SUPPLEMENTARY INFORMATION

doi:10.1038/nature18299

CONTENTS

S01 Ethical approvals in relation to sampling in Australia

Craig Muller, Michael C Westaway, Joanne L Wright, Tim H Heupink, Anna-Sapfo Malaspinas, Eske Willerslev, David M Lambert

Background

From its inception, this research project has been a collaboration between research partners at the Centre for GeoGenetics at the University of Copenhagen and Griffith University, together with a number of Aboriginal individuals and groups. Initially in 2010, researchers from Griffith University, on behalf of the Copenhagen / Griffith University team, established research in collaboration with the Paakantji, Ngyiampaa and later with the Mutthi Mutthi peoples to study ancient remains from the Willandra Lakes area. Similarly, in 2011 Professor Eske Willerslev from GeoGenetics initiated discussions with Wongatha, Ngadju and other Aboriginal Australian peoples in Western Australia. These discussions were intended to gauge support for the publication of the first Aboriginal genome, obtained from a hair sample collected in the Goldfields region in the 1920s. This permission was agreed to and the genome was published in 2011 (Rasmussen et al. 2011). Subsequently, both researchers and the Aboriginal groups involved in that project expressed interest in additional research. Hence, collaborations were expanded to include the range of groups represented in this study.

Sampling

Aboriginal Australians from numerous language groups across Australia were approached to participate in the research project. Research team members from Griffith University collected samples from Eastern Australia (BDV, CAI and WPA), while research team members from the University of Copenhagen collected samples from Western Australia and the Riverine area of New South Wales (ENY, WCD, WON, NGA and RIV) (Figure S01.1). Each institution obtained their own ethics approval and samples were collected under the ethical guidelines set forward by the researcher's home institution.

DNA samples from the BDV, CAI and WPA Aboriginal Australians collected by Griffith University

Sample collection was planned with the guidance of Aboriginal Elders from each community. These Aboriginal Elders, or their representatives, joined the research team in order to initiate contact with the community members interested in participating.

In accordance with the National Statement on Ethical Conduct in Human Research, we submitted a Human Ethics Research Application (Ref No: ENV/20/13/HREC) with the Griffith University Human Research Ethics Committee (HREC). This application included the submission and subsequent approval of the consent package: a plain English information sheet, which was provided to all members of the community who were interested in the project, and a consent form.

Before collecting a sample from a potential participant, researchers spoke to the community and outlined the expected benefits of the research. Discussions outlined the possible risks and explained in depth how their genetic data would be treated confidentially and anonymously, using a de-identification system from the time of collection before sending the original consent forms (see at end of section for the "consent package" - Appendix S01.1- that was shared with each participant) to a third-party to hold on our behalf. It was stressed that their participation was voluntary and the participants were advised they could withdraw from the research at any time by contacting the third-party.

DNA samples from the ENY, NGA, PIL, WCD, RIV and WON Aboriginal Australians collected by the University of Copenhagen

Project ethics were constructed using the research guidelines set by the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) and the Free, Prior and Informed Consent (FPIC) protocols for working with Indigenous peoples set by the *United Nations Declaration on the Rights of Indigenous Peoples 2007*. Following Danish law, the project proposal was submitted to the The National Committee on Health Research Ethics, Denmark (H-3-2014-FSP26). Initial meetings were held with key individuals of Aboriginal communities, and whenever possible, a senior person from the group was engaged as a consultant and culturally appropriate liaison. The ideas and suggestions put forward by these representatives were incorporated into the planning stages of this research. Discussions with potential participants included a background to the genetic research. Participants were made aware that while the results would be published, their identities would remain anonymous. Participants were advised that if they wished to withdraw from the study at any time they may do so by contacting the elder from their group or the locally-based researcher without having to offer any explanation for their decision.

Plain English consent forms were provided to and signed by each participant - Appendix S01.2 - who were also filmed giving their consent. To protect anonymity, the filmed consents are held securely and are not directly accessible to anyone outside the immediate research team. If there was a challenge to the process of obtaining consent, an arrangement will be made for a mutually acceptable third party to view the footage and confirm that consent was freely given.

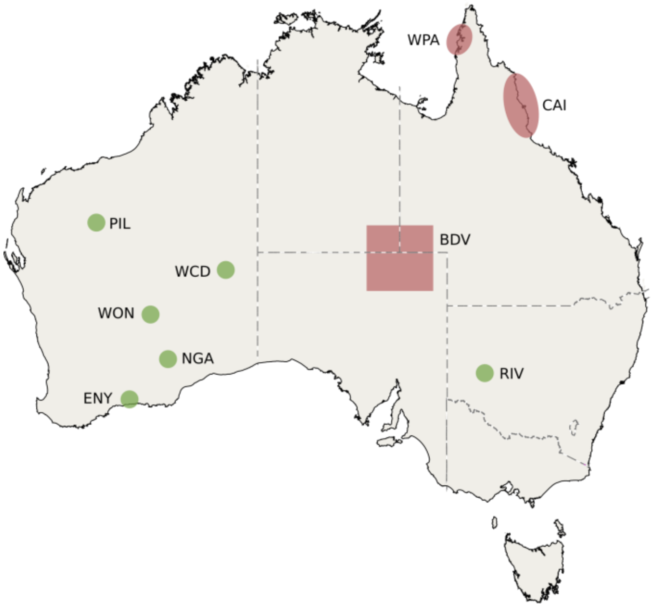


Figure S01.1 Contemporary sampling localities. In red are the broad areas covered by Griffith University's ethics approval and in green are localities covered by the University of Copenhagen's ethics approval. Basemap: © OpenStreetMap.org contributors. Regions are indicated by the following abbreviations: BDV (Birdsville), CAI (Cairns), WPA (Weipa), ENY (Esperence Nyungar), WCD (Western central desert), WON (Wongatha), NGA (Ngadju) and RIV (Riverine).

Appendix S01.1 **Griffith University consent package & consent form.**

The peopling of Australia

INFORMATION SHEET

Who is conducting the research?

Chief investigator:
Prof David M Lambert
Environmental Research Centre
07 373 55298
d.lambert@griffith.edu.au

Other investigators:

Griffith University, AU: Prof Adrian Miller (Professor of Indigenous Research), Prof Paul Tacon, Prof Brian Fry, Dr Michael Westaway, Dr Tim Heupink, Dr Subashchandran Sankarasubramanian, Ms Joanne Wright.

University of Copenhagen, DK: Prof Eske Willerslev, A/Prof Martin Sikora, Dr Craig Muller, Dr Anna-Sapfo Malaspinas.

Australian National University, AU: Dr Duncan Wright.

University of Otago, NZ: Prof Lisa Matisoo-Smith

University of Auckland, NZ: Dr Craig Millar.

Natural History Museum, UK: Dr Margaret Clegg.

Peking University, CH: Dr Ruiqiang Li.

Queensland Museum, AU: Mr Nicholas Hadnutt.

University of Western Australia, AU: A/Prof Joe Dortch.

University of New South Wales, AU: A/Prof Darren Curnoe, Dr Sheila van Holst-Pellekaan.

