Late Holocene periods. The dataset also includes one amphibian (frog) and one reptile (crocodile). Values of terrestrial fauna presented by De Massi (2009) and Colonese et al. (2014) from Santa Catarina and Sao Paulo are also included.

The terrestrial values show significant variability, with a range from -24.6‰ to -11.1‰ for $\delta^{13}C$, and 2.4‰ to 13.3‰ for $\delta^{15}N$. Some of the samples show enriched values in ^{13}C, probably related to the consumption of C_4 plants, and others with higher $\delta^{15}N$ values are associated with a more carnivorous diet. Regarding consumption of C_4 plants by some of the specimens, e.g. *Dasypus* sp., these cannot be associated with the consumption of maize because these individuals date to the Mid-Holocene, before maize was introduced to this region. The plants consumed by *Dasypus* sp., therefore, probably correspond to other C_4 plants present in the tropical forests of Minas Gerais. In contrast, the values of *Cavia* sp. can be related to consumption of maize, because these animals postdate the introduction of domesticated plants in Brazil. Other dietary changes through time were observed for *Tayassu* sp., which change from an herbivore to carnivore diet between the Middle and Late Holocene.

Marine fauna from modern and archaeological samples from Santa Catarina have been analyzed by de Massi (2009) and Colonese et al. (2014). These samples included fish, shellfish, reptiles, seabirds and sea mammals. As expected, the values for marine fauna present high nitrogen values together with high carbon values. The range observed for $\delta^{13}C$ values is -20.2‰ to -9.0‰ while $\delta^{15}N$ values range from 5.7‰ to 18.9‰. The lowest values observed correspond to shellfish while the highest come from seabirds.

Regarding aquatic flora and fauna, Brito et al. (2006) analyzed modern samples from the small stream of Córrego da Andorinha located in Ilha Grande, Rio de Janeiro. These authors analyzed mainly algae and microalgae and suggest that aquatic fauna of the stream consumed microalgae rather than other types of algae. The fauna analyzed included fish and shrimps, as well as some insects. The range of values for $\delta^{13}C$ and $\delta^{15}N$ observed in fish and shrimps correspond to -24.8‰ to -17.1‰ and 8.2‰ to 6.6‰, respectively. The $\delta^{13}C$ values for the algae and microalgae range from -30.7‰ to -19‰, while the $\delta^{15}N$ values range from 1.4‰ to 2.6‰.

All stable isotopic values of terrestrial, marine and aquatic fauna were plotted together with the values obtained for individuals Bot15 and Bot17 (Supplementary Figure 8.1). In addition, stable isotope values from other Botocudo individuals (Bot13 and Bot65) were included in the graph to compare their values with the local fauna. While there is wide variability in the terrestrial fauna as mentioned before, the marine fauna forms a clear group, with some outliers (shellfish). The aquatic faunal values overlap partially with the terrestrial data. It can be seen that individual Bot15 (as well as Bot13 and Bot65) is not directly related to the type of marine diet attested from the coast of Brazil, since its carbon isotope ratio should be more enriched in ^{13}C compared with the marine fauna and clearly is not. Although it is difficult to entirely rule out the consumption of shellfish, this would not be consistent with the nitrogen values observed for the shellfish and individuals Bot15 and Bot65, which differ by around 7‰. It is expected that differences in trophic level between the animal and its consumer will be around 2 to 6‰ (Sponheimer et al. 2003), and in this case the difference is higher than the range expected. If there was consumption of marine resources this must have been in a moderate proportion. Since the values of Bot13, Bot15 and Bot65 fall in between the carbon and nitrogen isotope values for terrestrial and marine fauna, a possible consumption of a mixed diet cannot be ruled out. Aquatic fauna could also have been consumed by these individuals but probably as a secondary resource. It is likely though that the variation in the stable isotopic composition of stream and lake fauna might be higher. Hence it is not possible to eliminate the consumption of aquatic resources completely and more information is needed in order to be more confident in this interpretation.

The more enriched ^{13}C values observed for Bot15, Bot13 and Bot65 may be also related to the consumption of animals such as *Dasypus* sp. and *Cavia* sp. These terrestrial species have been recovered from archaeological sites in Minas Gerais and show values very enriched in ^{13}C. The range observed for $\delta^{13}C$ values in *Dasypus* sp. ranges from -20.5‰ to -12.6‰ while for *Cavia* sp. the range is in between -21.3 to -11.1‰. Yet, both animals were consuming different types of C_4 plants, with *Dasypus* sp. probably ingesting wild plants while *Cavia* sp. is known to consume mainly maize (Hermenegildo 2009). It is important to note that by the time of the occupation by Botocudo people, maize was already integrated in the horticultural systems by some groups in Brazil (De Massi 2009). Another animal that also yielded positive carbon values was *Euphractus* sp. with a maximum $\delta^{13}C$ of -15.4‰, a value probably related to the consumption of wild C_4 plants. These resources may be an influence on the human values in terms of their slight carbon isotope enrichment.

A similar situation occurs with the $\delta^{15}N$ values for *Dasypus* sp. and *Euphractus sexcinctus*. The values are quite positive for terrestrial animals, with the $\delta^{15}N$ values for *Dasypus* sp. falling in a range between 7.8 and 11.3‰ (De Massi 2009), while the maximum value for *Euhpractus sexcinctus* is 13.3‰ (Colonese et al. 2014). It is possible that these might also be an influence on enriched human values.

The isotopic evidence for individual Bot17 is different from the other three archaeological human values we collected. It is strongly suggestive of a marine diet, albeit one that does not seem to be associated with the consumption of Brazilian coastal marine resources. High values obtained for Bot17 could be related to the values observed for sea mammals (max. values of -11.4‰ for δ^{13}C and 16.4‰ for δ^{15}N), sharks (max. values of -9.5‰ for δ^{13}C and 16.0‰ for δ^{15}N), anchovies (max. values of -11.4‰ for δ^{13}C and 13.3‰ for δ^{15}N), magellanic penguin (max. values of -11.2‰ for δ^{13}C and 14.5‰ for δ^{15}N) and seabirds (max. values of -9.8‰ for δ^{13}C and 18.9‰ for δ^{15}N) rather than small fish, shellfish or turtles (De Massi 2009; Colonese et al. 2014). This might support the notion that Bot17 is an outsider or migrant compared to the areas for which we found reference isotope data in Brazil. Unfortunately, Brazil is still poorly mapped; it is therefore impossible to be certain that Bot17 is a migrant, let alone a migrant from outside Brazil (as the genetic data could suggest) based on the δ^{13}C and δ^{15}N isotope data.

Isotopic analysis of ancient human remains from Minas Gerais, Sao Paulo, Santa Catarina and Amazonia

Stable isotope data of terrestrial human samples from Minas Gerais (Hermenegildo 2009), Sao Paulo (Colonese et al. 2014), Santa Catarina (De Massi 2009) and Amazonia (De Massi 2009) are presented in Supplementary Figure 8.2. In addition, isotopic data from different coastal sites of Santa Catarina (De Massi 2009; Colonese et al. 2014) are shown. A third group ("Transition Inland-Coastal population") was also included in the comparisons. This group is composed of human samples from the site of Piaçaguera in Sao Paulo (Colonese et al. 2014). The site is located around 12 km away from the coast and shows a mix of marine and terrestrial fauna.

Three different groups are clearly seen, which shows that it is possible to differentiate terrestrial, coastal and transition inland-coast groups. There are some terrestrial individuals that yield higher δ^{13}C values, which are probably related to the consumption of maize. In addition, there are some individuals from the coast that have a terrestrial dietary signal with the consumption of maize, rather than a coastal one. This is in accordance with the archaeological data from the sites (De Massi 2009). It is interesting that individual Bot15 does not group with the individuals from the coast or the inlands, fitting better with the values of δ^{13}C and δ^{15}N observed for the transitional inland-coastal group.

Individual Bot13 presents more positive carbon isotope values than the ones observed for the terrestrial populations from Brazil. Nevertheless, this value is still lower than the values from the coastal populations. For this reason, we suggest that individuals Bot13 and Bot65 follow a similar trend as individual Bot15, plotting their carbon and nitrogen values closer to the transitional inland-coastal group, probably consuming a mixed diet. We are not able to exclude the possibility that the higher δ^{13}C values observed for Bot15 and Bot65 could be related to the consumption of maize - either directly or indirectly via animals consuming the crop. By 1400 AD maize agriculture was already established in the region (Hermenegildo 2009).

With respect to individual Bot17, its δ^{15}N value is higher than the maximum values observed for the coastal individuals. There is a difference of 3.7‰ between Bot17 and the individuals with the highest nitrogen value analysed by (De Massi 2009). At the same time, a significant difference can also be seen between the value of δ^{13}C from this individual (-14.8‰) and the average carbon value observed for the coastal populations of Brazil (-11.8 ±1.7‰). Once again, with respect to the archaeological stable isotope data, the δ^{13}C for Bot17 suggests the consumption of a different type of marine fauna than that exploited by the coastal populations of Brazil, at least from Sao Paulo and Santa Catarina.

It is difficult to draw any definitive conclusions from a limited dataset. However, in instances such as this one, we would usually invoke a mixed diet for values that sit so clearly apart from the major marine/terrestrial groupings. This interpretation agrees with the similar values observed for Bot13, Bot15 and Bot65 when compared to the transitional inland-coastal group, which has been characterized as having a mixed diet. In summary, we conclude that Bot13, Bot15 and Bot65 probably subsisted on a mixed diet, perhaps taking foods from a mix of terrestrial, aquatic and marine sources. Bot17, however, likely consumed more marine protein.

Sulphur isotope analysis

We also measured δ^{34}S values for Bot17 and Bot15 to confirm whether there were grounds to consider a reservoir effect in their radiocarbon ages or not. Sulphur isotopes are useful because in marine ecosystems δ^{34}S values derived from producers are consistently elevated (ca. +20‰) and consistent with values of oceanic sulphates, whilst the δ^{34}S values of terrestrial and freshwater producers are far more variable (ca. -20‰ to +20‰) (Craig et al. 2006). For this reason the measurement of δ^{34}S from bone collagen, where it is present at around 0.2%, has become useful for allowing human diets to be broadly discriminated between marine and terrestrial partitioned sources. Sulphur contamination from salt-spray, the effects of coastal precipitation and, possibly, the burning of recent fossil fuels is a potential problem, however. A further complication is the general lack of baseline sulphur isotope data from modern and archaeological material in many regions including Minas Gerais. We obtained δ^{34}S values from Bot15 and 17 to explore further the potential for these samples being derived from coastal or marine humans, rather than terrestrial consumers living in the interior of Brazil. A significant problem with the analysis of course is the lack of modern or archaeological analogues for comparison. In the light of the carbon and nitrogen isotope values for Bot17, however, we can be sure that this individual is a marine consumer. Comparing the δ^{34}S value of this individual with the value for Bot15 is useful in order to place that individual into a marine or terrestrial context.

The results are shown in Supplementary Table 8.2. The value for Bot17 is enriched, and close to the values obtained for the seal bone collagen standard we used from the South African Cape. The δ^{34}S value of 14.3‰ is consistent with a marine input to the diet as expected on the basis of the other stable isotopes. The value for Bot15 is marginally lower, at 13.8‰, but still enriched and in our view indicative of a partial marine input to the diet with the caveats outlined above. This interpretation is in agreement with our conclusion about a mixed diet for Bot15. Of course more research is required to confirm this, to obtain other values from archaeological remains, and to explore the wider issue of regional sulphur isotope variability, but for the time being the values suffice to strongly imply a marine contribution.

Calibration of the radiocarbon dates

The calibrated dates are shown in Supplementary Table 8.3. Bot13 and Bot65 were calibrated using the SHCAL04 curve (McCormac et al. 2004), accounting for the Southern Hemisphere offset to the Northern Hemisphere radiocarbon concentration. We then calibrated the radiocarbon age for Bot17 by incorporating an estimated contribution of ^{14}C-depleted marine carbon. This was estimated using a linear interpolation model in which we assume that collagen with 100% terrestrial carbon would have a δ^{13}C value of -20 ± 1‰ and collagen from a fully marine organism would have a value of -11 ± 1‰. This provided an estimate of marine carbon protein in the Bot17 individual of 60 ± 16%. We used the mixed curves method in OxCal (Ramsey 2009) and the SHCAL04 curve for the terrestrial component and the Marine09 curve for the oceanic equivalent. In addition, we assumed no local ΔR offset, and therefore simply used the average world ocean calibration curve with a ΔR of 0±20 years to account for the uncertainty (see below). If we model the Bot17 age using these estimates the corrected calibrated range under these assumptions is shown in Supplementary Figure 8.3. The range is imprecise but broadly spans the range c. 1500—1840 AD (at 95.4%). Work some of our group have recently undertaken suggests that δ^{13}C values are enriched by 1.8‰ for every 100 radiocarbon years in the reservoir effect (Craig et al. 2013), which fits quite well with our assumed values. Of course it is important to remember that radiocarbon reservoir effects vary, sometimes dramatically, due to localized upwelling zones, the presence of hardwater effects, exchange by shellfish with atmospheric CO_2 through increased aeration, as well as the direct ingestion of old carbon by live shellfish (Petchey et al. 2011).

We calculated a combined age for the two determinations on the Bot15 specimen. On the basis of the stable isotopes, and the sulphur result, we incorporated a marine reservoir correction for it as well. We used the same linear approximation as previously applied, and estimate a 30±16% marine protein uptake and a calibrated age range of 1479-1708 (83.9%) and 1730-1804 (11.5%) at 95.4% probability (Supplementary Figure 8.3). The same caveats as outlined above apply, radiocarbon reservoir effects vary widely and this would introduce the possibility of some variation outside the quoted range.

Comparing the calibrated dates with historical events

As discussed in the main text, several historical events are of importance to rule out scenarios related to the origin of the individuals. We therefore compared the calibrated dates with the dates of those events to evaluate the probability that Bot15 and Bot17 were alive at the time of those events. Specifically we used OxCal (Ramsey 2009) to compute the probability density function (PDF) of the difference ("difference PDF") between the calibrated dates and the beginning of established contact between Europeans and Polynesians (Thomas 2010), the beginning of the Madagascar-Brazil slave trade in 1718AD[1] (Eltis 2013) and the beginning of the Polynesia-Peru slave trade in 1862 AD (Maude 1981). When calculating the difference PDF for Bot15 and Bot17 the marine reservoir offset was included. We then computed the probability that the difference is larger than 0 by numerically integrating over the values of the difference PDF above 0. This is essentially the same routine as using the Order function in OxCal.

We found that Bot15 (respectively, Bot17) was still alive by the time the Europeans established contacts with Polynesians with probability 0.08 (respectively, 0.19), by the beginning of the Madagascar-Brazil slave trade with probability 0.13 (respectively, 0.31) and by the beginning of the Peru-Brazil slave trade with probability 0.006 (respectively, 0.027, Supplementary Figure 8.4).