Simon Fraser University, CA: Dr Mark Collard.

Why is the research being conducted?

This study investigates the history of Aboriginal and Torres Strait Island People in Australia. This is done by characterising the DNA of both contemporary people and those that lived up to 45,000 years ago. We will compare the DNA of all these individuals and also compare it with other people from all over the world. We aim to investigate the origin of the First Peoples of Australia and study any subsequent migrations within, to and from Australia. We will also investigate how ancient and contemporary Australians compare and are related to each other. These genetic data will also reveal how and when other populations have been in contact with Australian Aboriginals. The study of both modern and ancient Australian Aboriginals may also reveal how the Australian Peoples interacted with each other and how cultures and technologies were exchanged within Australia. This research will not investigate disease related questions.

What you will be asked to do

DNA is a molecule that contains the genetic information, describing much of an organism or individual. DNA exists throughout the body, particularly in cells, some of which get deposited in the saliva. The bulk of a person’s DNA has been inherited from both parents, the DNA therefore not just reveals information about the individual but also about the parents, grandparents and earlier ancestors.

DNA will be collected using spit sample kits, this is an hygienic and safe way to collect DNA. A funnel helps you deposit your saliva in a collection tube. After having deposited sufficient saliva (up to the fill line) the funnel can be discarded and the tube capped and deposited in the collection box. The sealed tube will be transported to the Brisbane laboratory where it will be prepared for shipment to our colleagues in Copenhagen and Beijing. There will be no other transfer of samples.

The DNA is multiplied through amplification to create enough synthetic DNA for future analyses and is stored in a freezer. The original and the copies of your DNA will be stored for a maximum of 5 years and are destroyed afterwards. The genome, a person's complete set of genetic material, will be characterised using technology available in Copenhagen and Beijing. After characterising the genome it will be analysed by members of the research team. The genome will be compared with that of other people from other groups and areas and with ancient First Australians. The relation of this DNA will reveal how people and populations are related and may reveal when and where they migrated to and from.

The basis by which participants will be selected or screened

We are particularly interested in obtaining DNA samples from those individuals whose immediate ancestors (parents and grandparents) are most likely of direct Australian Aboriginal descent. These results provide us with the most information about the history of the Australian Peoples. It is for this reason we will ask about your direct ancestry. It is voluntary to provide this information. We are not able to accept samples from minors.

The expected benefits of the research

This study may reveal the history of Australia’s First People in that it may indicate their origins and how they interacted with other people in other parts of the world. We aim to investigate the number of individuals and populations that gave rise to Australia’s First People. We will also study the ancient migrations of these individuals and populations and their ancestors within and outside of Australia. In addition the study may reveal how certain cultural traditions and technologies have been exchanged across Australia.

The project also holds the potential to create a DNA map of Australia’s First People and help identify the origin of Aboriginal skeletal remains that are being returned to Australia by museums.

Risks to you

The saliva sampling kit we use prevents any potential risk to you. With regards to privacy please refer to the following section.

Your confidentiality

DNA can also hold information that can be considered more private, for example in relation to genetic disease (although we will not investigate this aspect). The sample you give is immediately de-identified, the sample tubes are mixed with others and your consent form is kept separately in a locked box. A third party will ensure the consent forms are kept safe and separate from the samples afterwards. This de-identification ensures to the maximum extent that all results are published and reported anonymously and cannot be retraced to the individual. Despite our careful de-identification, your characterised genetic material is in principle re-identifiable. This means that someone with access to the data could in theory link the DNA data to you, despite the de-identification; we try to prevent this from happening in every possible way. The resulting de-identified data are available for researchers wishing to verify the results of this study only with an ethics approval. Any other research will have to be approved first by you, then by the research team and an appropriate ethics committee.

Your participation is voluntary

You are advised that your participation is voluntary and are free to decline without giving reasons. Also, if you agree to participate, you are free to withdraw from the study at any time, at which point your DNA will be destroyed. An independent third party will hold files that enable the cross referencing of names to individual samples, so that these can be destroyed in the event that a participant wants to withdraw from the study. Please contact Dr Donald R. Love at the Auckland City Hospital, New Zealand on donaldl@adhb.govt.nz or +64 9 307 4949 22013 in the event that you would like to withdraw from this study. Dr. Donald R. Love is not part of the research group, but an independent third party that will look after the consent forms and the numbers that associate these with the samples in the laboratory.

Questions / further information

You can contact any member of the research team that is present when you receive this sheet with questions. You can also contact the Chief Investigator at any other point in time, contact information is provided above.

The ethical conduct of this research

Griffith University conducts research in accordance with the *National Statement on Ethical Conduct in Human Research*. If you have any concerns or complaints about the ethical conduct of the research project you should contact the Manager, Research

Ethics on 3735 5585 or research-ethics@griffith.edu.au.

Feedback to you

The results obtained from this study will be published in peer-review scientific journals. As a result of the de-identification we can not report any individual results back to you, it is for this reason that we can not report any medical or family related results back to you. Instead, you and your community will be invited to a local presentation where we will present the results of this study.

Privacy Statement

The conduct of this research involves the collection, access and / or use of de-identified personal information only. Participants' anonymity will at all times be safeguarded. For further information consult the University's Privacy Plan at www.griffith.edu.au/ua/aa/vc/pp or telephone (07) 3735 5585.

The peopling of Australia
CONSENT FORM

Research Team

Chief investigator:
Prof David M Lambert
Environmental Research Centre
07 373 55298
d.lambert@griffith.edu.au

By signing below, I confirm that I have read and understood the information package and in particular have noted that:

- I understand that my involvement in this research will involve providing a saliva sample from which a complete genome will be characterised;
- I understand that this study will not undertake any form of health testing;
- I understand that my DNA may be frozen for future use in this study;
- I agree the sample may be sent to members of the research team in other overseas centres for the purposes of this study;
- I have had any questions answered to my satisfaction;
- I understand the risks involved;
- I understand that there will be no direct benefit to me from my participation in this research;
- I understand that my participation in this research is voluntary;
- I understand that, because all samples have been de-identified prior to analysis, it is not possible to receive individual results;
- I understand that the information gained from this research may result in improved methods for analysis, but as an individual I do not have ownership of these results, the research records, or the sample that I give;
- I understand that if I have any additional questions I can contact the research team;
- I understand that I am free to withdraw at any time, in which case my DNA will be destroyed, without comment or penalty;
- I understand that I can contact the Manager, Research Ethics, at Griffith University Human Research Ethics Committee on 373 54375 (or research-ethics@griffith.edu.au) if I have any concerns about the ethical conduct of the project;
- **I agree to participate in the project.**

Name	
Signature	
Date	

Sample #	

Appendix S01.2 **Copenhagen University consent form**

CONSENT FORM FOR ABORIGINAL AUSTRALIAN DNA RESEARCH
CENTRE FOR GEOGENETICS, UNIVERSITY OF COPENHAGEN

Explanation

I have been asked to participate in a research study regarding the genetic characteristics of Aboriginal populations. The nature of the study has been fully explained to me and I have had the opportunity to ask questions and express my opinions.