B. Strontium isotope analysis

Strontium isotopes in prehistoric human teeth provide a geochemical signature of the place of birth or, depending on which tooth is sampled, the area where a person spent their childhood. In archaeology, strontium isotope analysis is used to track human mobility and to identify possible non-local individuals within a burial population. The technique has been in use for over 20 years and is described in detail in a number of articles (Price, Grupe, and Schroeter 1994; Price, Burton, and Bentley 2002; Bentley 2006; Price et al. 2008). One of the underlying principles of strontium isotope analysis in archeology is that strontium isotope ratios (^{87}Sr/^{86}Sr) vary with the age and geochemical composition of the rock (Faure 2001). As a result, the ^{87}Sr/^{86}Sr ratios of the earth's crust are highly variable. For example, very old (> 100 my) rocks, such as shales and granites, usually have high ^{87}Sr/^{86}Sr ratios typically above 0.710 and as high as 0.740. In contrast, geologically young rocks (< 1-10 my) typically have lower ^{87}Sr/^{86}Sr ratios below 0.706. These variations may seem small, but they are exceptionally large from a geological standpoint. Strontium is taken up by plants from soil and water, moving then into the food chain and being finally incorporated into human tissues such as bones and teeth (Bentley 2006). Tooth enamel mainly forms during infancy and early childhood. A person's tooth enamel ^{87}Sr/^{86}Sr ratios will therefore reflect the geological values of the area in which they grew up. If a person's enamel ^{87}Sr/^{86}Sr value does not match the local geology where the individual was buried we can conclude that he or she did not grow up in the area. However, to reliably identify an individual as non-local we need to be able to establish a local base-line of biologically available ^{87}Sr/^{86}Sr values. This is complicated by the fact that geological ^{87}Sr/^{86}Sr values can vary quite substantially over a relatively small geographic area. One way of dealing with this has been to measure ^{87}Sr/^{86}Sr values found in local plants or low-mobility animal species, which tend to provide a more reliable baseline than geological ^{87}Sr/^{86}Sr values (Bentley 2006).

In principle, it is also possible to determine where a person originated by matching human and geological or biologically available ^{87}Sr/^{86}Sr values. However, in practice this is complicated by the fact that different geographic areas may yield the same or very similar ^{87}Sr/^{86}Sr values because of a similar geology. Although it is relatively straightforward to identify non-locals, this

[1] http://slavevoyages.org/tast/database/search.faces?yearFrom=1514&yearTo=1866&mjbyptimp=60811&mjslptimp=50000

“equifinality problem” makes it generally impossible to determine their precise geographic origins with any degree of confidence (see, *e.g.*, Schroeder et al. (2009)).

Analytical procedure

Strontium isotope ratios were obtained from enamel samples from four Botocudo individuals. We deliberately chose enamel over bone because we were interested in establishing the individuals’ childhood origins. Moreover, enamel has been shown to be generally more resistant to diagenesis than bone and is therefore a more reliable indicator of biogenic strontium values (e.g., (Kohn, Schoeninger, and Barker 1999; Budd et al. 2000; Hoppe, Koch, and Furutani 2003)). Unfortunately, it was not possible to sample the same tooth for all four individuals. For Bot15 and Bot17, we sampled an upper left 1st pre-molar and an upper left 1st molar, which start forming in the first two years of life and around birth, respectively. For Bot13 and Bot65 we sampled a lower right and left 3rd molar, respectively, which start forming much later, around age 8-10. Samples were analyzed at the Department of Geological Sciences, University of North Carolina at Chapel Hill using Thermal Ionization Mass Spectrometry (TIMS). Enamel samples were mechanically cleaned and then around 2-5 mg was dissolved in 5 M nitric acid. The strontium fraction was purified using EiChrom Sr-Spec resin and eluted with nitric acid followed by water. Isotope ratios were obtained on this strontium fraction using a VG (Micromass) Sector 54 thermal ionization mass spectrometer. Analyses of strontium standard NIST987 averaged 0.71026 ± 0.00001 (2s; n = 30). The relative standard errors for those four samples ranged between 0.0006 to 0.0009 %.

Results and Discussion

The strontium isotope measurements yielded two different kinds of $^{87}Sr/^{86}Sr$ ratios for the four Botocudo individuals (Supplementary Table 8.4). Bot15 and Bot17 both yielded a value of around 0.7083, which lies towards the lower end of the $^{87}Sr/^{86}Sr$ range, just below the value for modern seawater (0.7092). By contrast, Bot65 and Bot13 yielded much higher $^{87}Sr/^{86}Sr$ ratios of around 0.7271 and 0.7273, respectively. These two sets of values are clearly distinct and we can, therefore, conclude that these four individuals did not grow up in the same place. However, it is possible that Bot15 and Bot17, and Bot13 and Bot65 did grow up in the same area as their $^{87}Sr/^{86}Sr$ values are very similar. This is consistent with the fact that the remains were found at different sites in Espírito Santo and Minas Gerais (S1).

While it seems safe to say that the Botocudos did not grow up in the same place, it is much more difficult to determine *where* they grew up. The problem we face is that the two regions that are relevant to this study, namely Brazil (where the samples were found) and Polynesia (which includes populations with the closest genetic affinity to two of them) are vast and relatively poorly understood in terms of their strontium isotope geology. With the exception of New Zealand, the islands of Polynesia are primarily volcanic and are often capped in coral limestone. As such, they tend to display geological $^{87}Sr/^{86}Sr$ values that are intermediate between volcanic basalts (0.702-0.704) and marine-derived carbonates (0.707-0709, Nishio et al. 2005; Shaw et al. 2009; Kinaston et al. 2013). By contrast, Brazil is geologically much more complex and although its strontium isotope geology remains only poorly understood, it is probably safe to say that it encompasses a much wider range of geological $^{87}Sr/^{86}Sr$ values. Unfortunately, we do not have any $^{87}Sr/^{86}Sr$ data for the Rio Doce Basin or the areas of Espírito Santo were the remains were found (S1), but a recent study of geological and biological samples from the site of Lagoa Santa (Machado 2013), which is located around 200 km away from the Rio Doce Basin, yielded a wide range of $^{87}Sr/^{86}Sr$ values (0.708-0.737). As can be seen in Supplementary Figure 8.5, which also includes geological and biological $^{87}Sr/^{86}Sr$ values from two other sites in Brazil (Mantovani et al. 1985; Bastos et al. 2011), this range encompasses the two sets of $^{87}Sr/^{86}Sr$ values obtained for the Botocudos and we, therefore, have to conclude that their $^{87}Sr/^{86}Sr$ values are consistent with an origin in Brazil.

References

Bastos, Murilo Q. R., Sheila M. F. Mendonça de Souza, Roberto V. Santos, Bárbara A. F. Lima, Ricardo V. Santos, and Claudia Rodrigues-Carvalho. 2011. “Human Mobility on the Brazilian Coast: An Analysis of Strontium Isotopes in Archaeological Human Remains from Forte Marechal Luz Sambaqui.” *Anais Da Academia Brasileira de Ciências* 83 (2): 731–43. doi:10.1590/S0001-37652011000200030.

Bentley, R. Alexander. 2006. “Strontium Isotopes from the Earth to the Archaeological Skeleton: A Review” 13 (3): 135–87.

Brito, Ernesto Fuentes, Timothy P. Moulton, Marcelo L. De Souza, and Stuart E. Bunn. 2006. “Stable Isotope Analysis Indicates Microalgae as the Predominant Food Source of Fauna in a Coastal Forest Stream, South-East Brazil.” *Austral Ecology* 31 (5): 623–33. doi:10.1111/j.1442-9993.2006.01610.x.

Brock, Fiona, Christopher Bronk Ramsey, and Thomas Higham. 2007. “Quality Assurance of Ultrafiltered Bone Dating.” *Radiocarbon* 49 (2): 187–92. doi:10.2458/azu_js_rc.v49i2.2917.

Budd, Paul, Janet Montgomery, Barbara Barreiro, and Richard G. Thomas. 2000. “Differential Diagenesis of Strontium in Archaeological Human Dental Tissues.” *Applied Geochemistry* 15 (5): 687–94. doi:10.1016/S0883-2927(99)00069-4.

Colonese, André Carlo, Matthew Collins, Alexandre Lucquin, Michael Eustace, Y. Hancock, Raquel de Almeida Rocha Ponzoni, Alice Mora, et al. 2014. “Long-Term Resilience of Late Holocene Coastal Subsistence System in Southeastern South America.” *PLoS ONE* 9 (4): e93854. doi:10.1371/journal.pone.0093854.

Craig, O. E., L. Bondioli, L. Fattore, T. Higham, and R. Hedges. 2013. “Evaluating Marine Diets through Radiocarbon Dating and Stable Isotope Analysis of Victims of the AD79 Eruption of Vesuvius.” *American Journal of Physical Anthropology* 152 (3): 345–52. doi:10.1002/ajpa.22352.

Craig, O.E., R. Ross, Søren H. Andersen, N. Milner, and G.N. Bailey. 2006. “Focus: Sulphur Isotope Variation in Archaeological Marine Fauna from Northern Europe.” *Journal of Archaeological Science* 33 (11): 1642–46. doi:10.1016/j.jas.2006.05.006.

De Massi, M. 2009. “Aplicações de Isótopos Estáveis de 18/16O, 13/12C E 15/14N Em Estudos de Sazonalidade, Mobilidade E Dieta de Populações Pré-Históricas No Sul Do Brasil.” 22 (2): 55–76.

Eltis, D. 2013. “A Brief Overview of the Trans-Atlantic Slave Trade,” Voyages: The Trans-Atlantic Slave Trade Database,” Accessed May 27. http://www.slavevoyages.org/tast/assessment/essays-intro-01.faces.

Faure, Gunter. 2001. *Origin of Igneous Rocks - The Isotopic Evidence*. Springer Verlag Berlin Heidelberg New York. http://www.springer.com/earth+sciences+and+geography/geochemistry/book/978-3-540-67772-7.

Hermenegildo, T. 2009. “Reconstitução Da Dieta E Dos Padres de Subsistencia Das Populações Pré-Históricas de Caçadores-Colectores Do Brasil Central Através Da Ecología Isotópica.” Dissertação apresentada para obtenção da título de Mes em Ecología Aplicada., São Paulo, Brazil: Universidade de São Paulo.

Hoppe, K. A., P. L. Koch, and T. T. Furutani. 2003. “Assessing the Preservation of Biogenic Strontium in Fossil Bones and Tooth Enamel.” *International Journal of Osteoarchaeology* 13 (1-2): 20–28. doi:10.1002/oa.663.

Kinaston, Rebecca L., Richard K. Walter, Chris Jacomb, Emma Brooks, Nancy Tayles, Sian E. Halcrow, Claudine Stirling, et al. 2013. “The First New Zealanders: Patterns of Diet and Mobility Revealed through Isotope Analysis.” *Plos One* 8 (5): e64580. doi:10.1371/journal.pone.0064580.

Kohn, Matthew J., Margaret J. Schoeninger, and William W. Barker. 1999. “Altered States: Effects of Diagenesis on Fossil Tooth Chemistry.” *Geochimica et Cosmochimica Acta* 63 (18): 2737–47. doi:10.1016/S0016-7037(99)00208-2.

Machado, M., C. 2013. “Metodologias Isotópicas 87Sr/86Sr δ13C E δ18O Em Estudos Geológicos E Arqueológicos”. Tese de Doutorado, Porto Alegre, Brazil: Universidade Federal do Rio Grande do Sul.

Mantovani, M. S. M., L. S. Marques, M. a. De Sousa, L. Civetta, L. Atalla, and F. Innocenti. 1985. “Trace Element and Strontium Isotope Constraints on the Origin and Evolution of Paraná Continental Flood Basalts of Santa Catarina State (Southern Brazil).” *Journal of Petrology* 26 (1): 187–209. doi:10.1093/petrology/26.1.187.

Maude, Henry Evans. 1981. *Slavers in Paradise: The Peruvian Slave Trade in Polynesia, 1862-1864*. Stanford University Press.

McCormac, F. G., Alan G. Hogg, Paul G. Blackwell, Caitlin E. Buck, Thomas F. G. Higham, and Paula J. Reimer. 2004. “SHCal04 Southern Hemisphere Calibration, 0–11.0 Cal Kyr BP.” *University of Waikato Research Commons*. http://researchcommons.waikato.ac.nz/handle/10289/3687.

Nishio, Yoshiro, Shun’ichi Nakai, Tetsu Kogiso, and Hans G. Barsczus. 2005. “Lithium, Strontium, and Neodymium Isotopic Compositions of Oceanic Island Basalts in the Polynesian Region: Constraints on a Polynesian HIMU Origin.” *GEOCHEMICAL JOURNAL* 39 (1): 91–103. doi:10.2343/geochemj.39.91.

Petchey, Fiona, Atholl Anderson, Albert Zondervan, Sean Ulm, and Alan Hogg. 2011. “New Marine ΔR Values for the South Pacific Subtropical Gyre Region.” *Radiocarbon* 50 (3): 373–97. doi:10.2458/azu_js_rc.v50i3.3221.

Price, T. D., J. H. Burton, and R. A. Bentley. 2002. “The Characterization of Biologically Available Strontium Isotope Ratios for the Study of Prehistoric Migration.” *Archaeometry* 44 (1): 117–35. doi:10.1111/1475-4754.00047.

Price, T. Douglas, James H. Burton, Paul D. Fullagar, Lori E. Wright, Jane E. Buikstra, and Vera Tiesler. 2008. “Strontium Isotopes and the Study of Human Mobility in Ancient Mesoamerica.” *Latin American Antiquity* 19 (2): 167–80. doi:10.2307/25478222.

Price, T.D., G. Grupe, and P. Schroeter. 1994. “Reconstruction of Migration Patterns in the Bell Beaker Period by Stable Strontium Isotope Analysis” 9: 413–17.

Ramsey, Christopher Bronk. 2009. “Bayesian Analysis of Radiocarbon Dates.” *Radiocarbon* 51 (1): 337–60. doi:10.2458/azu_js_rc.v51i1.3494.

Schroeder, Hannes, Tamsin C. O’Connell, Jane A. Evans, Kristrina A. Shuler, and Robert E.M. Hedges. 2009. “Trans-Atlantic Slavery: Isotopic Evidence for Forced Migration to Barbados.” *American Journal of Physical Anthropology* 139 (4): 547–57. doi:10.1002/ajpa.21019.

Shaw, Ben J., Glenn R. Summerhayes, Hallie R. Buckley, and Joel A. Baker. 2009. "The Use of Strontium Isotopes as an Indicator of Migration in Human and Pig Lapita Populations in the Bismarck Archipelago, Papua New Guinea." *Journal of Archaeological Science* 36 (4): 1079–91. doi:10.1016/j.jas.2008.12.010.

Sponheimer, M., T. Robinson, L. Ayliffe, B. Roeder, J. Hammer, B. Passey, A. West, T. Cerling, D. Dearing, and J. Ehleringer. 2003. "Nitrogen Isotopes in Mammalian Herbivores: Hair δ15N Values from a Controlled Feeding Study." *International Journal of Osteoarchaeology* 13 (1-2): 80–87. doi:10.1002/oa.655.

Thomas, Nicholas. 2010. *Islanders: The Pacific in the Age of Empire*. Yale University Press.

Supplementary Table 8.1 Radiocarbon date and analytical data for the samples dated at the ORAU and Aarhus. Stable isotope values are reported in per mille notation at ±0.2‰ and ±0.3‰ precision for C and N respectively. 'Yield' is yield of collagen and %C is the percentage of carbon recorded during the combustion of the collagen. This value is consistent with our expected values. The range of acceptability at ORAU for C/N atomic ratios is 2.9-3.5. Modern collagen is 3.21.

Lab number	Specimen	Radiocarbon age BP	Dentine weight used (mg)	Yield (mg)	%Yld	%C	δ^{13}C (‰)	δ^{15}N (‰)	C/N
AAR-17656	Bot13	195±25	173	16.9	9.8	41.9	-16.2	13.1	3.2
OxA-27184	Bot15	408 ±24	250	34.2	13.7	42.1	-17.3	13.2	3.2
AAR-17522	Bot15	417±25	708	9.9	1.4	38.8	-17.0	13.9	3.2
AAR-17657	Bot17	487±25	267	29.8	11.2	37.8	-14.8	18.2	3.2
AAR-17658	Bot65	185±25	196	22.8	11.7	41.4	-17.1	12.7	3.2

Supplementary Table 8.2 Sulphur and associated analytical data for the samples analysed by IsoAnalytical (UK). The bone collagen samples were analysed alongside standards composed of seal bone (obtained from the Cape Province, South Africa) and cow bone collagen (Oxford, England). Sulphur Isotope values are reported in per mille notation at ±0.3‰.

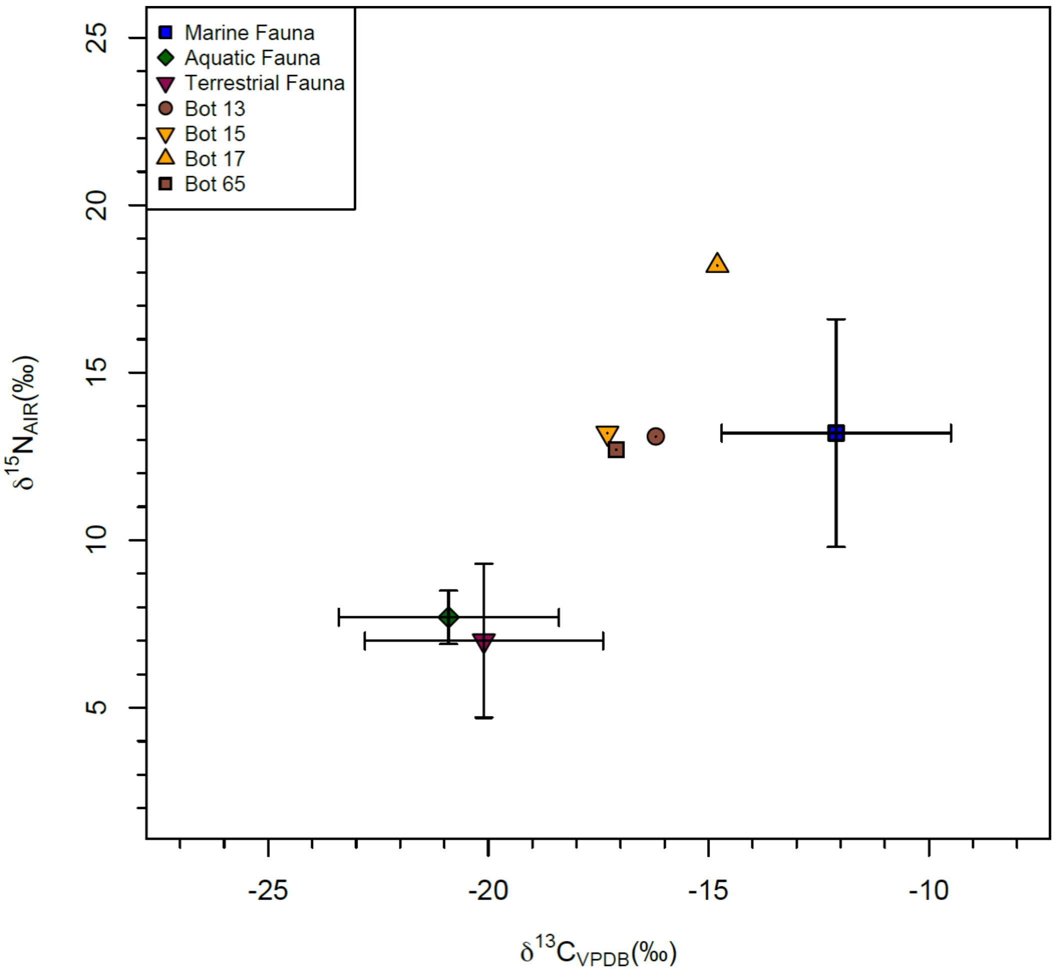
Sample Code	Sulphur content (%)	$\delta^{34}S_{V\text{-}CDT}$ (‰)
Bot13	0.21	12.0
Bot15	0.24	13.8
Bot17	0.20	14.3
Bot65	0.20	11.4
SEA1 seal bone standard	0.22	15.6
“	0.21	15.5
“	0.24	15.3
SMBG Cow bone standard	0.19	8.3
“	0.19	8.1
“	0.22	8.2
“	0.22	8.0

Supplementary Table 8.3 Summary of all four calibrated dates mentioned throughout the text.

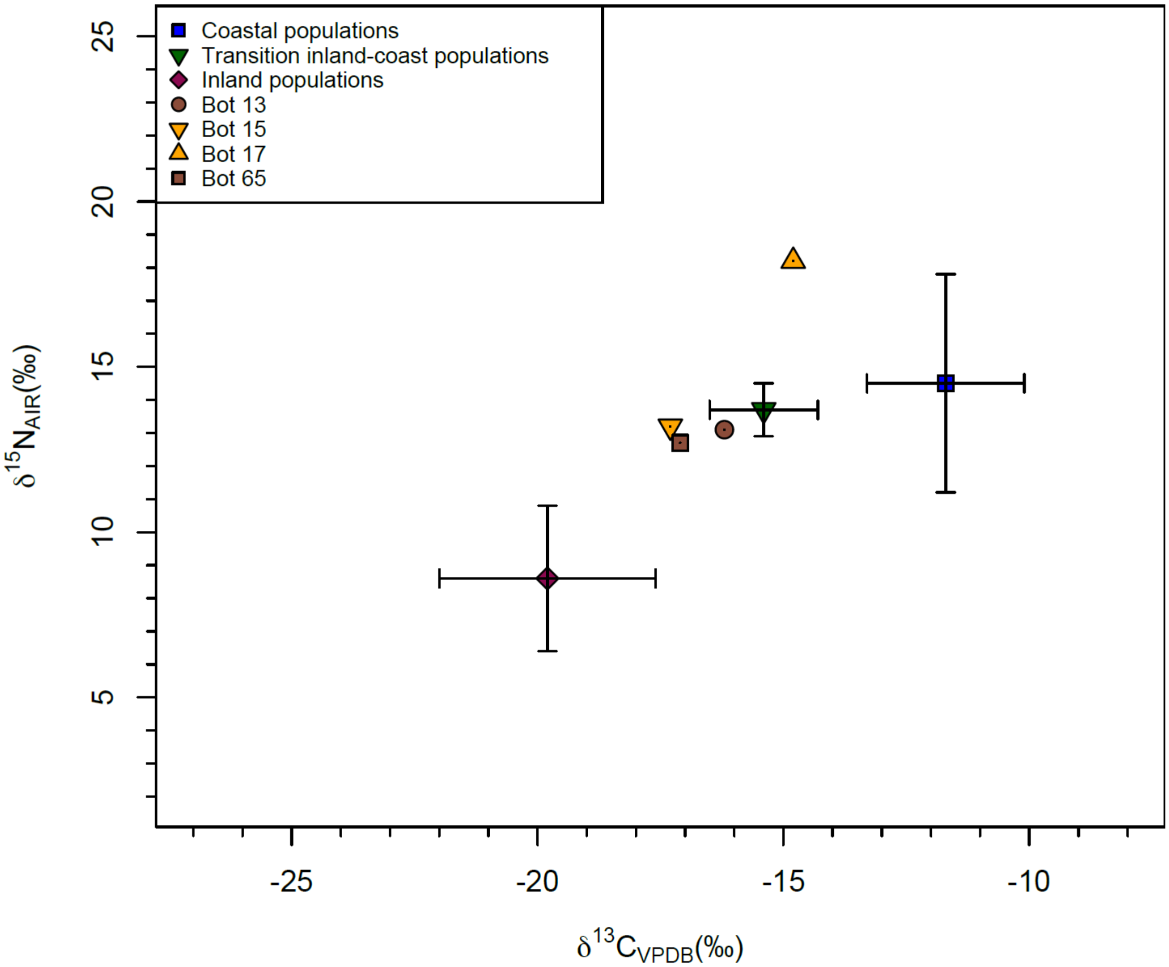
calibrated dates	SHCal04	incl. marine reservoir effect
Bot13	1663 AD (23.5%) 1712 AD 1718 AD (49.9%) 1813 AD 1836 AD (12.9%) 1890 AD 1922 AD (9.1%) 1954 AD	NA
Bot15	1452 AD (66.3%) 1510 AD 1579 AD (29.1%) 1620 AD	1464 AD (84.6%) 1702 AD 1732 AD (10.8%) 1802 AD
Bot17	1419 AD (95.4%) 1477 AD	1495 AD (95.4%) 1833 AD
Bot65	1668 AD (50.8%) 1786 AD 1793 AD (10.1%) 1815 AD 1829 AD (21.3%) 1893 AD 1921 AD (13.1%) 1954 AD	NA

Supplementary Table 8.4 Strontium results for all four individuals included in the study.

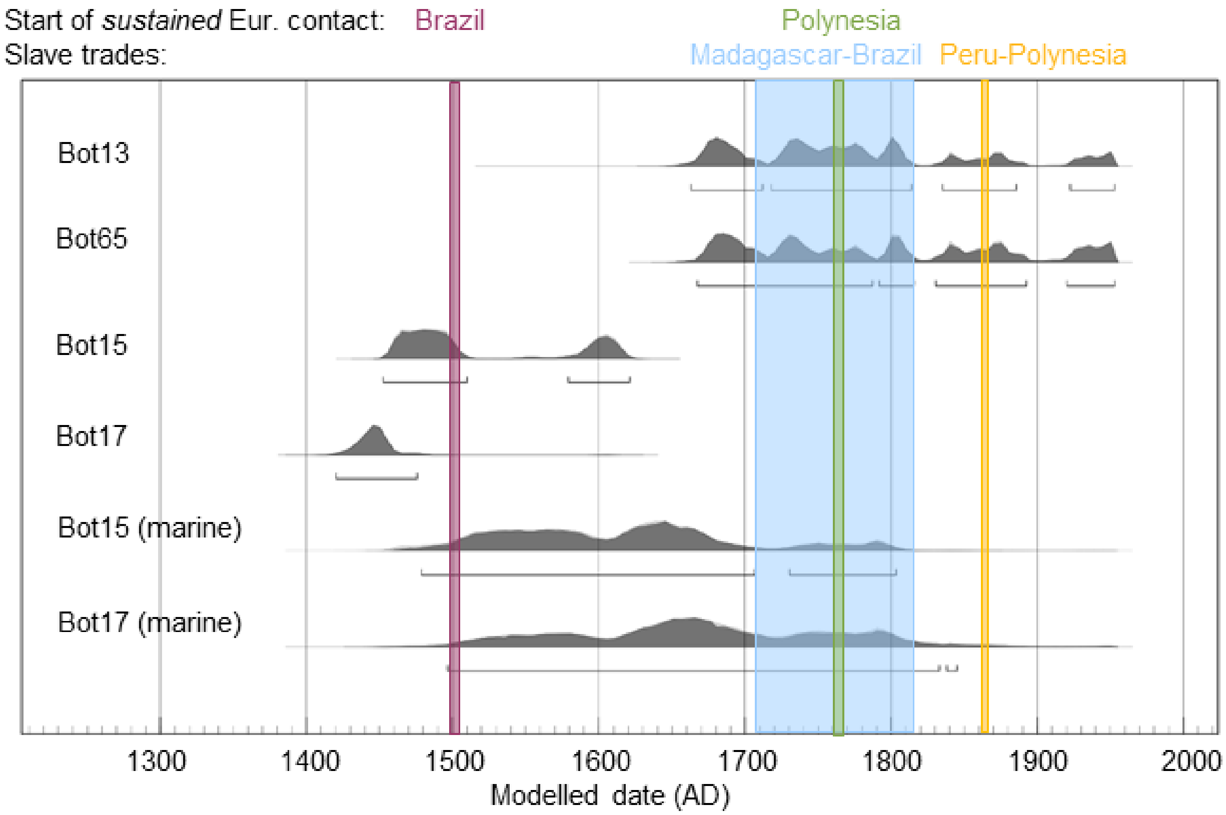
Sample Code	$^{87}Sr/^{86}Sr$	%std err
Bot13	0.727304	0.0008
Bot15	0.708312	0.0008
Bot17	0.708319	0.0009
Bot65	0.727097	0.0006



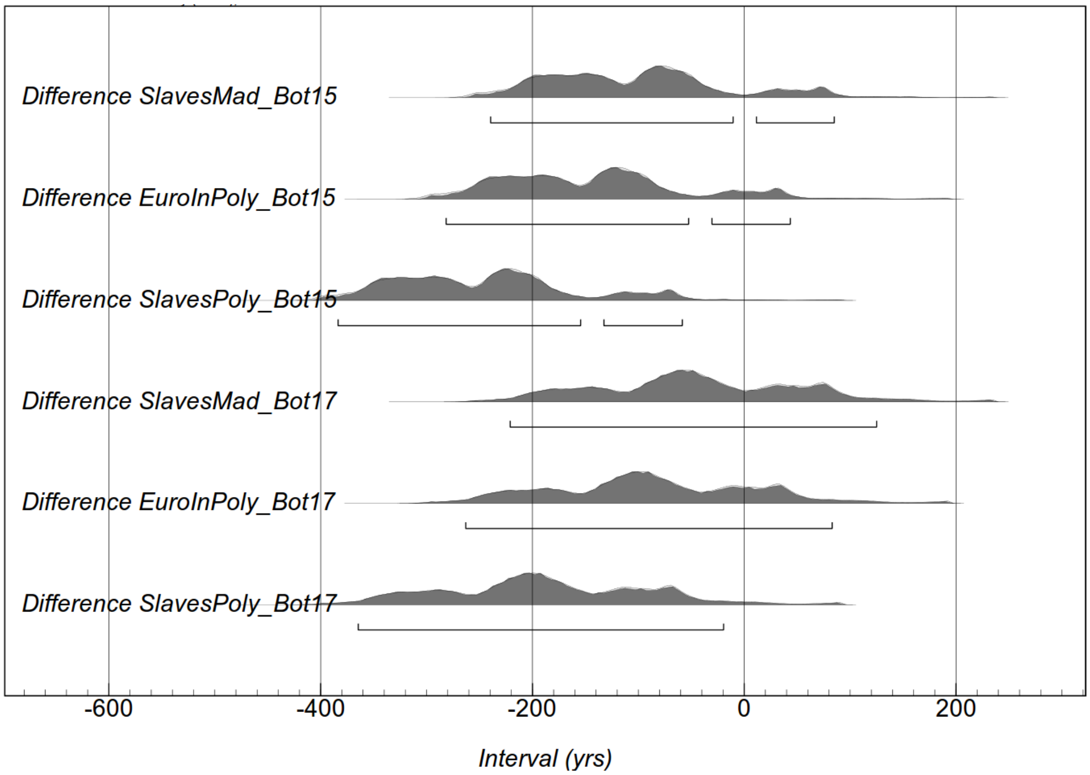
Supplementary Figure 8.1 Values of δ^{13}C and δ^{15}N plotted for modern and archaeological fauna and flora from Brazil together with values from Botocudo individuals.