Assurances

I have volunteered to participate in this study and I understand I may choose to stop participating in the study at any time.

Procedure

I understand that my participation in this study will last for a brief period (usually 20-30 minutes). During that time, I will provide saliva samples. The saliva will be used for full genome sequencing and analysis and it will be compared to other genomic sequences. This will determine broad population characteristics and histories.

I understand that this information will eventually be published and available, including on the internet, but my personal identity will not be revealed and my name will not be used without my permission.

I also understand that I will be filmed giving my consent and that this film will only be available to people directly involved in this study.

Physical risk/discomfort in the procedure

None

Costs

There will be no cost to me.

Expenses

I have not been paid to participate in this study.

Other

____

____

Authorization

I have read the explanation above and I understand it; *or*
This form has been read/translated to me and I understand it.

I agree to take part in this study, and I have not been pressured or made to feel obligated to take part.

Signed overleaf

Participant

______________________________ ______________________________________

Name Address

______________________________ ______________

Signature Date

Investigator

____

Name

______________________________ _____________

Signature Date

Witness

____

Name

______________________________ _____________

Signature Date

S01 References

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, et al. 2011. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334:94–98.

S02 Ethnography and linguistics for the Aboriginal Australian individuals

Craig Muller, Claire Bowern

Ethno-historic information on the participant language groups

The participating groups represent a wide range of cultures and languages across Australia. Each region has its own contact history, the complexity of which includes post-contact (i.e. during the last three to ten or so generations) gene flow. The following offers brief descriptions of the language groups, their territories and some comment on what gene flow is likely to have occurred pre- and post-contact, based on historical information.

KEY:
code; (number of participants); main language group/s; regional cultural groupings*; geographic area
*per AIATSIS map (Horton 1994)

WPA; (6 individuals); Yupangati and Thanakwithi; West Cape; northeastern Australia

The six samples donors from this area primarily belong to the Yupangati and/or Thanakwithi language groups, which occupy the country on the western side of Cape York between Albatross Bay and Cullen Point. These groups traditionally relied heavily on marine and coastal swamp resources (Meston 1986). They had extensive contact and trade with people in the Torres Strait (Chase, 1981; Haddon, 1901b) and Papua New Guinea (Macknight, 1972) in the period before European settlement. In particular, a regular maritime trade route existed running from Papua New Guinea through the Torres Strait to the “main point of contact” at Batavia River (the mission at Mapoon) on the west coast of Cape York (Haddon, 1901a; McCarthy, 1939b, p. 182). On the eastern side of Cape York there was considerable intermarriage between the Torres Strait and Aboriginal groups (McCarthy, 1939a).

Indonesian fishermen from Makassar, Sulawesi, visited northern Australia on a regular basis from at least the 1720s (some 170 years before European settlement in the Gulf of Carpentaria region) until the early twentieth (Macknight, 1986, 1972). Generally, they are reckoned to have travelled regularly east as far as the southern coast of the Gulf of Carpentaria (e.g. (Russell, 2004)) and contact with Aboriginal groups on the west coast of Cape York, while not conclusive, is a distinct possibility (Tacon and May 2013).

Cattle stations were established in Cape York in the 1860s but the whole region, and in particular the west coast, remained very sparsely populated by non-Aboriginal people. Contact between Aborigines and non-Aboriginal (other than Torres Strait Islander, Papuan and perhaps Indonesian) people would likely have been rare until church missions were established at Mapoon in 1891 and Weipa in 1898. Today, most Yupangati and Thanakwithi live in the main regional town of Weipa with some others at Mapoon which remains an Aboriginal community. Samples were taken at those two places.

CAI; (10 individuals); Yidindji and Gungandji; Rainforest; northeastern Australia

With one exception, the participants belong to the Yidindji and Gungandji groups, whose country lies just north and south of the north Queensland town of Cairns. These groups have the strongest connections to their northern and southern coastal neighbours and have lesser connections westward to the West Cape York groups (McCarthy, 1939a). The Yidiny and Gunggay languages were close enough to be regarded as dialects of a single language (Dixon, 1977).

Until the 1850s, the Yidindji and Gungandji had little contact with non-Aboriginal people other than the occasional maritime exploration venture along the eastern Australia seaboard. At that time, the commercial harvesting of *bêche-de-mer* began in the region, a process that included the use of local Aboriginal labour. This interaction almost certainly included relationships between Yidindji and Gungandji women and non-Aboriginal men (Yarrabah Aboriginal Shire Council, *http://www.yarrabah.qld.gov.au/en_US/history*).

Gold was discovered inland of Yidindji and Gungandji country in the 1870s and the subsequent arrival in of thousands of non-Aboriginal people and the establishment of the port of Cairns to support the mining led to extensive population exchanges. The goldfields, the port and later the agriculture in the area attracted a variety of outside groups, including British, Irish, Italian, Chinese and Kanakas from Melanesia.

Yarrabah mission was established as a home for Aboriginal people just south of Cairns in the 1890s and was the largest mission in Queensland by 1903 (Yarrabah Aboriginal Shire Council n.d.). People from many different language groups were forced to relocate to Yarrabah (Tindale, 1938), introducing the likelihood of considerable mixing of groups that were previously isolated from one another.

BDV; (10 individuals); Wangkangurru and Yarluyandi; Eyre; northeastern central Australia

Samples were taken from Aboriginal people traditionally from the Birdsville area, an outback town near the borders of New South Wales (NSW), Queensland and South Australia. The participants identified as members of the Wangkangurru and/or Yarluyandi language groups, both members of the Karnic subgroup of Pama-Nyungan (Bowern, 2009). Their home country lies in the very far north of the state of South Australia crossing into the Northern Territory and Queensland (South Australia Museum Archive). Hercus (1987) documents extensive trade and ceremonial networks across this region in traditional times, spanning multiple language groups.

The Birdsville region was crossed by numerous overland exploration expeditions in the middle of the nineteenth century. Some limited intermarriage between the men of those expeditions and Aboriginal women may have occurred. A cattle industry was established in the region in the 1870s and Birdsville was at the centre of a cattle-driving route. Greater number of mixed-population relationships could be expected from this time although the non-Aboriginal population has always been thinly spread in this region.

RIV; (8 individuals); Barkindji; Riverine; southeastern Australia

The participants from the Riverine region are all members of or have strong connections to the Barkindji language group, although some identified primarily with the neighbouring

Maraura, Ngiyambaa and Kurnu groups. The Barkindji (also known as Paakantyi) language group occupy an almost 800 kilometre stretch of country along the Darling River from its junction with the Murray River north almost to the Queensland border (Hercus 1986; National Native Title Tribunal Geospatial Analysis & Mapping Branch, 2004). Barkindji means river people, where the river, the Barka, is the centre of economic, cultural and spiritual life (Murdi Paaki 2014, *http://www.mprec.org.au/*). It is semi-arid country, yet was relatively densely populated by related groups whose members were able to move with considerable freedom between territories (Allen, 1974). The river (Allen, 1974) and a strip of land 30-50 kilometres each side (South Australia Museum Archives) provided the majority of resources for the Barkindji. Alone among the participant groups here, the Barkindji have previously worked with geneticists (van Holst Pellekaan et al., 2006). Barkindji and Kurnu are closely related varieties of a single language (Hercus 1986).