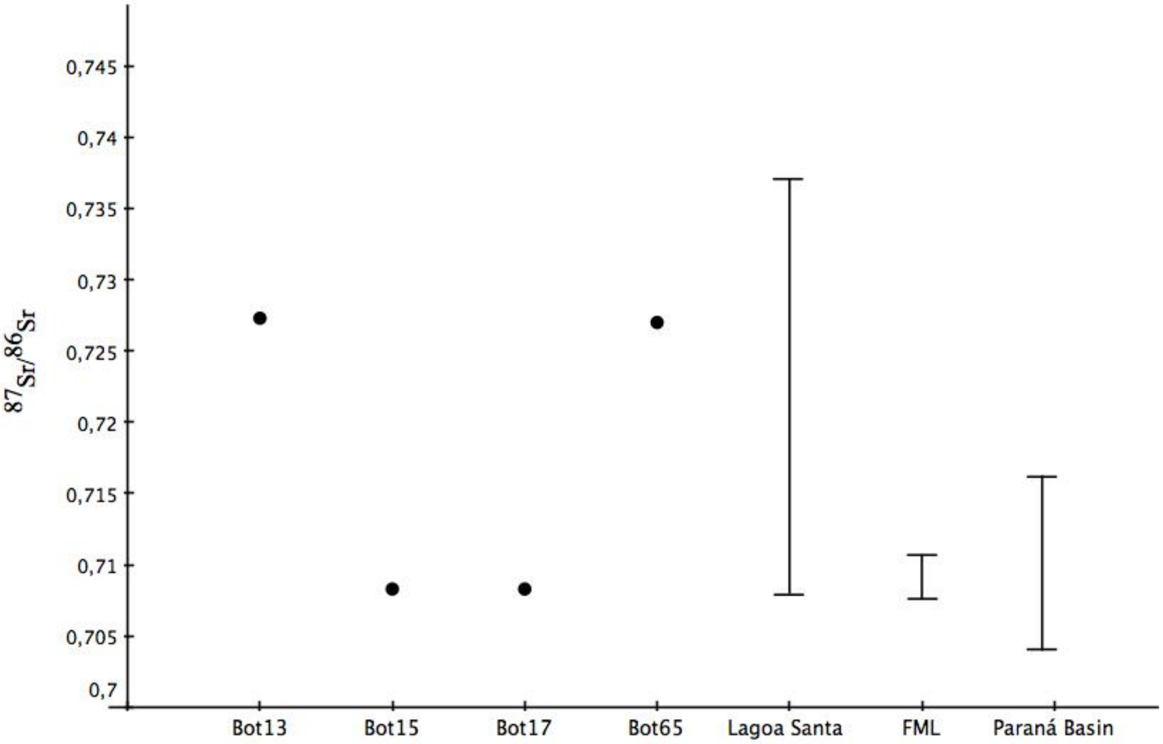
Supplementary Figure 8.2 Values of δ^{13}C and δ^{15}N plotted for archaeological human remains from the coast, interior and inland-coast transition from Brazil together with values from Botocudo individuals.



Supplementary Figure 8.3 Calibrated ^{14}C dates for Bot13, Bot15, Bot17 and Bot65 individuals. For Bot17(marine) and Bot15 (marine), we assumed 60±16% and 30±16% marine carbon dietary uptake respectively, and a reservoir age that is equivalent to the average world ocean reservoir. The colored bars indicate four pertinent historical events: (1) the European discovery of Brazil (purple), (2) the established European contact with Polynesian islands (green), (3) the Madagascar-Brazil slave trade (blue), and (4) the Peru-Polynesia slave trade (orange).



Supplementary Figure 8.4 Probability density function of the difference between the calibrated dates of Bot15 (first three rows) respectively Bot17 (last three rows) and three historical events: the beginning of the Madagascar-Brazil slave trade (row 1 and 4), the beginning of the Madagascar-Brazil slave trade (row 2 and 5) and the beginning of the Peru-Polynesia slave trade (row 3 and 6).



Supplementary Figure 8.5 Values of $^{87}Sr/^{86}Sr$ for Bot13, Bot15, Bot17 and Bot65 individuals compared with values obtained from human tooth samples from the coastal site Forte Marechal Luz (FML, Santa Catarina, Bastos et al. 2011). Local values for Parana Basin (Santa Catarina (Mantovani et al. 1985; Bastos et al. 2011)) and Lagoa Santa (Minas Gerais, Machado 2013)) are also shown for comparison.

Supplementary 9

Craniometry of the Botocudo individuals from the Museu Nacional, Rio de Janeiro, collection

Danilo V. Bernardo*, Walter Neves*

*to whom correspondence should be addressed (danvb@ib.usp.br, waneves@ib.usp.br)

Before any multivariate technique treatment we performed a standardization by means of the indexation of each craniometric variable to the geometric mean of each specimen (Darroch and Mosimann 1985). This procedure, a size correction, was necessary once measurements of skulls are known to be extremely influenced by the size of the individuals (Corruccini 1973). In other words, the dataset was transformed to shape alone data. In order to maximize the sample size we analyzed males and females together.

The quantitative method employed to infer the intrapopulational morphological diversity of the Botocudos, the Principal Components Analysis (PCA), is a multivariate statistic technique of dimension reduction. The basic principle is to generate new axes of variables that explain more of the total variance than the original ones alone. These new axes are the so called “principal components” and their mathematical construction assures that they are orthogonal (uncorrelated) to each other. This method is particularly useful to make hidden patterns explicit. In the particular case of morphometric studies the space delimited by the principal components defines the so called morphospace, generally represented by a bidimensional topology.

The analysis performed by us produced a morphospace composed by the two first Principal Components (PC) that summarize around 37 % of original variance present in the Botocudo samples (see Supplementary Table 9.2 for correlation between the first three PCs and Supplementary Table 9.3 for eigenvalues and variance explained by the PCs). As can be seen in Supplementary Figure 9.1, all Botocudo specimens appear morphologically integrated, without any outliers. Bot15 and Bot17 are perfectly integrated to the other specimens of the sample. It is important to emphasize the virtually perfect cleavage of the sample between male (represented by filled dots, concentrated in the right side of the graph) and female fraction (empty triangles, concentrated in the left side of the graph). This partition is due to the differential correlation among the original variables and each PC utilized to construct the morphospace. In this analysis, high positive values of first PC are associated with high values of GLS (Glabella projection) and vice-versa, while high negative values of the same PC are associated with high values of FMR, NOL, XCB, DKR, EKR, EKB and NAR, and vice-versa (see Supplementary Table 9.1). The second PC was not suitable to morphological interpretation. All this information means that the specimens located in the right side of the graph are represented by skulls with prominent glabelar projections, while the specimens located in the left side of the graph are represented by slightly longer skulls, with marked facial projection.

The second analysis performed was designed to infer the morphological affinities of the Botocudo Indians in a world-wide perspective. In order to do this comparison, we assembled a dataset composed by 2,542 specimens from Howells databank. This sample is divided in 30 series of wide dispersion, representative of the recent global cranial variation. These specimens, represented by 40 craniometric variables, taken in accordance with Howells protocol (1973), are entirely complete, without any occurrence of missing-values. We then added to this dataset 35 individuals (19 males and 16 females) of late Paleoindians from Lagoa Santa, Minas Gerais, Brazil. The Botocudo skulls as well as the Lagoa Santa specimens were both measured by one of us (WN) (see Supplementary Table 9.7 for the measurements for the Lagoa Santa Paleoindians and Supplementary Table 9.8 for the measurements for the Botocudo indians).

Due to their archeological nature, specimens from Lagoa Santa (Paleoindian) sample presented a high percentage of missing-values. To take this into account, the first filtering step was the removal of specimens and variables showing high percentages of missing-values. After this filtering step some individuals exhibiting missing-values still remained in the dataset (0.3% of the total). The replacement of these residual missing values was carried out using the most conservative strategy available in the literature for this purpose: the grand mean calculated over all samples represented in the dataset. The Botocudo Indians sample did not need any replacement procedure (see Supplementary Table 9.4 for details on the craniometric variables considered).

After the screening and replacement of missing information, all specimens were standardized to the shape alone transformation, following the strategy proposed by (Darroch and Mosimann 1985).

To assess the morphological affinities of the specimens of the dataset, we performed a Cluster Analysis, based on Mahalanobis's distances. The principle of this type of analysis is to create several sub-sets or sub-groups within the total database and hierarchically locate these groups in relation to each other based on their degree of resemblance. This can be done using several distinct algorithms. In this study the "Ward method" algorithm was adopted (Ward 1963). Also known as "minimum variance method", this algorithm combines in each step of the amalgamation the two groups that present the minor value for the squared deviation distance to the centroid of each group.

Cluster Analysis is based on a dissimilarity matrix (*i.e.*, a distance matrix). There are several possible dissimilarity matrices, the euclidean distance being the most common among them. However, as said above, in this work Mahalanobis Distance was adopted (Mahalanobis 1936). This particular kind of dissimilarity measurement takes into consideration the correlation among variables and is by far the most popular algorithm in biodistance studies (see Supplementary Table 9.5 for the distances obtained). Another important advantage is that according to principles of quantitative genetics the Mahalanobi's Distances correspond to minimum genetic distances, being more suitable to establish biological affinities among populations (Williams-Blangero and Blangero 1989; Williams-Blangero and Blangero 1990; Relethford and Harpending 1994; Relethford and Blangero 1990; Relethford, Crawford, and Blangero 1997).

As can be seen in Supplementary Figure 9.2 - the topology resulting from the Cluster Analysis - the 32 samples included in the study were grouped, in general, in accordance with their geographic location. The major cleavage observed in this topology opposed a cluster formed for population from Africa (BUS, DOG, TEI, ZUL) and Australo-melanesia (AUS, TAS, TOL) from other geographical groups represented in the analysis. Botocudos (BOT) present a clear association with the Paleoindians from Lagoa Santa (LST), and both are part of a cluster formed primarily by Polynesians (EAS, MOK, MOR, SMA, NMA) in a slightly outgroup cluster.

References

Corruccini, Robert S. 1973. “Size and Shape in Similarity Coefficients Based on Metric Characters.” *American Journal of Physical Anthropology* 38 (3): 743–753. doi:10.1002/ajpa.1330380313.

Darroch, John N., and James E. Mosimann. 1985. “Canonical and Principal Components of Shape.” *Biometrika* 72 (2) (August 1): 241–252. doi:10.2307/2336077.

Howells, William White. 1973. *Cranial Variation in Man: A Study by Multivariate Analysis of Patterns of Difference Among Recent Human Populations*. Harvard University Press.

Mahalanobis, PC. 1936. “On the Generalised Distance in Statistics.” In , 2:49–55. http://ir.isical.ac.in/dspace/handle/1/1268.

Relethford, J H, M H Crawford, and J Blangero. 1997. “Genetic Drift and Gene Flow in Post-famine Ireland.” *Human Biology* 69 (4) (August): 443–465.

Relethford, J. H., and J Blangero. 1990. “Detection of Differential Gene Flow from Patterns of Quantitative Variation.” *Human Biology* 62 (1) (February): 5–25.

Relethford, J. H., and H. C. Harpending. 1994. “Craniometric Variation, Genetic Theory, and Modern Human Origins.” *American Journal of Physical Anthropology* 95 (3): 249–270. doi:10.1002/ajpa.1330950302.

Ward, Joe H. 1963. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association* 58 (301) (March): 236–244. doi:10.1080/01621459.1963.10500845.

Williams-Blangero, S, and J Blangero. 1989. “Anthropometric Variation and the Genetic Structure of the Jirels of Nepal.” *Human Biology* 61 (1) (February): 1–12.

———. 1990. “Effects of Population Structure on Within-group Variation in the Jirels of Nepal.” *Human Biology* 62 (1) (February): 131–146.

Supplementary Table 9.1 – Craniometric variables (in accordance with Howells, 1973) used in the Principal Components Analysis.

Abb.	Description
GOL	Glabello-occipital length
NOL	Nasio-occipital length
BNL	Basion-nasion length
BBH	Basion-bregma height
XCB	Maximum cranial breadth
XFB	Maximum frontal breadth
STB	Bistephanic breadth
ZYB	Bizygomatic breadth
AUB	Biauricular breadth
WCB	Minimum cranial breadth
ASB	Biasterionic breadth
BPL	Basion-prosthion length
NPH	Nasion-prosthion height
NLH	Nasal height
OBH	Orbit height
OBB	Orbit breadth
JUB	Bijugal breadth
NLB	Nasal breadth
MDB	Mastoid breadth
ZMB	Bimaxillary breadth
SSS	Zygomaxillary subtense
FMB	Bifrontal breadth
NAS	Nasio-frontal subtense
EKB	Biorbital breadth
DKS	Dacryon subtense
DKB	Interorbital breadth
WNB	Simotic cord
IML	Malar length, inferior
XML	Malar lenght, maximum
MLS	Malar subtense
WMH	Cheek height
SOS	Supraorbital projection
GLS	Glabella projection
FOL	Foramen magnum length
FRC	Frontal cord
FRS	Frontal subtense
FRF	Nasion-subtense fraction
PAC	Parietal cord
PAS	Parietal subtense
PAF	Bregma-subtense fraction
OCC	Occipital cord
OCS	Occipital subtense
OCF	Lambda-subtense fraction
VRR	Vertex radius
NAR	Nasion radius
SSR	Subspinale radius
PRR	Prosthion radius
DKR	Dacryon radius
ZOR	Zygoorbitale radius
FMR	Frontalmalare radius
EKR	Ectoconchion radius
ZMR	Zygomaxillare radius
BRR	Bregma radius
LAR	Lambda radius
OSR	Ophistion radius
BAR	Basion radius

Supplementary Table 9.2 – Correlation between the first three Principal Components generated (based on covariance matrix) and the 56 original craniometric variables.

Abb.	PC 1	PC 2	PC 3
GOL	-0.62893	-0.11766	-0.02297
NOL	-0.76242	-0.21814	0.11315
BNL	-0.49824	-0.51922	0.19463
BBH	-0.39592	-0.10645	-0.71918
XCB	-0.74155	0.21904	-0.21933
XFB	-0.67022	0.42182	-0.19743
STB	-0.66460	0.49071	-0.05951
ZYB	-0.30355	0.09754	0.35175
AUB	-0.51396	0.30328	0.07224
WCB	-0.45671	0.26774	0.44380
ASB	-0.35658	0.19593	-0.05448
BPL	-0.43564	-0.47803	-0.01315
NPH	-0.28353	-0.59619	-0.06685
NLH	-0.32251	-0.46944	0.36433
OBH	-0.56825	-0.23839	0.10955
OBB	-0.53494	0.11057	0.25813
JUB	-0.52361	0.14055	0.41288
NLB	-0.13163	0.18646	0.17500
MDB	0.50773	0.06949	0.08806
ZMB	-0.50578	0.08561	0.19656
SSS	-0.24865	-0.55339	-0.23727
FMB	-0.58035	0.42927	0.35617
NAS	0.10335	-0.04707	0.37235
EKB	-0.71539	0.39949	0.37633
DKS	0.03603	0.20664	0.39539
DKB	-0.02713	0.22339	0.46017
WNB	0.19495	0.07287	0.05810
IML	0.29742	0.12325	0.26001
XML	-0.13835	-0.19872	0.04065
MLS	0.02423	-0.14423	-0.23643
WMH	0.11705	0.05631	0.21412
SOS	0.46873	0.14584	0.07380
GLS	0.84222	0.04922	-0.13545
FOL	-0.32729	-0.16252	0.18007
FRC	-0.28545	0.24907	-0.68004
FRS	-0.36267	0.19193	-0.44943
FRF	0.23122	0.03969	-0.39113
PAC	-0.35472	0.53942	-0.05859
PAS	-0.12581	0.49925	-0.03124
PAF	-0.43948	0.39398	-0.36307
OCC	-0.49898	-0.52025	-0.32834
OCS	0.12079	-0.36959	-0.00745
OCF	0.26638	-0.14032	-0.04134
VRR	-0.55851	0.15715	-0.53182
NAR	-0.71341	-0.26587	0.39461
SSR	-0.56733	-0.54479	-0.01412
PRR	-0.58855	-0.55327	-0.17259
DKR	-0.72105	-0.23176	0.41531
ZOR	-0.66710	-0.23563	0.22270
FMR	-0.79927	-0.14798	0.30690
EKR	-0.71927	-0.26643	0.27868
ZMR	-0.47795	-0.31431	0.28819
BRR	-0.48858	0.26460	-0.51196
LAR	-0.56189	-0.35349	-0.13150
OSR	-0.26486	-0.25180	0.06389
BAR	0.03139	-0.42471	-0.55029

Supplementary Table 9.3 – Non-zero eigenvalues of each Principal Component extracted from covariance matrix from 56 craniometric variables and related statistics.

PC	Eigenvalue	% Total – variance	Cumulative - Eigenvalue	Cumulative - %
1	0.05958	25.32250	0.05958	25.32250
2	0.02817	11.97291	0.08776	37.29540
3	0.02189	9.30269	0.10965	46.59810
4	0.01917	8.14639	0.12881	54.74450
5	0.01571	6.67646	0.14452	61.42100
6	0.01445	6.14146	0.15897	67.56240
7	0.01119	4.75472	0.17016	72.31710
8	0.00948	4.02683	0.17964	76.34400
9	0.00829	3.52491	0.18793	79.86890
10	0.00673	2.85817	0.19466	82.72700
11	0.00660	2.80594	0.20126	85.53300
12	0.00534	2.26927	0.20660	87.80220
13	0.00439	1.86348	0.21098	89.66570
14	0.00385	1.63541	0.21483	91.30110
15	0.00319	1.35508	0.21802	92.65620
16	0.00310	1.31815	0.22112	93.97440
17	0.00270	1.14859	0.22382	95.12300
18	0.00240	1.01947	0.22622	96.14240
19	0.00181	0.76910	0.22803	96.91150
20	0.00153	0.65039	0.22956	97.56190
21	0.00148	0.62811	0.23104	98.19000
22	0.00111	0.47089	0.23215	98.66090
23	0.00086	0.36686	0.23301	99.02780
24	0.00062	0.26425	0.23363	99.29200
25	0.00059	0.24903	0.23422	99.54110
26	0.00047	0.19754	0.23469	99.73860
27	0.00034	0.14591	0.23503	99.88450
28	0.00027	0.11549	0.23530	100.00000

Supplementary Table 9.4 – Craniometric variables (in accordance with Howells, 1973) represented in the dataset used to perform the Mahalanobis Distance Matrix.

Abb.	Description
GOL	Glabello-occipital length
NOL	Nasio-occipital length
BNL	Basion-nasion length
BBH	Basion-bregma height
XCB	Maximum cranial breadth
XFB	Maximum frontal breadth
STB	Bistephanic breadth
ZYB	Bizygomatic breadth
AUB	Biauricular breadth
WCB	Minimum cranial breadth
ASB	Biasterionic breadth
BPL	Basion-prosthion length
NPH	Nasion-prosthion height
NLH	Nasal height
OBH	Orbit height
OBB	Orbit breadth
NLB	Nasal breadth
ZMB	Bimaxillary breadth
SSS	Zygomaxillary subtense
FMB	Bifrontal breadth
NAS	Nasio-frontal subtense
EKB	Biorbital breadth
DKS	Dacryon subtense
DKB	Interorbital breadth
WMH	Cheek height
SOS	Supraorbital projection
FRC	Frontal cord
FRF	Nasion-subtense fraction
PAC	Parietal cord
PAF	Bregma-subtense fraction
OCC	Occipital cord
OCF	Lambda-subtense fraction
VRR	Vertex radius
NAR	Nasion radius
SSR	Subspinale radius
PRR	Prosthion radius
ZOR	Zygoorbitale radius
FMR	Frontalmalare radius
EKR	Ectoconchion radius
ZMR	Zygomaxillare radius

Supplementary Table 9.5 – Mahalanobis Distance Matrix of 32 populations, performed over 40 craniometric variables (in accordance with Howells, 1973).

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.	20.	21.	22.	23.	24.	25.	26.	27.	28.	29.	30.	31.	32.
1.BUS	0.000	25.140	19.712	16.363	43.873	39.275	35.977	54.092	53.094	24.048	34.875	29.788	30.928	26.817	33.552	33.595	47.909	31.468	25.757	22.883	24.040	33.210	36.986	30.148	34.980	28.389	40.674	50.894	42.865	46.878	44.080	58.194
2.DOG		0.000	15.082	12.210	43.873	34.938	42.506	61.437	60.436	20.090	22.320	26.723	22.702	38.275	41.447	30.476	43.583	40.242	22.611	30.286	32.465	20.260	29.458	23.674	25.814	20.059	44.039	46.006	36.281	36.881	44.075	52.664
3.TEI			0.000	11.547	31.057	30.778	33.236	57.433	37.872	19.254	24.642	27.133	20.937	23.579	35.958	22.064	38.222	32.966	17.796	18.403	23.801	25.389	27.737	23.259	23.470	21.242	26.865	33.633	35.417	31.586	31.522	46.650
4.ZUL				0.000	33.517	26.902	33.482	62.480	50.165	15.697	22.559	22.392	19.785	23.104	26.583	19.412	39.364	30.270	18.181	20.429	17.768	24.801	27.569	20.205	24.577	17.661	32.845	36.082	29.384	32.784	35.138	47.791
5.ARI					0.000	12.591	15.649	27.503	32.553	18.817	16.723	18.239	16.950	39.588	33.354	26.970	40.140	18.645	26.022	20.949	17.064	24.894	18.634	20.773	17.540	16.459	41.353	36.955	21.208	12.970	18.809	21.165
6.PER						0.000	13.292	35.947	38.079	16.543	15.716	17.211	15.591	39.347	36.066	27.868	39.943	19.936	23.194	18.313	16.432	21.188	19.075	18.405	18.059	17.889	32.931	38.363	20.367	16.836	22.376	23.334
7.SCR							0.000	40.769	36.578	20.621	25.256	20.119	21.906	29.484	33.592	26.677	45.814	21.120	25.983	19.924	19.245	21.042	30.760	26.185	23.799	20.136	36.848	49.769	31.360	22.111	28.764	30.806
8.BUR								0.000	43.923	38.363	28.737	28.728	33.082	74.638	68.866	61.855	62.121	24.544	49.803	38.210	35.896	51.631	32.524	38.460	35.977	32.492	68.229	66.970	43.451	42.919	55.736	49.813
9.ESK									0.000	28.977	33.286	28.427	27.189	41.559	56.931	34.525	44.690	46.334	40.789	34.103	34.853	48.620	33.753	35.873	28.864	40.455	36.095	38.092	33.264	32.579	30.629	37.975
10.AIN										0.000	13.972	12.552	10.999	25.997	27.720	20.991	37.636	20.188	14.385	11.761	10.179	17.575	16.449	12.805	14.918	14.847	33.353	28.076	19.969	18.300	23.342	27.308
11.HAI											0.000	7.076	4.184	46.486	39.539	25.991	29.802	24.269	21.495	22.680	18.151	16.059	4.048	9.070	7.028	5.711	29.334	28.150	15.906	20.022	28.055	27.168
12.NJA												0.000	3.059	40.627	37.472	24.445	29.967	21.787	19.770	20.454	15.890	20.296	9.272	9.014	12.384	9.809	29.814	27.921	18.439	17.953	24.135	26.603
13.SJA													0.000	40.990	38.208	22.057	26.892	23.409	17.885	19.359	16.104	16.572	6.906	6.480	8.339	8.098	25.489	22.950	15.509	17.031	21.967	26.481
14.AUS														0.000	16.123	16.074	55.690	39.767	33.237	28.734	25.902	39.326	49.565	36.624	36.843	33.296	40.640	48.060	45.772	39.803	32.205	43.787
15.TAS															0.000	12.148	51.107	31.439	36.168	32.094	24.733	34.059	40.643	32.619	37.047	28.519	43.321	36.149	31.201	29.894	25.946	33.573
16.TOL																0.000	34.975	34.570	32.968	29.130	25.245	27.897	27.887	24.420	24.623	21.308	26.405	27.230	23.824	22.764	18.988	29.704
17.BBC																	0.000	41.940	42.510	39.404	39.499	41.057	30.785	30.476	27.040	35.717	15.420	32.294	28.086	32.458	38.388	34.431
18.BER																		0.000	20.349	11.793	8.243	27.438	31.536	23.064	28.838	20.529	41.150	47.450	32.046	28.476	35.729	36.435
19.EGY																			0.000	6.167	9.714	18.315	27.800	25.121	23.649	20.131	38.128	40.259	31.186	22.847	33.588	39.482
20.NOR																				0.000	5.855	25.098	27.569	25.145	22.432	22.401	33.587	39.091	30.152	20.340	29.822	34.679
21.ZAL																					0.000	22.805	22.694	17.110	19.366	15.253	34.021	37.445	25.686	24.494	30.058	33.055
22.AND																						0.000	25.830	20.559	21.538	14.075	38.475	38.571	22.069	24.419	33.304	35.028
23.ANY																							0.000	10.905	8.938	9.407	31.157	25.088	17.756	22.020	25.303	27.633
24.ATA																								0.000	12.127	8.710	25.428	27.010	22.517	25.485	25.409	28.237
25.GUA																									0.000	8.313	26.394	25.206	13.908	18.616	22.006	21.802
26.PHI																										0.000	30.776	33.522	19.504	24.212	27.965	30.656
27.LST																											0.000	28.383	30.916	35.914	34.177	35.764
28.EAS																												0.000	13.701	20.886	17.833	21.576
29.MOK																													0.000	12.841	16.631	17.181
30.MOR																														0.000	10.851	10.557
31.NMA																															0.000	12.776
32.SMA																																0.000

Supplementary Table 9.6 – Name, abbreviation and size of sample of each population of the dataset.

Series	Abb	N. Masc.	N. Fem.
BUSHMAN	BUS	41	49
DOGON	DOG	47	52
TEITA	TEI	33	50
ZULU	ZUL	55	46
ARIKARA	ARI	42	27
PERU	PER	55	55
SANTA CR	SCR	51	51
BURIAT	BUR	55	54
ESKIMO	ESK	53	55
AINU	AIN	48	38
ANDAMAN	AND	11	6
HAINAN	HAI	45	38
NORTHERN JAPAN	NJA	55	32
SOUTHERN JAPAN	SJA	50	41
AUSTRALIA	AUS	52	49
TASMANIA	TAS	44	42
TOLAI	TOL	56	54
BOTOCUDO FROM CENTRAL BRAZIL	BOT	15	14
BERG	BER	56	53
EGYPT	EGY	58	53
NORSE	NOR	55	55
ZALAVAR	ZAL	53	44
ANDAMAN	AND	24	29
ANYANG	ANY	42	-
ATAYAL	ATA	29	18
GUAM	GUA	30	27
PHILLIPINES	PHI	50	-
PALEOINDIAN FROM CENTRAL BRAZIL (LAGOA SANTA)	LST	19	16
EASTER ISLAND	EAS	49	37
MOKAPU	MOK	51	49
MORIORI	MOR	57	51
S MAORI	SMA	20	20
		1401	1205

Supplementary Table 9.7 – Craniometric measurements of 35 specimens from Lagoa Santa, taken in accordance with Howells protocol (1973), used in the dataset.