WCD; (13 individuals); Ngaanyatjarra; Desert; western central Australia

Participants are Ngaanyatjarra, one of the language groups that make up the Western Desert Cultural Bloc, an area of common culture and language that covers the central arid zone of Australia – approximately one third of the continent. Their country extends from just west of the Western Australia/South Australia border in the east, to approximately Cosmo Newbery. They practice a desert subsistence economy.

Initial occupation of the Western Desert arid zone dates to the Pleistocene and possibly as early as 39 000 BP (Smith et al., 1997). Abandonment of much or perhaps the entire desert zone, with the exception of refugia mainly on the margins, is likely to have occurred during the Last Glacial Maximum (Veth, 1989). Permanent settlement of the arid interior resumed or expanded from a small base by the terminal Pleistocene/early Holocene (Smith et al., 1998) and the entire zone was occupied by the mid-Holocene (O'Connor and Veth, 1996; Veth, 1993). Considerable inward migration and population growth is believed to have occurred during the last two millennia (Smith, 1996). Archaeological and linguistic evidence suggests the Hamersley Range area in the Pilbara is favoured as a major source for this repopulation (McConvell, 1996; Veth, 2000), a theory backed up by ethnography recorded in the Western Desert (Tindale, 1966).

Several exploration parties visited the Warburton Range area between the 1890s and 1920s but contact was minimal until the 1930s by which time the area was visited regularly by prospectors and would-be pastoralists. In 1934 the Mount Margaret mission opened a branch at Warburton, which became an independently operating institution three years later (Green, 1983; Neville, 1935). Beginning about 1906, Western Desert Aborigines began coming in to the mining towns of the northern Goldfields and a few cross-cultural sexual relations occurred. Since the 1970s there has been a move back to the outstations on the groups' original country and the Western Desert and non-Aboriginal populations have tended to increase their separateness. Today, most Ngaanyatjarra are based at the Aboriginal communities of Warburton, Jameson and Blackstone. The non-Aboriginal population in the Western Desert has remained very small and there are a small minority of mixed-descent people now living at the communities. Most of the sample donors related details of their parents and grandparents showing their families were not admixed.

WON; (11 individuals); Wongatha, Tjupan and Koara; Desert; western central Australia

Most of the group labelled WON are primarily speakers of the Wongatha (Wangkatja) dialect but include a minority of individuals who belong to the closely related language groups Tjupan and Koara (see also Extended Data Table 1). These are the three most south-western of the Western Desert (Wati) dialect groups. Pre-contact occupation of this country is broadly as for the Ngaanyatjarra above.

Wongatha country was crossed by numerous overland exploration parties from 1869 until the 1890s. In total these expeditions would have numbered several dozen men, though contact between Aboriginal people and the exploration parties was minimal and sexual relations would likely have been rare.

Thousands of non-Aboriginal people began arriving in Wongatha country in 1892, following discoveries of gold at various places north of Kalgoorlie. The majority of the newcomers were from the British Isles but there were also people from northern, central and south-eastern Europe as well as small numbers of camel drivers, brought from what is now Pakistan and Afghanistan, a few Chinese and even fewer Japanese and Filipino people. It is known that many relationships between Aboriginal women and non-Aboriginal men date from the mid-1890s.

Today, many of the groups' members live in the towns of Kalgoorlie, Leonora and Laverton in the Goldfields region. They are related to the WCD participants although each of the groups identify with a particular area of country and have certain restrictions on entering the neighbouring areas.

PIL; (12 individuals); Yinhawangka, Banjima and Guruma; Northwest; northwestern Australia

The Yinhawangka, Banjima and Guruma are inland Pilbara groups occupying the Hamersley Range area, thought to have been a place of refuge from dryer conditions during the LGM (Veth, 1993). This is a semi-arid area but with irregular, sometimes high summer rainfall. There is today little direct connection between these language groups and those to the south.

Non-Aboriginal people began a permanent presence in the Pilbara in the 1860s. Prior to the beginnings of the pastoral stations and the small ports to service them there was very little contact between Pilbara Aborigines and non-Aboriginal people. Pastoralism and the ports declined in the early twentieth century and the area's non-Aboriginal population remained small until mining began in the 1930s and greatly expanded in the 1960s. Nonetheless, much of the Pilbara's current Aboriginal population includes European and other non-Aboriginal ancestry, often dating back to the late nineteenth century.

NGA; (6 individuals); Ngadju; Southwest; southwestern Australia

The Ngadju occupy the country south of Kalgoorlie across to the Southern Ocean at Israelite Bay (which derives its name from being the limit of the country of circumcision-practising groups at the time of first contact). Very limited archaeological research has happened in the area. The group has strongest cultural links with the Kalamaia-Gubrun to the northwest and the Mirning to the east, with less strong connections to the Nyungar to the southwest and Western Desert groups to the north (e.g. (Bates, 1985)). Ngadju country is mostly inland semi-arid woodland but includes some coastline. Linguistically, Ngadjumaya and Mirniny are closely related varieties (von Brandenstein, 1980).

The Ngadju made early and fleeting contacts with non-Aboriginal people between the 1840s and 1890s. Pastoralism was established in the area in 1872 when Fraser Range sheep station was built in the heart of Ngadju country. Further stations were begun in the eastern part of the country during the 1870s. Non-Aboriginal employees living on the stations probably began relationships with Aboriginal women soon after that time. In 1892-93, gold was discovered in the north-western part of Ngadju country and several thousand non-Aboriginal people arrived over the next few years. By the middle of the twentieth century most of the Ngadju population was of mixed-descent.

ENY; (8 individuals); Nyungar, Southwest; southwestern Australia

Participants are Esperance Nyungar, the easternmost subgroup of the large Nyungar language group that occupies the southwest of Western Australia. This is a coastal group with strong ties extending to the west and weaker links to the east and north. To the west of Esperance, the south-western tip of Australia is known to have been occupied at 48 000BP (Turney et al., 2001) and to the east the Nullarbor Plain was occupied at 40 000BP. The oldest date yet obtained for the Esperance area is about 13 000 years (Smith, 1993).

French and British maritime explorers landed along the south coast of Western Australia in 1792 and 1802, respectively. Contact with Aboriginal people occurred but sexual relationships are unlikely. In 1826 a military outpost was set up at Albany on the south coast of Western Australia, approximately 400 kilometres west of Esperance. Some marriages between British soldiers and Aboriginal women may have occurred but this was probably limited. From the 1830s or earlier, there was intimate contact between whalers and sealers and Aboriginal people along the south coast of Western Australia. The crews of the whaling and sealing ships were from widely differing origins, probably including British, various European-American, African American, Filipino and perhaps Malay, Chinese, Russian and others. In some cases, these men brought with them Aboriginal women from Kangaroo Island in South Australia, Tasmania and the islands of the Bass Straight, all to the east.

Between the 1840s and the 1870s, agriculture spread eastward along the south coast. Although the non-Aboriginal population was small and scattered there was undoubtedly intermarriage between Aboriginal women and non-Aboriginal men at various places. Esperance town was settled in 1866 and grew as a port servicing the gold rush centres to the north from the early 1890s. Aboriginal and non-Aboriginal people mixed in practice, if not officially acknowledged. As with the neighbouring Ngadju, most of the Esperance Nyungar population was of mixed-descent by the middle of the twentieth century.

S02 References

Allen, H., 1974. The Bagundji of the Darling basin: cereal gatherers in an uncertain environment. World Archaeol. 5, 309–322. doi:10.1080/00438243.1974.9979576

Bates, D., 1985. The native tribes of Western Australia, in: White, I. (Ed.), The Native Tribes of Western Australia. National Library of Australia, Canberra.

Bellwood, P. 2014 First migrants: ancient migration in global perspective. John Wiley & Sons.

Bowern, C., 2009. Reassessing Karnic: A Reply to Breen (2007). Aust. J. Linguist. 29, 337–348. doi:10.1080/07268600903232733

Chase, A., 1981. "All Kind of Nation": Aborigines and Asians in Cape York Peninsula. Aborig. Hist. 5, 7–19.

Dixon, R.M.W., 1977. A Grammar of Yidiny. Cambridge University Press, Cambridge.

Green, N., 1983. Desert school. Fremantle Arts Centre Press, South Fremantle.

Haddon, A.C., 1901a. Reports of the Cambridge Anthropological Expedition to Torres Straits, vol. 1, General Ethnography. Cambridge University Press, Cambridge.

Haddon, A.C., 1901b. Reports of the Cambridge Anthropological Expedition to Torres Straits, vol. 5, Sociology, Magic and Religion of the Western Islanders. Cambridge University Press, Cambridge.

Horton, D. 1994 Map Aboriginal Australia, in Horton D (gen. ed.) *The encyclopaedia of Aboriginal Australia*, The Australian Institute of Aboriginal and Torres Strait Islander Studies, Canberra.

Hercus, L.A., 1987. Linguistic diffusion in the Birdsville area, in: A World of Language: Papers Presented to Professor S.A. Wurm on His 65th Birthday.

Hercus, L.A., 1986. The Baagandji language. PL, Canberra.

Macknight, C.C., 1986. Macassans and the Aboriginal Past. Archaeol. Ocean. 21, 69–75.

Macknight, C.C., 1972. Macassans and Aborigines. Oceania 42, 283–321.

McCarthy, F.D., 1939a. "Trade" in Aboriginal Australia, and "Trade" Relationships with Torres Strait, New Guinea and Malaya, part 1. Oceania 9, 405–438.

McCarthy, F.D., 1939b. "Trade" in Aboriginal Australia, and "Trade" Relationships with Torres Strait, New Guinea and Malaya, part 2. Oceania 10, 171–195.

McConnel, U., 1934. The Wik-Munkan and Allied Tribes of Cape York Peninsula, N.Q. Oceania 4, 310–367.

McConvell, P., 1996. Backtracking to Babel: the chronology of Pama-Nyungan expansion in Australia. Archaeol. Ocean. 31, 125–144.

Meston, A., 1986. Report on the Aboriginals of Queensland. Queensland, Government Printer, Brisbane. online version <http://archive.aiatsis.gov.au/removeprotect/92163.pdf> on 31.10.2013

National Native Title Tribunal Geospatial Analysis & Mapping Branch, 2004, <http://www.nntt.gov.au/searchRegApps/NativeTitleRegisters/RNTC%20Extracts/NC1997_032/NC1997_032%201.%20Map%20of%20External%20Boundary.pdf> on 11.7.2015.

Neville, A.O., 1935. Aborigines Department annual report. Western Australia Government Printer, Perth.

O'Connor, S., Veth, P., 1996. A preliminary report on recent archaeological research in the semi-arid/arid belt of Western Australia. Aust. Aborig. Stud. 2, 42–50.

Roberts, R., Spooner, N., Jones, R., Cane, S., Olley, J., Murray, A., Head, J., 1996 Preliminary luminescence dates for archaeological sediments on the Nullarbor Plain, South Australia. Australian Archaeology, 7-16.

Russell, D., 2004. Aboriginal-Makassan interactions in the eighteenth and nineteenth centuries in northern Australia and contemporary sea rights claims. Australian Aboriginal Studies 1, 3-17.

Smith, M.A., 1996. Prehistory and human ecology in central Australia: an archaeological perspective, in: Morton, S.R., Mulvaney, D.J. (Eds.), Exploring Central Australia: Society, the Environment and the 1894 Horn Expedition. Surrey, Beatty and Sons, Sydney, pp. 61–73.

Smith, M., Fankhauser, B., Jercher, M., 1998. The changing provenance of red ochre at Puritjarra rock shelter, Central Australia: Late Pleistocene to present. Proc. Prehist. Soc. 64, 275–292.

Smith, M., Prescott, J.R., Head, M., 1997. Comparison of 14 C and luminescence chronologies at Puritjarra rock shelter, central Australia. Quat. Sci. Rev. 16, 299–320.

Smith, M.V., 1993. Recherche a l'Esperance: a prehistory of the Esperance region of south-western Australia. unplished dissertation, University of Western Australia, Perth.

South Australia Museum Archives <http://archives.samuseum.sa.gov.au/tindaletribes/barkindji.htm> on 11.7.2015, 2000.

Tacon, P., May, S.K., 2013. Rock art evidence for Macassan - Aboriginal contact in northwestern Arnhem Land, Griffith University. Australian National University E Press. <http://epress.anu.edu.au/apps/bookworm/view/Macassan+History+and+Heritage/10541/ch08.xhtml> on 5.11.2013

Tindale, N.B., 1966. Journal of a trip to Western Australia in search of Tribal Data. unpublished, S. Aust. Mus. AA 338/1/27.

Tindale, N.B., 1938. Harvard and Adelaide Universities Anthropological Expedition genealogies. S. Aust. Mus. AA 346/5/3.

Turney, C.S., Bird, M.I., Fifield, L.K., Roberts, R.G., Smith, M., Dortch, C.E., Grün, R., Lawson, E., Ayliffe, L.K., Miller, G.H., 2001. Early human occupation at Devil's Lair, southwestern Australia 50,000 years ago. Quat. Res. 55, 3–13.

van Holst Pellekaan, S.M., Ingman, M., Roberts-Thomson, J., Harding, R.M., 2006. Mitochondrial genomics identifies major haplogroups in Aboriginal Australians. Am. J. Phys. Anthropol. 131, 282–294. doi:10.1002/ajpa.20426

Veth, P., 2000. Origins of the Western Desert language: convergence in linguistic and archaeological space and time models. Archaeol. Ocean. 35, 11–19.

Veth, P., 1989. Islands in the Interior: A Model for the Colonization of Australia's Arid Zone. Archaeol. Ocean. 24, 81.