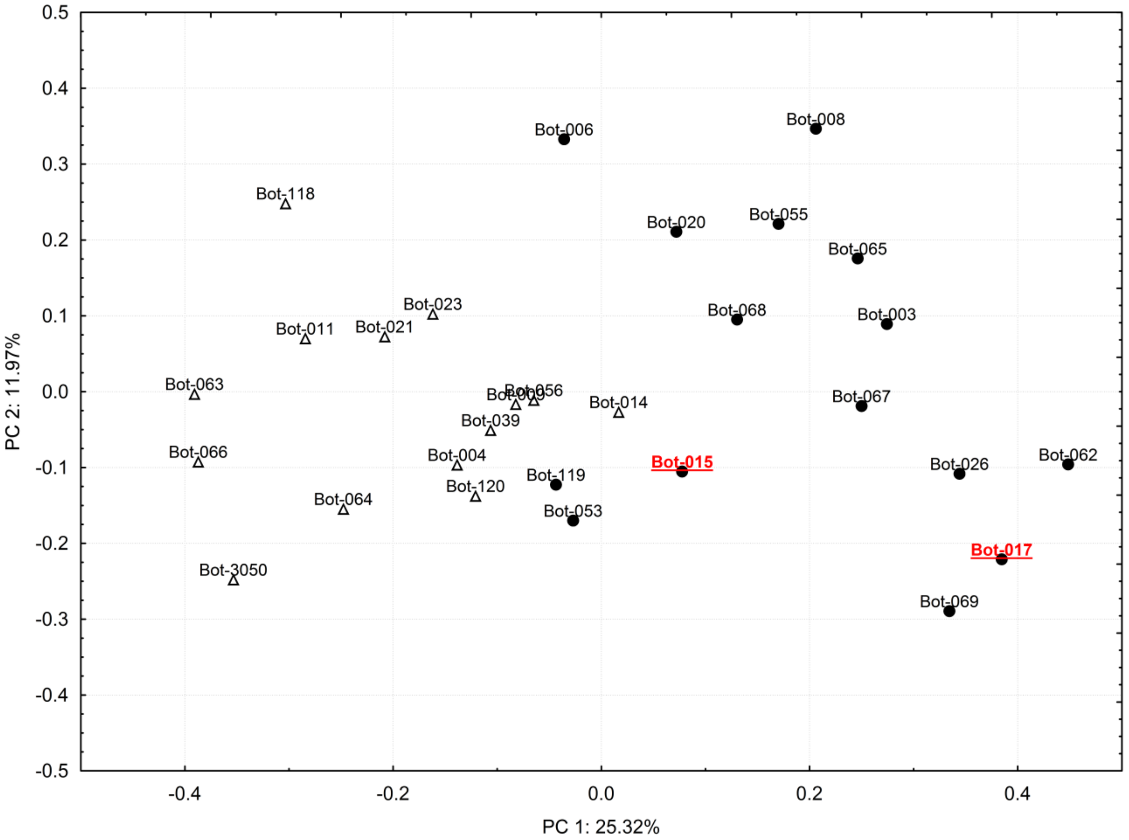
Serie	#id	Archeolog. site	Sex	GOL	NOL	BNL	BBH	XCB	XFB	STB	ZYB	AUB	WCB	ASB	BPL	NPH	NLH	OBH	OBB	NLB	ZMB	SSS	FMB	NAS	EKB	DKS	DKB	WMH	SOS	FRC	FRF	PAC	PAF	OCC	OCF	VRR	NAR	SSR	PRR	ZOR	FMR	EKR	ZMR
LST	CONFINS	-	M	179	178	-	-	126	112	112	-	120	79	110	-	76	54	37	36	31	-	-	-	-	-	-	24	24	6	110	45	109	57	-	-	125	95	94	101	79	-	76	72
LST	HW 009	-	M	186	181	-	-	128	105	-	127	120	68	113	-	56	47	34	39	26	98	21	99	17	98	12	22	26	7	-	-	-	-	-	-	131	99	108	112	88	81	78	85
LST	HW 010	-	M	183	180	104	139	125	108	103	133	124	75	102	97	72	56	30	39	26	96	21	98	16	94	7	20	25,5	7	110	50	109	56	102	50,5	125	96	93	104	84	81	77	80
LST	HW AN14	-	M	183	181	-	-	126	107	-	-	123	70	-	-	60	43	33	39	25	97	17	95	11	95	6	24	24	6	114	45	122	-	-	-	132	89	92	101	82	79	73	75
LST	MN 1355	Cerca Grande 6	M	185	183	-	-	127	105	102	-	119	69	112	-	-	-	33	39	-	-	-	100	19	100	-	26	23	5	111	46	111	57	98	34	121	94	-	-	82	79	74	77
LST	MN 1357	Cerca Grande 6	M	184	180	-	-	124	106	97	128	118	65	101	-	63	49	32	40	24	-	-	102	15	-	-	23	23	5	112	50	127	67	92	49	128	95	99	107	84	81	77	79
LST	MN 629	Lapa de Carrancas	M	182	178	-	-	134	118	-	-	130	-	115	-	-	-	33	40	23	-	-	100	13	99	6	25	-	7	109	50	111	55	93	43	123	95	-	-	86	84	79	-
LST	MN 630	Lapa de Carrancas	M	187	182	101	136	137	116	112	134	123	72	114	99	60	43	29	40	24	102	18	102	16	101	8	28	26	5	113	-	122	67	99	47	127	95	93	100	86	81	77	77
LST	MN 804	Lapa do Caetano	M	187	185	105	128	132	114	111	-	132	71	113	-	-	53	36	41	26	-	-	108	18	-	-	24	26	7	108	44	112	61	-	-	123	101	99	-	87	83	78	-
LST	MN 805	Lapa Mortuária	M	176	171	-	-	120	103	102	126	114	71	101	-	59	44	32	39	21	104	19	100	14	93	8	22	27	7	106	42	111	54	90	41	120	92	93	97	81	81	75	76
LST	MN 807	Lapa Mortuária	M	182	178	100	137	132	110	104	-	123	72	110	96	61	48	33	39	26	-	-	100	15	-	-	25	23	7	104	42	121	69	94	40	126	94	100	104	87	81	74	78
LST	SR1-1	Santana do Riacho 1	M	196	190	108	134	125	108	108	134	122	73	112	103	72	55	35	45	27	106	-	106	19	100	-	26	27	9	114	47	118	60	97	52	-	-	-	-	-	-	-	-
LST	SH-02	Sumidouro	M	185	180	-	-	128	113	110	-	121	73	114	-	63	47	38	39	-	-	-	-	-	-	-	26	23	7	110	47	109	48	-	-	121	92	102	108	85	81	77	79
LST	SH-03	Sumidouro	M	191	186	98	134	132	110	110	135	118	69	111	96	67	45	31	40	22	95	23	98	14	96	8	21	24	8	110	49	115	64	110	64	121	93	98	101	83	78	74	76
LST	SH-04	Sumidouro	M	182	179	102	137	128	114	114	140	126	73	113	-	-	50	33	40	24	103	22	103	17	100	7	27	27	10	111	52	114	62	97	44	127	96	97	-	85	82	78	79
LST	SH-05	Sumidouro	M	183	180	93	132	129	110	109	127	120	71	112	91	57	45	32	39	24	95	17	97	13	98	9	25	22	7	111	46	122	60	95	45	124	91	93	98	81	79	74	76
LST	SH-09	Sumidouro	M	182	180	98	139	129	114	114	134	119	73	113	93	65	47	35	41	25	103	24	103	15	101	10	26	23	9	112	48	110	64	104	49	128	91	94	100	80	79	72	72
LST	SH-11	Sumidouro	M	197	191	110	141	140	117	117	-	128	79	113	-	-	-	-	-	-	-	-	101	19	-	-	28	-	7	124	56	108	54	107	53	131	104	-	-	-	87	-	-
LST	SH-16	Sumidouro	M	180	177	102	139	131	113	111	134	123	72	108	95	61	49	33	40	25	100	22	101	15	99	8	26	23	7	110	46	117	62	94	51	130	96	98	100	83	82	76	78
LST	HW 001	-	F	181	178	95,5	138	132,5	117	117	130	117	68	105	89,5	56	46	31	37,5	26,5	99,5	21	99,5	13	96	6	23	19	7	112	46,5	119	60	99	43	125	91	91,5	94	80	79	75,5	71
LST	HW 004	-	F	179	175	-	-	-	105	109	127	116	65,5	104	-	61	45	33	38	24	103	23	97	19	94	12	24	25	3	109	49	115	66	102	53	127	88,5	94	99,5	76	77,5	69	71
LST	HW 005	-	F	178	174	97	129	130	105	98	124	104	74	105	98	55	43	31	37	24	94	14	92	10	93	5	21	20	5	105	43	108	-	101	43	121	88	86	96	78	80	75	74
LST	HW 006	-	F	183	180	99	138	124	108	108	-	-	-	-	93	59	49	30	36	24	-	-	94	14	93	13	22	24	5	102	45	118	62	-	-	-	-	-	-	-	-	-	-
LST	HW S/N	Centro Lagoa Santa	F	176	174	-	-	133	115	114	130	120	67	105	-	60	46	32	37	25	96	20	95	14	94	9	24	21	6	110	50	120	65	-	-	126	90	90	91	77	77	72	72
LST	MN 1325	Cerca Grande 6	F	183	177	-	-	129	104	-	-	107	-	97	-	52	41	34	-	24	-	-	-	-	-	-	-	18	5	107	47	124	70	94	48	123	95	94	100	-	-	-	-
LST	MN 1353	Cerca Grande 6	F	181	178	96	127	134	111	105	-	122	65	103	-	-	-	33	39	-	90	-	97	9	-	-	20	23	7	108	42	105	52	105	49	124	90	-	-	82	84	-	75
LST	MN 1388	Cerca Grande 7	F	174	170	100	131	130	108	107	-	117	71	104	-	-	44	34	38	-	92	17	96	15	96	12	20	22	8	108	44	111	59	92	45	124	94	90	-	-	78	71	72
LST	MN 806	Lapa Mortuária	F	171	168	94	126	127	106	103	-	110	65	102	-	-	-	-	-	-	-	-	94	13	-	-	-	-	4	102	42	113	56	88	37	123	89	-	-	-	76	-	-
LST	MN 1959	Lapa Vermelha IV	F	185	181	102	133	126	108	103	123	118	73	106	94	61	46,5	33	39	26	-	-	100	14	-	-	-	23	5	109	45	112	65	93	42	120	92	93	100	89	83	77	77
LST	SR1-3	Santana do Riacho 1	F	177	174	93	123	122	101	98	-	114	61	102	-	-	-	-	-	26	-	-	95	14	-	-	-	-	5	104	45	111	62	93	37	114	92	-	-	-	83	-	-
LST	SR1-6	Santana do Riacho 1	F	171	169	-	-	118	-	-	120	-	-	105	-	54	42	30	34	20	-	-	-	-	-	-	-	-	5	106	45	115	59	90	41	-	-	-	-	-	-	-	-
LST	SH-07	Sumidouro	F	176	172	-	-	123	105	105	-	113	70	106	-	63	47	35	36	25	-	-	-	-	-	-	-	24	5	104	51	107	64	-	-	114	90	90	94	80	79	73	74
LST	SH-08	Sumidouro	F	185	180	97	136	131	116	113	-	121	71	113	-	-	-	34	39	27	-	-	-	-	-	-	26	24	6	113	51	113	61	101	51	127	94	-	-	-	85	-	-
LST	SH-10	Sumidouro	F	179	177	-	-	132	110	108	-	127	80	111	-	64	47	33	39	25	-	-	100	13	-	-	25	-	6	115	56	124	63	-	-	131	94	99	105	84	77	72	79
LST	SH-15	Sumidouro	F	172	169	92	130	124	107	107	-	115	69	94	-	-	-	-	-	-	-	-	94	11	-	-	22	-	6	106	49	116	60	93	47	122	89	-	-	-	81	-	-

Supplementary Table 9.8 – Craniometric measurements of 29 specimens of Botocudo from Central Brazil, taken in accordance with Howells protocol (1973), used in the dataset.

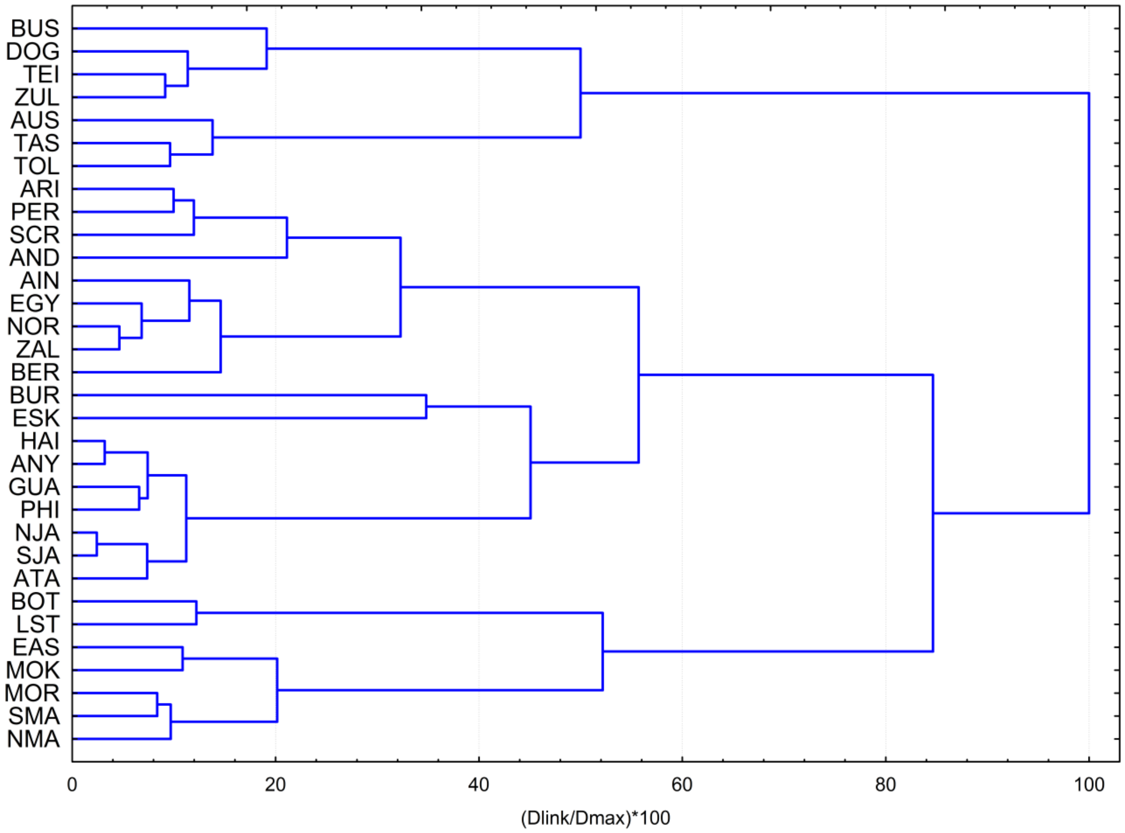
Serie	BBC	BBC	BBC	BBC	BBC	BBC	BBC	BBC	BBC	BBC
#id	MN-068	MN-055	MN-065	MN-069	MN-067	MN-053	MN-008	MN-015	MN-017	MN-119
Region	Botocudo - ES - Aldeiamento Mutum	Botocudo - ES - Cachoeira do Itapeirim, Caverna de Penedias	Botocudo - ES - Cachoeira do Itapemirim	Botocudo - ES - Rio Mucuri	Botocudo – MG	Botocudo - MG - Fazenda Santanna / Rio Novo Caverna da Babilônia	Botocudo - MG - Rio Doce	Botocudo - MG - Rio Doce	Botocudo - MG - Rio Doce	Botocudo - MG - Rio Doce
Sex	M	M	M	M	M	M	M	M	M	M
GOL	181	176	184	182	183	179	188	196	190	183
NOL	177	173	179	178	180	177	182	190	187	177
BNL	97	101	101	109	104	103	102	106	105	101
BBH	128	135	143	143	140	140	146	141	138	144
XCB	134	142	131	133	139	138	138	137	133	132
XFB	110	118	113	112	113	115	122	112	108	108
STB	108	116	113	112	112	113	120	111	108	108
ZYB	133	135	132	134	135	136	146	142	142	130
AUB	122	126	124	121	127	123	129	131	128	119
WCB	69	73	78	67	71	65	75	72	77	65
ASB	108	111	103	102	108	112	113	110	111	106
BPL	94	100	95	103	96	96	95	101	103	102
NPH	68	71	65	75	66	68	67	75	74	69
NLH	52	53	48	55	52	50	50	57	55	51
OBH	34	34	31	33	35	37	36	38	34	33
OBB	42	42	41	43	40	43	44	45	41	39
JUB	115	120	117	119	118	119	128	121	125	112
NLB	25	29	24	24	25	25	27	24	27	25
MDB	10	13	15	11	13	13	14	15	18	13
ZMB	97	103	100	98	99	96	107	96	104	100
SSS	21	26	19	25	24	22	22	27	25	26
FMB	100	103	103	101	101	97	112	108	103	98
NAS	17	15	17	16	15	13	15	17	17	11
EKB	98	103	100	100	100	99	110	104	103	96
DKS	12	11	11	11	9	11	11	9	9	4
DKB	23	28	25	24	24	24	25	25	26	21
WNB	9	10	5	8	9	8	6	5	7	5
IML	42	34	36	40	43	34	41	39	40	33
XML	52	50	53	60	62	53	57	55	58	49
MLS	10	8	11	10	15	8	8	11	11	11
WMH	26	26	24	27	24	18	29	21	32	27
SOS	6	7	6	5	7	7	9	6	9	7
GLS	3	4	5	6	4	3	7	7	6	5
FOL	38	35	41	39	39	39	35	37	38	34
FRC	107	115	116	114	115	112	120	123	113	115
FRS	24	24	28	24	22	26	26	27	22	27
FRF	46	54	51	45	47	45	50	49	52	53
PAC	113	113	121	106	113	100	128	124	114	113
PAS	23	23	29	24	22	14	27	27	23	21
PAF	59	61	67	54	59	50	64	58	51	63
OCC	85	87	94	94	93	102	97	105	102	97
OCS	29	23	24	28	28	30	25	26	30	27
OCF	40	42	55	51	47	58	54	52	48	37
VRR	116	123	128	126	124	123	133	135	129	125
NAR	95	93	98	97	97	94	97	103	99	88
SSR	93	96	96	103	100	93	98	103	103	98
PRR	99	103	104	109	103	101	101	107	107	104
DKR	85	81	85	87	88	82	85	90	88	77
ZOR	83	78	84	89	87	80	85	88	87	80
FMR	82	76	79	81	85	81	85	84	85	77
EKR	76	70	74	77	79	71	77	81	79	72
ZMR	71	69	76	80	77	73	76	78	79	73
BRR	116	121	126	121	124	123	131	131	125	123
LAR	99	104	101	105	106	108	107	114	115	108
OSR	43	44	47	46	43	45	43	45	45	46
BAR	13	14	19	24	18	17	16	12	14	24