Veth, P.M., 1993. Islands in the interior: the dynamics of prehistoric adaptations within the arid zone of Australia, in: Archaeological Series 3. International Monographs in Prehistory, Ann Harbor, Michigan.

von Brandenstein, C.G., 1980. Ngadjumaja: an Aboriginal language of south-east Western Australia. Institut für Sprachwissen-schaft der Universität Innsbruck, Innsbruck.

S03 Sample location and collection, DNA extraction, array genotyping, whole-genome sequencing and processing

Simon Rasmussen, Anders Bergström, Anna-Sapfo Malaspinas, Ashot Margaryan, Stephen J Oppenheimer, Sturla Ellingvåg, Andrea B Migliano, Francois-Xavier Ricaut

Sampling locations and historical context

The geographical positions and the archaeological context of the samples discussed below are shown in Figure 1 and Figure S03.1.

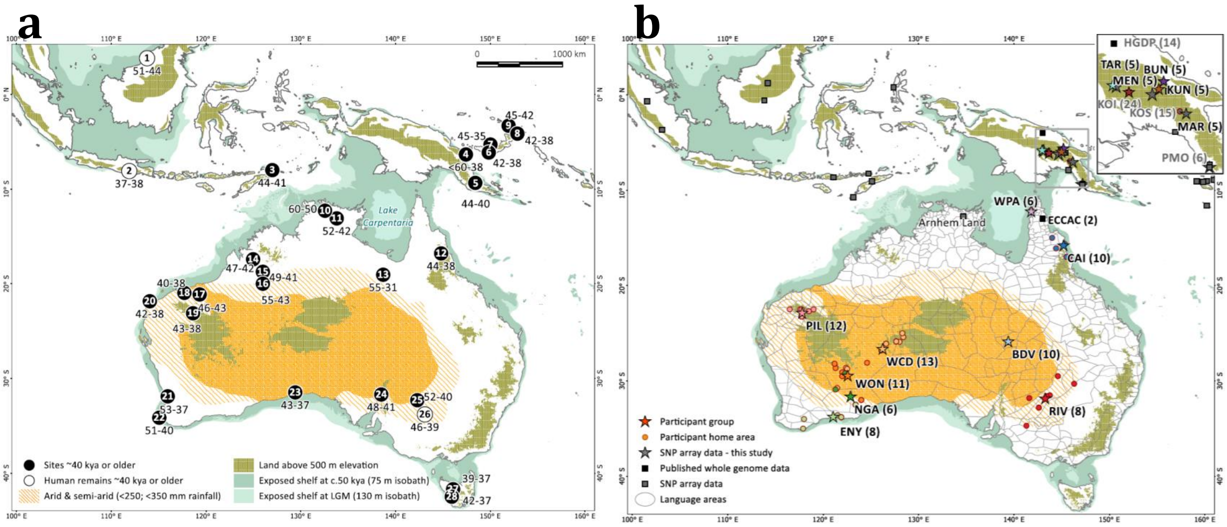


Figure S03.1 Sampling locations and historical context. This figure corresponds to Figure 1 in the main text but includes two subpanels. **a, Archaeological sites and human remains dated to ~40 kya or older in southern Sunda and Sahul.** The sites with dated human remains are shown as white circles and the archaeological sites as black circles. All dates are calibrated. See Allen and O'Connell (2014) and citations therein and Table S03.1. Lake Carpentaria, which covered a significant portion of the land bridge between Australia and New Guinea 11.5-40 kya and thus potentially acted as a barrier to gene flow, is also indicated. **b, Aboriginal Australians and Papuans samples used in this study.** The stars indicate the average sampling location for each group. SNP array data are shown in grey, and whole genomes are represented by coloured stars. Also shown on this figure are the coordinates for each participant - when available - computed as the mean between the parents' birth sites (filled circles). Publically available genetic data (see S04) used as a reference panel in this study shown as squares. The grey boundaries correspond to territories defined by the language groups provided by the Australian Institute of Aboriginal and Torres Strait Islander studies (Horton 1994). Sampled Aboriginal Australians self-identify primarily as: Yidindji and Gungandji from the Cairns region (CAI, 10, see also S02); Yupangati and Thanakwithi from northwest Cape York (WPA, 6), Wangkangurru and Yarluyandi from the Birdsville region (BDV, 10, 9 sequenced at high depth), Barkindji from southeast (RIV, 8); Pilbara area Yinhawangka and Banjima (PIL, 12), Ngaanyatjarra from central Australia (WCD, 13), Wongatha from WA's northern Goldfields (WON, 11), Ngadju from WA's southern Goldfields (NGA, 6); and Nyungar from southwest Australia (ENY, 8). Papuans include samples from the locations Bundi (BUN, 5), Kundiawa (KUN, 5), Mendi (MEN, 5), Marawaka (MAR, 5) and Tari (TAR, 5) - all whole genome sequenced. Additionally, we generated SNP array data for 45 Papuan samples including 24 Koinambe (KOI) and 15 Kosipe (KOS) - described before (Migliano et al. 2013) - and 6 individuals with highland ancestry sampled in Port Moresby (PMO).

Table S03.1 Ages associated to each archeological site or human remain shown in Figure S03.1. Site ages as listed by (Allen and O'Connell 2014), references cited therein, and (Clarkson et al. 2015)**.**

Number in Figure S03.1	Site	Uncalibrated basal age	Calibrated basal age 1 sigma	Calibrated basal age range 2 sigma	Other age est. (e.g. OSL)	Long.	Lat
1	Niah Cave	45.9±0.8	47.6±1.6	50.8-44.4	45-39	113.78	3.81
2	Wajak				37.4-28.5	111.9	-8.1
3	Jerimalai	38.26±0.6	42.48±0.85	44.2-40.8		127	-8
3	Lene Hara	38.21±0.61	42.41±0.86	44.1-40.7		127	-8
4	Bobongara				<60-38	147.43	-6.35
5	Ivane-Vilakauv	41.95±1.57	46.13±3.01	52.1–40.1		144	-5
5	Ivane-South Kov	40.3±0.96	44.16±1.59	47.3–41.0		144	-5
5	Ivane-Airport	39.84±0.91	43.81±1.47	46.7-40.9		144	-5
5	Ivane-AER	35.05±0.67	39.80±1.41	42.6–37.0		144	-5
6	Yombon	35.57±0.48	40.23±0.50	42.4–38.0		144	-5
7	Kupona na Dari				39.8±5.2	144	-5
8	Matenkupkum	35.41±0.43	40.01±0.96	41.9–38.1		144	-5
9	Buang Merabak	39.59±0.55	43.46±0.92	45.3–41.6		152	-3
10	Madjedbebe				44.2±4.7	132.56	-12.29
11	Nawarla Gabarnmang	42.87±1.45	46.95±2.75	52.4–41.5		133.8	-13.1
12	Ngarrabullgan	35.46+0.75/-0.69	40.10±1.40	43.7–37.9		144.82	-16.8
13	GRE 8	37.11±2.95	43.10±6.12	55.3–30.9		138.7	-19.05
14	Carpenters Gap	40.6±0.8	44.28±1.36	47.0–41.6		124.97	-17.42
15	Riwi	41.3±1.0	44.98±1.91	48.8–41.2		126	-18.74
16	Parnkupirti				48.7±6.0	126	-20
17	Yurlu Kankala	40.44±0.91	44.22±0.76	45.7–42.7		119.13	-21.13
18	Karriyarra	33.98±0.35	38.9 ± 0.6	40.0-37.7		118.46	-20.94
19	Djadjiling	35.75±0.55	40.37±1.16	42.7–38.0		118.63	-23.12
20	Jansz	35.23±0.45	39.83±1.11	42.0–37.6		114.07	-21.85
21	Upper Swan	39.5+2.3/-1.8	44.77±3.92	52.6–36.9		115.96	-31.89
22	Devils Lair	41.46+1.4/-1.19	45.44±2.57	50.6–40.3		115.07	-34.15
23	Allens Cave				39.8±3.1	129.42	-31.48
24	PACD H1	40.5±0.95	44.29±1.60	47.5–41.1		138.5	-31.7
25	Menindee	41.53±1.63	45.86±3.10	52.1–39.7		142.3	-32.3
26	Lake Mungo	38.1±1.1	42.56±1.92	46.4–38.7	42-40	143.1	-33.8
27	Parmerpar Meethaner	33.85±0.45	38.13±0.63	39.4–36.9		146.08	-41.7
28	Warreen	34.79±0.51	39.46±1.07	41.6–37.3		145.94	-42.49