Serie	BBC	BBC	BBC	BBC	BBC	BBC	BBC	BBC	BBC	BBC
#id	MN-062	MN-026	MN-003	MN-006	MN-020	MN-014	MN-021	MN-066	MAE-3050	MN-004
Region	Botocudo - MG - Rio Mucuri (Itambacuri)	Botocudo Bahia = Nack-Nanuk	Botocudo ES - Poaia, Mutum	Botocudo -ES - São Mateus	Botocudos - ES - Rio Mucuri	Botocudo - ES - Mutum	Botocudo - ES - Mutum	Botocudo - ES - Aldeiamento Mutum	Botocudo - Gutucracy (ES)	Botocudo - MG
Sex	M	M	M	M	M	F	F	F	F	F
GOL	185	187	178	190	185	172	172	173	168	172
NOL	180	180	174	187	179	170	170	171	167	168
BNL	112	108	98	105	106	100	93	104	96	98
BBH	153	140	132	140	144	129	134	128	135	133
XCB	135	135	135	141	139	130	134	128	129	133
XFB	120	117	110	128	122	108	109	112	111	111
STB	115	102	110	127	121	107	107	110	104	110
ZYB	142	133	139	137	141	130	127	132	122	125
AUB	129	124	125	127	127	120	121	117	113	117
WCB	71	68	70	73	75	72	69	80	63	67
ASB	110	112	103	112	105	102	99	97	94	101
BPL	103	105	100	96	100	91	87	101	95	95
NPH	75	74	69	70	65	75	61	66	78	71
NLH	54	51	51	53	52	54	48	52	49	50
OBH	33	35	34	35	31	36	32	34	33	35
OBB	44	43	40	45	42	41	40	41	36	40
JUB	118	117	120	122	125	112	111	115	110	108
NLB	25	26	23	23	25	23	23	25	21	22
MDB	17	9	18	12	13	11	11	14	8	10
ZMB	106	102	104	101	102	92	92	101	94	91
SSS	26	26	23	17	24	20	24	23	24	24
FMB	102	102	100	108	103	96	94	100	93	93
NAS	18	108	15	14	14	14	12	12	11	14
EKB	101	98	98	108	103	97	94	101	92	94
DKS	11	12	12	10	12	9	9	9	7	10
DKB	22	26	23	26	25	22	19	25	20	23
WNB	11	11	9	10	11	8	7	7	8	7
IML	35	37	34	40	38	34	35	32	33	34
XML	55	55	48	57	60	51	49	48	56	50
MLS	10	11	9	13	7	10	8	8	13	9
WMH	25	22	27	27	31	25	22	23	25	22
SOS	7	11	6	8	7	5	6	6	4	5
GLS	5	6	5	3	4	3	2	1	1	2
FOL	38	35	40	41	37	36	38	35	36	35
FRC	124	116	113	116	121	105	113	103	106	105
FRS	29	27	24	25	29	24	26	24	24	23
FRF	50	50	54	45	42	41	42	39	46	40
PAC	111	116	108	123	118	108	106	106	106	103
PAS	23	26	23	28	28	25	22	21	23	21
PAF	61	63	62	69	61	58	56	57	63	57
OCC	96	97	89	92	93	91	94	90	94	94
OCS	29	26	27	27	25	23	24	24	24	27
OCF	53	43	45	40	45	51	45	42	39	42
VRR	131	126	129	130	128	117	122	117	118	120
NAR	99	98	92	96	97	95	91	97	91	91
SSR	101	99	95	96	100	89	92	97	95	90
PRR	109	107	102	101	102	96	96	103	104	99
DKR	86	84	81	86	88	82	81	86	83	78
ZOR	85	82	80	86	88	79	78	85	84	78
FMR	82	82	78	83	84	82	80	86	80	77
EKR	76	75	72	76	79	72	72	77	75	71
ZMR	76	75	74	79	74	71	63	77	72	69
BRR	128	123	128	126	127	115	119	114	116	117
LAR	105	106	104	112	104	99	104	101	100	101
OSR	48	45	45	47	48	42	41	41	41	43
BAR	24	18	12	15	13	13	16	14	20	18

Serie	BBC	BBC	BBC	BBC	BBC	BBC	BBC	BBC	BBC
#id	MN-011	MN-064	MN-120	MN-118	MN-056	MN-039	MN-063	MN-009	MN-023
Region	Botocudo - MG - Rio Doce	Botocudo - MG - Rio Doce	Botocudo - MG - Rio Doce	Botocudo - MG - Rio Doce (1882)	Botocudo - MG - Rio Mucuri (Itambacuri)	Botocudo ES - Aldesmento Poaiae Mutum	Botocudo -ES - São Mateus	Botocudos - ES - Rio Mucuri	Índio Poxixa e Rio Mucuri
Sex	F	F	F	F	F	F	F	F	F
GOL	165	182	171	170	175	182	174	167	168
NOL	163	181	169	165	171	179	173	165	164
BNL	92	107	100	97	100	102	101	100	93
BBH	126	136	126	132	130	130	133	132	130
XCB	126	134	123	131	125	134	133	130	131
XFB	105	117	103	114	107	115	109	110	111
STB	102	116	102	113	105	111	109	110	110
ZYB	126	136	124	125	124	125	130	133	116
AUB	115	122	111	118	115	120	120	117	110
WCB	61	75	63	69	69	70	71	71	70
ASB	100	107	102	104	102	102	103	101	99
BPL	90	99	92	93	97	103	94	100	91
NPH	67	76	63	60	64	71	67	62	62
NLH	49	55	47	46	49	51	49	49	43
OBH	32	35	32	32	32	34	34	31	31
OBB	39	43	41	40	41	41	42	39	39
JUB	110	119	109	110	111	111	130	113	105
NLB	24	22	24	26	26	26	23	23	21
MDB	10	10	10	10	12	9	8	10	9
ZMB	96	103	95	94	95	96	102	94	89
SSS	24	27	24	21	22	27	21	23	22
FMB	92	101	95	96	98	97	101	95	92
NAS	12	21	14	14	14	18	15	11	13
EKB	91	101	93	97	95	97	100	96	92
DKS	9	11	10	9	9	14	12	11	10
DKB	19	25	21	24	21	24	23	23	21
WNB	6	7	6	5	7	8	8	8	9
IML	30	35	33	33	30	34	32	33	32
XML	52	58	48	50	50	50	50	55	48
MLS	8	9	7	10	9	8	9	9	10
WMH	25	28	22	18	25	23	22	23	21
SOS	3	6	4	5	6	7	6	5	5
GLS	2	1	2	2	2	2	1	3	2
FOL	34	42	38	33	39	34	37	37	33
FRC	103	115	102	110	107	112	106	106	108
FRS	23	24	24	28	26	31	23	26	29
FRF	49	44	42	41	40	41	43	44	42
PAC	112	110	105	108	106	115	107	100	105
PAS	27	21	24	26	24	21	22	21	24
PAF	60	60	53	54	55	62	58	60	62
OCC	85	100	88	84	85	90	93	90	93
OCS	19	22	25	24	23	29	25	26	24
OCF	40	45	46	41	37	41	36	41	41
VRR	118	125	115	121	119	121	118	120	119
NAR	88	100	92	91	89	97	96	92	89
SSR	90	100	92	90	90	98	93	95	90
PRR	97	104	96	97	96	108	100	100	96
DKR	78	88	80	79	79	86	86	82	80
ZOR	78	81	80	80	80	83	83	80	78
FMR	78	83	79	79	77	81	81	78	77
EKR	72	78	71	71	72	78	74	70	70
ZMR	68	76	68	71	70	72	75	71	70
BRR	113	122	110	119	115	116	117	119	115
LAR	97	112	100	94	102	99	101	101	100
OSR	40	51	45	39	49	43	45	44	38
BAR	14	15	17	13	16	11	17	13	16



Supplementary Figure 9.1 PCA of the Botocudo specimens based on 56 cranial measurements. The specimens that were included for further genetic analysis are indicated in red, while Bot065 (MN065) was measured for isotopes. See S1 for a description of those three samples.



Supplementary Figure 9.2 Cluster analysis for 32 variables for a wordwide sample including a defined cluster grouping Botocudos (BOT), Lagoa Santa (LST), Easter Islanders (EAS), Mokapu (MOK), Moriori (MOR), Southern Maori (SMA), Northern Maori (NMA) see Supplementary Table 9.6 for a legend of all the population names.

Supplementary 10

Discussion of potential scenarios

Anna-Sapfo Malaspinas, Hannes Schroeder, Manfred Kayser, Mark Stoneking, Eske Willerslev

Several scenarios should be considered in explaining our finding of Polynesian ancestry in Botocudo individuals, some of which have been proposed before (Goncalves et al. 2013). We discuss five of them in this section, considering the new evidence that has come to light.

For what follows, we use the definition of (Patrick 2010) for Near Oceania (New Guinea and adjacent islands eastward, including the Solomon Islands up to Santa Cruz and the Reef Islands) and Remote Oceania (all islands from Santa Cruz and the Reef Islands eastward, including New Caledonia, Vanuatu, Fiji, and Polynesia).

We first discuss some facts about the peopling of Polynesia. Wollstein et al. have shown through genome wide data that present-day Polynesians are genetically distinct from other populations in Near and Remote Oceania. They found strongest support for a model involving a complex demographic scenario suggesting that Polynesians are the result of an admixture event between populations from Southeast Asia such as Borneo (87%) and Near Oceania such as Papua New Guinea Highlanders (13%) around 3,000 years ago. This admixture scenario in the population history of Polynesians was already pinpointed earlier based on Y-chromosome and mtDNA data (e.g., (Kayser et al. 2006)**)**, which demonstrated that 94% of Polynesian mtDNA are of Asian origin while only 6% are of New Guinea origin, with 65% of Polynesian Y-chromosomes being of New Guinean origin and 28% of Asian origin. Based on the "Slow Boat from Asia" hypothesis initially developed mainly based on Y and mtDNA data (Kayser et al. 2000) and later supported by genome-wide data (Kayser et al. 2008; Wollstein et al. 2010), Polynesian ancestors originated from East Asia and on their migration eastwards interacted with and admixed with local New Guineans before colonizing the Pacific. Moreover, a recent compilation of radiocarbon dates suggest that "East Polynesia", a region including some of the Cook Islands, New Zealand and Rapa Nui, was colonized in a very short period spanning a hundred years (around 1200-1300 AD). Our genetic results suggest that the two Botocudo individuals are closest to the Polynesian populations analyzed by Wollstein et al. and Xing et al. and that they are distinct from all other populations analyzed including those from Papua New Guinea, Fiji or Asia.

Scenario (A) Pre-Clovis migration. This scenario involves an ancient pre-Clovis migration wave into the Americas by a population with shared ancestry to Oceanians (Neves and Hubbe 2005). Some researchers have identified two broad craniofacial groups, including a "Paleoamerican" group with similarities to present-day Australians, Melanesians and sub-Saharan Africans (González-José et al. 2001; Neves and Hubbe 2005; Meltzer 2010). They have argued that there must have been an early migration into the Americas from East Asia. This would be the most parsimonious way to explain their findings. However, a direct migration from South Pacific to

South America could not be ruled out. The descendants of these two waves would then co-exist on the American continent in relative isolation from each other, resulting in disparate cranial morphologies. According to this scenario, the early Holocene Paleoamericans from Lagoa Santa would be the descendants of this earlier wave, and would be the ancestors of the Botocudos. The craniometric analysis we present here (S9) could be consistent with these ideas. However, there are no genetic results for the Lagoa Santa remains. Therefore we cannot assess whether the Botocudo and the Lagoa Santa people are genetically related. However, we believe that the "pre-Clovis" scenario can be ruled out as an explanation for the observations presented here, as our whole genome results suggest that the two Botocudo have specifically Polynesian ancestry. Thus, given that the Austronesian expansion began around 4,000 years ago, it is not possible for the Polynesian ancestors of Bot15 and Bot17 to have been involved in the early peopling of the Americas.

Scenario (B) Peru-Polynesia slave trade. It was initially suggested that the Oceanian signature in the Botocudo mtDNA derives from slaves brought to Brazil by Europeans (Goncalves et al. 2013). Between 1862 AD and 1864 AD, around 2,000 individuals were brought by Europeans from Polynesia and Micronesia to Peru and forced into labor. However, the ^{14}C dates for the skulls predate the start of this slave trade (S8) rule out this scenario.

Scenario (C) Madagascar-Brazil slave trade. More than three million slaves are recorded to have disembarked in Brazil, accounting for about forty percent of the transatlantic slave trade (Eltis 2013). Although the majority of slaves were from mainland Africa, a number also originated from Madagascar, which was initially colonized from South-Eastern Asia (Razafindrazaka et al. 2010; Pierron et al. 2014). Recent results pertaining to the transatlantic slave trade suggest that around 7,200 Malagasy slaves were taken to Brazil, and that the first voyage was in 1718 AD (Eltis 2013; 2014). The Polynesian mtDNA motif in the hypervariable region I is found at high frequency in Madagascar today (Soodyall, Jenkins, and Stoneking 1995) reason why this explanation was favored in the mtDNA-based study (Goncalves et al. 2013). However, we can now exclude this scenario for two primary reasons: (1) The ancestry of present-day Malagasy is at least 60% African (Pierron et al. 2014) and the genomes of Bot15 and Bot17 revealed no African ancestry component (S5). Moreover, the admixture event between the Austronesian and Bantu (African) ancestral populations was found to predate the beginning of the 17th century, *i.e.*, before the start of the slave trade. (2) The Austronesian-speaking individuals that colonized Madagascar are more closely related to the extant populations from the Java-Kalimantan-Sulawesi area (Pierron et al. 2014), not Polynesians as found here (S5, but see also S8).

Aside from these two historically-attested slave trades (scenarios B and C) there is - to our knowledge - no other relevant slave trade that could account for the presence of two Botocudo men of Polynesian ancestry in the interior of Brazil in the 15th-18th centuries.

Scenario (D) Polynesian crew, passengers, or stowaways. Another conceivable scenario is that the two individuals (or their ancestors), potentially independently, boarded European ships in Polynesia either as crew, passengers, or stowaways, disembarked somewhere in South America, and finally made their way to the interior of Brazil. To assess the likelihood of this scenario it is important to consider the time frame of European contacts with Polynesia.

While Fernão de Magalhães (Magellan) first spotted some seemingly uninhabited Polynesian islands in 1521 AD (Maude 1959), the earliest written records of European sighting of an inhabited Polynesian island in the Pacific dates to 1568 AD ("Isla de Jesus", which has been identified as Nui (Tuvalu) (Maude 1959; Baert 1999) during Álvaro de Mendaña's first voyage – but no contact with the locals was made during that trip. During Mendaña's second (and last) voyage, the inhabited islands in French Polynesia (the Marquesas) and Pukapuka (Cook Islands) were visited, and contact was made with the locals. Subsequently, the lack of precise navigational techniques (Maude 1959), as well as war between European nations, meant that many of those islands were not visited by Europeans again for another 150 years or more. Therefore, commerce, trade and empire involving Euroamerican ships in the Pacific only began after 1760 AD (Thomas 2010). By 1760 AD, Bot15 and Bot17 were already deceased with a probability of 0.92 and 0.81, respectively (S8). Moreover, although Rio de Janeiro was a regular port of call for the 18th century discovery expeditions bound for the Pacific, Polynesians were rare at the time and any disappearance would have been widely noticed (Glyn Williams, personal communication) – making it the more difficult that a certain number of individuals (presumably more than two since we detected Polynesian ancestry in two out of 35 Botocudo individuals in the Museu Nacional collection) with Polynesian ancestry were in Brazil at the time.

Scenario (E) Polynesian seafarers. Parallels between Polynesian and South American cultural traits were noted as early as 1837 AD, and the possibility of contact was discussed regularly in the first half of the 20th century (Jones et al. 2011). However, the idea of pre-European contact between Polynesians and South Americans was essentially disregarded in the second half of the 20th century, following Thor Heyerdahl's controversial voyage in 1947, which, together with his subsequent work (Heyerdahl 1952), transformed this issue into a highly contentious and divisive topic within the archaeological and anthropological communities (Jones et al. 2011). However, in recent years, linguistic, archaeological and genetic evidence have continued to accumulate in favor of such contact (Jones et al. 2011; Roullier et al. 2013), including linguistic data for the Botocudo population specifically (Pericliev 2006). Several of those results have also remained mired in controversy (*e.g.*, (Gongora et al. 2008; Thomson et al. 2014)). However, it has been well established that the Polynesian Pacific expansion from Southeast Asia covered distances of thousands of kilometers reaching New Zealand, Hawaii and Easter Island – an area approximately the size of North America - between ca. 1200-1300 AD (Wilmshurst et al. 2011). Similarly, it is accepted that Madagascar was originally peopled by South East Asians who would have sailed around 6,000 km (Fitzpatrick and Callaghan, Richard 2014), while, for example, the distance between Easter Island and Ecuador is "only" around 3,000 km (Supplementary Figure 10.1). Thus, Polynesians certainly had the necessary technology and skills to navigate from Polynesia to South America, as has been additionally demonstrated via simulations (Fitzpatrick and Callaghan 2009).