Aboriginal Australian samples

For all the Aboriginal Australian samples, after obtaining written consent (S01), saliva samples were collected from the donors with the Oragene Discover (OGR-500) collection kit (DNAgenotek, Canada). Genomic DNA was extracted from 500 µl aliquots of the samples using prepIT-L2P reagent (Oragene) for manual extraction following the manufacturer's instructions. DNA yield was measured using Qubit Fluorometer (Invitrogen). All extracts were sent to Macrogen (*http://www.macrogen.com/eng/*) for whole-genome sequencing. For each sample a Truseq Nano library kit (350 base pair inserts) was used to construct a single library. The libraries were sequenced on the Illumina HiSeq X platform (paired-end sequencing, 150 cycles).

DNA aliquots from the same extractions were also sent to AROS (*http://arosab.com/*) for array based genotyping. Samples were genotyped on the HumanOmniExpressExome BeadChip, containing 964,193 markers.

Papuan samples

Whole genome sequence (WGS) data

Sampling background

Samples of DNA from 25 Papuan Highlander individuals were selected from a larger collection of samples collected between 1980 and 1990 as part of a collaboration between the Institute of Medical Research (IMR) in Papua New Guinea and the University of Oxford in the United Kingdom. These two institutions have a long history of collaboration, dating from before the time of these particular sample collections to the present day, working with the indigenous groups investigating population genetics and susceptibility to diseases including thalassemia and malaria (Oppenheimer et al. 1984, 1987 and Flint et al. 1985). The regions of the New Guinea Highlands that were the focus of this study are intensively cultivated and constitute the most densely populated areas in Papua New Guinea. At the time of sampling this region had remained relatively isolated and had undergone little inward migration from the lowland regions of Papua New Guinea or from any foreign arrivals to the country. The final samples selected for whole-genome sequencing in this study comprise five Gende speakers from the Bundi area (BUN), five Kuman speakers from Kundiawa (KUN), five Angal speakers from Mendi (MEN), five Huli speakers from Tari (TAR) and five Angan speakers from Marawaka (MAR). All of these languages form part of the large and linguistically numerous Trans-New Guinea family.

Ethical Approvals and Consenting

The DNA was derived from blood samples that were all collected as part of a single expedition in 1984 which had regulatory approval from the Papua New Guinea Institute of Medical Research and the Papua New Guinea Medical Research Advisory Council (Official Ethical Board) and was endorsed by the Public Health Department. Ethical approval was also obtained from UK regulatory committees. All participants were verbally consented following a brief description of the work that would be undertaken using the sample that the participants provided. The description included the phrase 'this work will attempt to find out what was different between the blood of individuals from the Papuan Highlands and Coastal peoples' but was not necessarily limited to such a phrase dependent upon the further information required to facilitate the understanding of the project for each participant. All samples are anonymised and were coded at the time of collection according to the region of origin of the participant and whether both parents spoke the same language. The DNA samples have all been stored with approval from the IMR for future research. Contemporary approvals for this current work have been sought and received from the IMR and from the Oxford Tropical Research Ethics Committee (OxTREC).

DNA sequencing

DNA was extracted from the buffy coat fraction from 10 ml whole-blood using a salting-out technique. Whole-genome sequencing of the samples was performed at the Wellcome Trust Sanger Institute. A single library (650 base pair inserts) was constructed for each sample. The libraries were multiplexed and sequenced across several lanes on the Illumina HiSeq X platform (paired-end sequencing, 151 cycles).

SNP array data

Additionally, 45 Papuan samples, described below, were genotyped on the HumanOmniExpressExome BeadChip.

Port Moresby (PMO)

Six samples were collected during the field seasons of 2010 in Port Moresby. The samples were obtained with informed consent donor permission and authorization by local ethics committees and the scientific board of the Evolutionary Medicine Department at the University of Toulouse, France. Saliva samples were collected using the Oragene DNA Collection kit (*http://dnagenotek.com*) and DNA was extracted using prepIT-L2P reagent (Oragene). Subjects were surveyed for language affiliation, current residence, familial birthplaces, and a genealogy of four generations to establish lineage ancestry suggesting they all originated in the Highlands. A total of six samples were analysed representing six different ethnic groups or languages (Figure 1).

Koinambe (KOI)

A total of 24 KOI samples (Figure 1) were included in this study (Migliano et al. 2013). The expedition to Koinambe village in 2008 had as its main objective to find the Suitani pygmies described by (Vogel, A. A. 1953). They are small-scale horticulturalists and speak Kalam and Maring languages (Trans-New Guinea family). To find them, we followed the original route and map described by Vogel. In the location where these people were first found, there is now the Koinambe mission, which we used as a base for visiting the villages nearby (05°29′N 144°35′E). The villages around the mission are in the lower Jimi River area, which had an estimated 18,000 people. The main villages in the region spoke Kalam (Lewis, M.P. 2009), a language of the Western Highlands branch of the Trans-New Guinea family, although some of the villages, farther up the valley toward the provincial border with Madang, spoke Maring, a language also from the Western Highlands branch of the Trans-New Guinea family that is spoken more widely (Lewis, M.P. 2009).

Kosipe (KOS)

A total of 15 KOS samples (Figure 1) were used in this study. The first described group of pygmies in the Goilala District of the highlands of Central Province was the Mafulu (Fuyege speakers) (Williamson, R. W. 1912). Kosipe Mission and its surrounding villages, close to the described location where Williamson studied the Mafulu people, are inhabited by Fuyege and Tauade speakers both from the Goilalan branch of the Trans-New Guinea family (Lewis, M.P. 2009).They are small-scale horticulturalists. Sampling included members of the two linguistic groups, as they intermarried and inhabited the same villages.