Notably, to date, no direct evidence for Polynesian ancestry in contemporary Native Americans has been reported (*e.g.*, (Reich et al. 2012)), suggesting that they did not descend from (or admix with) Polynesians. This is perhaps not entirely surprising. First the admixture could have taken place during a short time period, *i.*e, between ~1200 AD and ~1800 AD, while the number of individuals potentially involved would have been relatively small compared to the population living in the Americas at the time. This would imply that only a small number of Native American populations could potentially carry the admixture signal. Since the number of native groups in South America to have become extinct in the last 400 years is known to be extremely high, and the remaining are relatively understudied, it could be that some South American populations carry or did carry such a Polynesian admixture signal, but that they are yet to be genetically characterized. It is for example accepted that the Vikings reached the Americas at a similar period (Hall 2013), while it remains unclear if their migration can be detected with genetic data.

If this scenario is indeed the correct one, it is very intriguing that the first possible genomic evidence of such Polynesian contact in South America is to be found in Brazil rather than on the west coast of South America. Based on the evidence at hand, whether these Polynesians arrived via land or by traveling around Tierra del Fuego is a matter of speculation.

Nonetheless, and regardless of how they got there (scenario D or E), the fact remains that at least two Brazilian Botocudo, who likely pre-dated European-Polynesian contact, are of Polynesian genetic ancestry.

It is our hope that these results will stimulate further research into transpacific contacts and open up avenues for additional sampling. For example, the generation of new isotopic data from reference populations as well as other Botocudo samples could help to re-evaluate the radiocarbon dates by determining if the marine correction applied here is indeed necessary and to test whether those individuals are first generation migrants (something that cannot be established with the data at hand). Moreover, genetic data on additional Botocudo samples would also aid in evaluating how widespread this Polynesian ancestry is among Botocudos.

References

Baert, Annie. 1999. *Le paradis terrestre : un mythe espagnol en Océanie - Les voyages de Mendana et de Quiros, 1567-1606*. Paris: Editions L'Harmattan.

Denis Pierron, Harilanto Razafindrazaka, Luca Pagani, Francois-Xavier Ricaut, Tiago Antao, Melanie Capredon, Clement Sambo, et al. 2014. "Genome-Wide Evidence of Austronesian-Bantu Admixture and Cultural Reversion in a Hunter-Gatherer Group of Madagascar." *Proceedings of the National Academy of Sciences*, January, 201321860. doi:10.1073/pnas.1321860111.

Eltis, D. 2013. "A Brief Overview of the Trans-Atlantic Slave Trade," Voyages: The Trans-Atlantic Slave Trade Database," Accessed May 27. http://www.slavevoyages.org/tast/assessment/essays-intro-01.faces.

Eltis, David. 2014. "Voyages: The Trans-Atlantic Slave Trade Database." Accessed July 21. http://slavevoyages.org/tast/database/search.faces?yearFrom=1514&yearTo=1866&mjbyptimp=60811&mjslptimp=50000.

Fitzpatrick, Scott, and Callaghan, Richard. 2014. “Seafaring Simulations and the Origin of Prehistoric Settlers to Madagascar.” Accessed February 6. http://www.academia.edu/1792262/Seafaring_Simulations_and_the_Origin_of_Prehistoric_Settlers_to_Madagascar.

Fitzpatrick, Scott M., and Richard Callaghan. 2009. “Examining Dispersal Mechanisms for the Translocation of Chicken (Gallus Gallus) from Polynesia to South America.” *Journal of Archaeological Science* 36 (2): 214–23. doi:10.1016/j.jas.2008.09.002.

Goncalves, V. F., J. Stenderup, C. Rodrigues-Carvalho, Hilton P. Silva, H. Goncalves-Dornelas, A. Liryo, T. Kivisild, et al. 2013. “Identification of Polynesian mtDNA Haplogroups in Remains of Botocudo Amerindians from Brazil.” *Proceedings of the National Academy of Sciences*, April. doi:10.1073/pnas.1217905110.

Gongora, Jaime, Nicolas J. Rawlence, Victor A. Mobegi, Han Jianlin, Jose A. Alcalde, Jose T. Matus, Olivier Hanotte, et al. 2008. “Indo-European and Asian Origins for Chilean and Pacific Chickens Revealed by mtDNA.” *Proceedings of the National Academy of Sciences* 105 (30): 10308–13. doi:10.1073/pnas.0801991105.

González-José, Rolando, Silvia L. Dahinten, María A. Luis, Miquel Hernández, and Hector M. Pucciarelli. 2001. “Craniometric Variation and the Settlement of the Americas: Testing Hypotheses by Means of R-Matrix and Matrix Correlation Analyses.” *American Journal of Physical Anthropology* 116 (2): 154–65. doi:10.1002/ajpa.1108.

Hall, Richard. 2013. *The World of the Vikings*. 1 edition. London: Thames & Hudson.

Heyerdahl, Y. 1952. *American Indians in the Pacific; the Theory Behind the Kon-Tiki Expedition.* Chicago: Rand McNally.

Jones, Terry L., Alice A. Storey, Elizabeth A. Matisoo-Smith, and José Miguel Ramírez-Aliaga. 2011. *Polynesians in America: Pre-Columbian Contacts with the New World*. Rowman Altamira.

Kayser, Manfred, Silke Brauer, Richard Cordaux, Amanda Casto, Oscar Lao, Lev A. Zhivotovsky, Claire Moyse-Faurie, et al. 2006. “Melanesian and Asian Origins of Polynesians: mtDNA and Y Chromosome Gradients Across the Pacific.” *Molecular Biology and Evolution* 23 (11): 2234–44. doi:10.1093/molbev/msl093.

Kayser, Manfred, Silke Brauer, Gunter Weiss, PeterA. Underhill, Lutz Roewer, Wulf Schiefenhövel, and Mark Stoneking. 2000. “Melanesian Origin of Polynesian Y Chromosomes.” *Current Biology* 10 (20): 1237–46. doi:10.1016/S0960-9822(00)00734-X.

Kayser, Manfred, Oscar Lao, Kathrin Saar, Silke Brauer, Xingyu Wang, Peter Nürnberg, Ronald J. Trent, and Mark Stoneking. 2008. “Genome-Wide Analysis Indicates More Asian than Melanesian Ancestry of Polynesians.” *The American Journal of Human Genetics* 82 (1): 194–98. doi:10.1016/j.ajhg.2007.09.010.

Maude, H. E. 1959. “Spanish Discoveries In The Pacific” 68 (4): 285–326.

Meltzer, David J. 2010. *First Peoples in a New World: Colonizing Ice Age America*. 1st ed. University of California Press.

Neves, Walter A., and Mark Hubbe. 2005. “Cranial Morphology of Early Americans from Lagoa Santa, Brazil: Implications for the Settlement of the New World.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (51): 18309–14. doi:10.1073/pnas.0507185102.

Patrick, V. Kirch. 2010. “Peopling of the Pacific: A Holistic Anthropological Perspective.” *Annual Review of Anthropology* 39 (1): 131–48. doi:10.1146/annurev.anthro.012809.104936.

Pericliev, Vladimir. 2006. “Significant Lexical Similarities between a Language of Brazil and Some Languages of Southeast Asia and Oceania: From Typological Perspective.” *Journal of Universal Language* 7 (2): 121–45.

Razafindrazaka, Harilanto, Francois-X Ricaut, Murray P Cox, Maru Mormina, Jean-Michel Dugoujon, Louis P Randriamarolaza, Evelyne Guitard, Laure Tonasso, Bertrand Ludes, and Eric Crubezy. 2010. “Complete Mitochondrial DNA Sequences Provide New Insights into the Polynesian Motif and the Peopling of Madagascar.” *European Journal of Human Genetics* 18 (5): 575–81. doi:10.1038/ejhg.2009.222.

Reich, David, Nick Patterson, Desmond Campbell, Arti Tandon, Stephane Mazieres, Nicolas Ray, Maria V. Parra, et al. 2012. “Reconstructing Native American Population History.” *Nature* 488 (7411): 370–74. doi:10.1038/nature11258.

Roullier, Caroline, Laure Benoit, Doyle B. McKey, and Vincent Lebot. 2013. “Historical Collections Reveal Patterns of Diffusion of Sweet Potato in Oceania Obscured by Modern Plant Movements and Recombination.” *Proceedings of the National Academy of Sciences* 110 (6): 2205–10. doi:10.1073/pnas.1211049110.

Soodyall, Himla, Trefor Jenkins, and Mark Stoneking. 1995. “‘Polynesian’ mtDNA in the Malagasy.” *Nature Genetics* 10 (4): 377–78. doi:10.1038/ng0895-377.

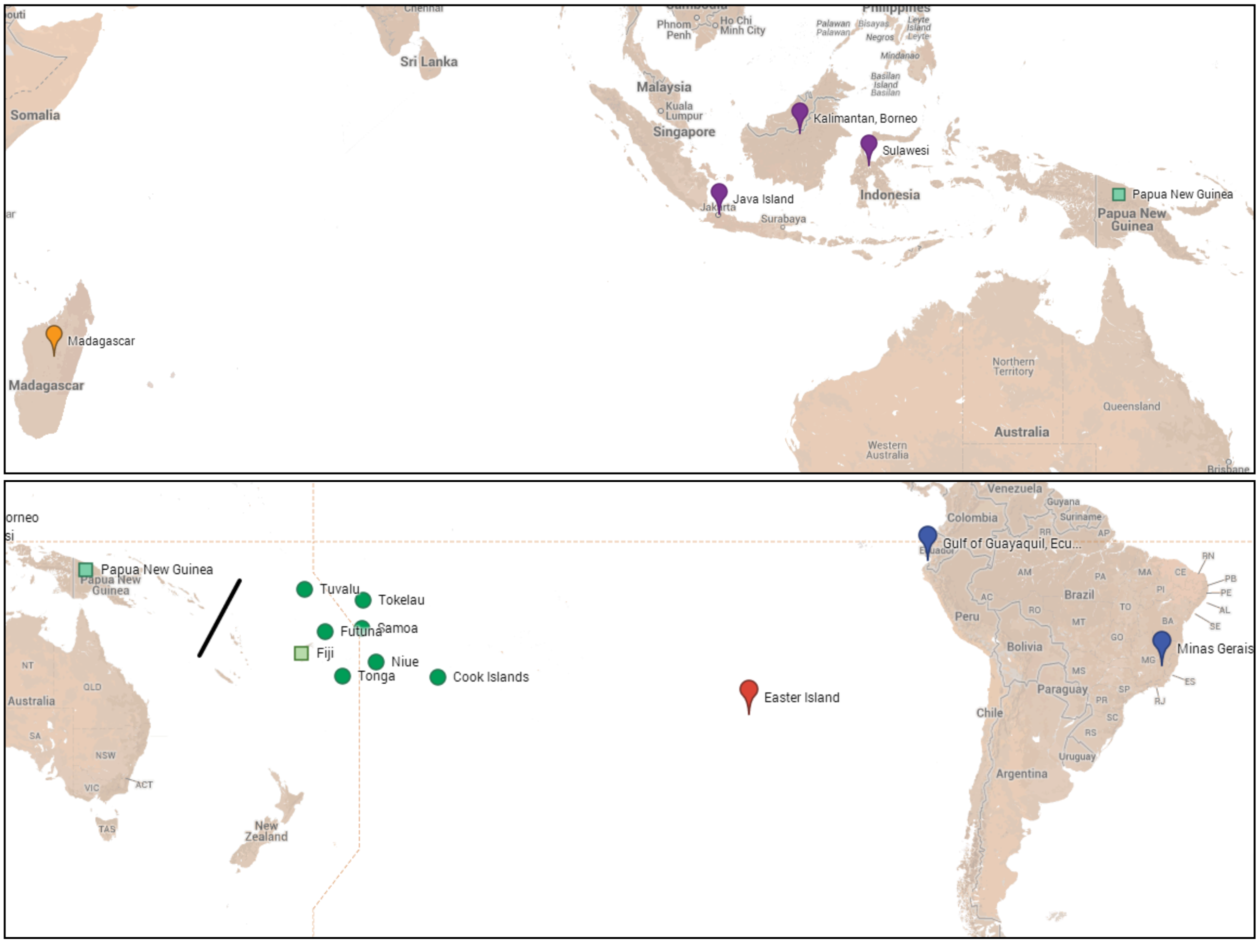
Thomas, Nicholas. 2010. *Islanders: The Pacific in the Age of Empire*. Yale University Press.

Thomson, Vicki A., Ophélie Lebrasseur, Jeremy J. Austin, Terry L. Hunt, David A. Burney, Tim Denham, Nicolas J. Rawlence, et al. 2014. “Using Ancient DNA to Study the Origins and Dispersal of Ancestral Polynesian Chickens across the Pacific.” *Proceedings of the National Academy of Sciences* 111 (13): 4826–31. doi:10.1073/pnas.1320412111.

Wilmshurst, Janet M., Terry L. Hunt, Carl P. Lipo, and Atholl J. Anderson. 2011. “High-Precision Radiocarbon Dating Shows Recent and Rapid Initial Human Colonization of East Polynesia.” *Proceedings of the National Academy of Sciences* 108 (5): 1815–20. doi:10.1073/pnas.1015876108.

Wollstein, Andreas, Oscar Lao, Christian Becker, Silke Brauer, Ronald J. Trent, Peter Nurnberg, Mark Stoneking, and Manfred Kayser. 2010. “Demographic History of Oceania Inferred from Genome-Wide Data.” *Current Biology* 20 (22): 1983–92. doi:10.1016/j.cub.2010.10.040.

Xing, Jinchuan, W. Scott Watkins, Adam Shlien, Erin Walker, Chad D. Huff, David J. Witherspoon, Yuhua Zhang, et al. 2010. “Toward a More Uniform Sampling of Human Genetic Diversity: A Survey of Worldwide Populations by High-Density Genotyping.” *Genomics* 96 (4): 199–210. doi:10.1016/j.ygeno.2010.07.004.



Supplementary Figure 10.1 Location of the regions mentioned in the text. Top: regions relevant to the colonization of Madagascar, notably regions within Indonesia (Kalimantan, Sulawesi, Java). Kalimantan refers to the Indonesian portion of the island of Borneo. Sulawesi and Java are islands in Indonesia. Bottom: (1) Islands for which genotype data exist (Wollstein et al. 2010; Xing et al. 2010) (green). (2) Solid black line indicates the divide between Near and Remote Oceania (Patrick 2010)(Wollstein et al. 2010)(Wollstein et al. 2010). (3) Position of Easter Island, the Polynesian island closest to South America. (4) A potential landfall in South America (the Gulf of Guayaquil in Ecuador, (Fitzpatrick and Callaghan 2009; Jones et al. 2011)). (5) Minas Gerais, the state in Brazil where the Botocudo skulls were found (S1). Both maps were drawn with Google maps Engine Lite (Map data © 2014 Google, INEGI).