New Zealand sample

The following cell DNA samples was obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research: NA04932. Detailed information including informed consent can be found at the website: *https://catalog.coriell.org/1/NIGMS*. Aliquots of DNA were sent to *http://www.macrogen.com/eng/* for whole-genome sequencing and AROS (*http://arosab.com/*) for array based genotyping (see details above).

Rapanui sample

Blood samples from Rapa Nui individuals were collected in 2008 (Thorsby et al. 2009), including sample P2077 sequenced in this study. These individuals were elders living on Rapa Nui and carefully selected from a wider population of elders. The selection of individuals was made in close collaboration with local historians, notable elders, scientists in connection with the Rapa Nui Museum, a nurse with knowledge among the elders and a few written sources. Each individual's genealogy was recorded with family lines going back as far as possible, in all cases before the latter half of the nineteenth century.

All participants were informed about the purpose of the study (specifically study population history with genetic data) with the help of an interpreter and gave their informed consent when contributing a blood sample. The project was approved by the local authorities on Rapa Nui, and specifically by the provincial governor. All samples were anonymized. Ethical approval for the project was also provided by The Regional Ethics Committee in Norway, while the project was evaluated by the National Committee on Health Research Ethics, Denmark (H-3-2012-FSP70). DNA was extracted as described in (Thorsby et al. 2009).

Aliquots of DNA were sent to *http://www.macrogen.com/eng/* for whole-genome sequencing and AROS (*http://arosab.com/*) for array based genotyping (see details above).

Processing of Australian, New Zealand and Rapanui modern genomes

BAM files with read alignments obtained from Macrogen were first converted to raw fastq files which were then processed from scratch as follows. Reads were trimmed using AdapterRemoval-1.5.4 (Lindgreen 2012) using the adapters AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC and AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAG. Additionally leading/trailing stretches of Ns and bases with quality 2 were trimmed from the reads, and only reads of 30 nucleotides or longer were kept. Reads were first aligned to the human reference genome build 37 using bwa-0.7.10 mem (Li 2013) and subsequently aligned using stampy-1.0.23 (Lunter and Goodson 2011) with the option to keep well-aligned reads from bwa. The stampy aligned BAMs were then processed using picard-1.127 (*http://picard.sourceforge.net*). Read groups were added, the reads sorted and merged at the library level and read duplicates were marked and merged at the sample level. The per sample alignments were realigned using GATK-3.3-0 (McKenna et al. 2010) using the Mills and 1000G gold standard as known indels. We finally removed all reads with a mapping quality lower than 30, recalculated the md-tags and added extended BAQs using samtools calmd. Depth and coverage were calculated using pysam (*http://code.google.com/p/pysam*) and BEDtools (Quinlan and Hall 2010). Statistics are available in S04. The sequence data and alignments for the Australian genomes are available under data access agreement.

Processing of Papuan highlander genomes

Reads were aligned with bwa mem and duplicates were marked, reads were then merged to sample level and another round of duplicate marking was applied. The same BAM improvement and filtering steps as described above were then applied. The sequence data and alignments for the Papuan genomes are available under data access agreement.

Genotyping and haplotype phasing

All genomes (including the publically available genomes described in S04) were genotyped individually using samtools-0.1.18 mpileup –C50 and bcftools-0.1.17 (Li and Durbin 2009). Calls from each genome were filtered for a minimum of 1/3 of the average sequencing depth of the sample and a maximum of 2 times the average depth, except for the mitochondrial genome which was filtered for a minimum of 10 and maximum of 10000 read depth. For males the X and Y chromosomes were filtered using half the autosomal threshold. Variants were subsequently filtered out if there were: (i) two variants called within 5 base pairs of each other, (ii) Phred posterior probability of less than 30 or (iii) strand bias or end distance bias of p<1e-4. Per individual calls were merged across all samples using GATK-3.3-0 CombineVariants (DePristo et al. 2011) and filtered for deviations from Hardy-Weinberg Equilibrium with p<1e-4 (Wigginton et al. 2005) (archaic genomes were excluded from this calculation). Comparison of the filtered genotype calls to the array genotypes generated for the Australian samples showed a very high concordance (mean across samples 99.97%, lowest 99.81%). We then put aside the archaic individuals and converted the biallelic variant sites to IMPUTE format for phasing. Phasing was performed in 5Mb windows with IMPUTE-2.3.2 (Howie et al. 2009) using the 1000 Genomes Phase3 reference panel and the options '-phase -no_remove -fill_holes' to ensure that variants not present in the reference panel were phased and kept. Because (at the time we phased the data) phased chromosome X haplotypes were not available for the 1000G Phase3 panel we used the 1000G phase 1 release 3 dataset for phasing this chromosome. We used the -Xpar option for the pseudo-autosomal regions 1-2699520 and 154931045-155260560 and the -chrX option for the remainder of chromosome X. Phased genotypes were converted to VCF format and were first merged with non-variant sites and thereafter with the archaic genotypes using GATK-3.3-0 CombineVariants. An overview of the heterozygosity for each individual genome included in the study is shown in S04.

S03 References

Allen J, O'Connell JF. 2014. Both half right: Updating the evidence for dating first human arrivals in Sahul. Aust. Archaeol.:86.

Clarkson C, Smith M, Marwick B, Fullagar R, Wallis LA, Faulkner P, Manne T, Hayes E, Roberts RG, Jacobs Z, et al. 2015. The archaeology, chronology and stratigraphy of Madjedbebe (Malakunanja II): A site in northern Australia with early occupation. J. Hum. Evol. 83:46–64.

DePristo MA, Banks E, Poplin RE, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas M., Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.

Flint J, Hill AV, Bowden DK, Oppenheimer SJ, Sill PR, Serjeantson SW, Bana-Koiri J, Bhatia K, Alpers MP, Boyce AP. 1985. High frequencies of alpha-thalassaemia are the result of natural selection by malaria. Nature 321:744-750.

Horton D. 1994. The Encyclopedia of Aboriginal Australia. Horton, D. Canberra: Australian Institute of Aboriginal and torres Strait Islander Studies

Howie BN, Donnelly P, Marchini J. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5:e1000529.

Lewis, M.P. 2009. Ethnologue: Languages of the World. Dallas, TX: SIL International.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio [Internet]. Available from: http://arxiv.org/abs/1303.3997

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Lindgreen S. 2012. AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res. Notes 5:337.

Lunter G, Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21:936–939.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Migliano A, Romero I, Metspalu M, Leavesley M, Pagani L, Antao T, Huang D-W, Sherman B, Siddle K, Scholes C, et al. 2013. Evolution of the Pygmy Phenotype: Evidence of Positive Selection from Genome-wide Scans in African, Asian, and Melanesian Pygmies. Hum. Biol. [Internet] 85. Available from: http://digitalcommons.wayne.edu/humbiol/vol85/iss1/12

Oppenheimer SJ, Weatherall DJ, Higgs DR, Barker J, Spark RA. 1984. Alpha thalassaemia in Papua New Guinea. The Lancet 323:424-426.

Oppenheimer SJ, Hill AV, Gibson FD, Macfarlane SB, Moody JB, Pringle J. 1987. The interaction of alpha thalassaemia with malaria. Transactions of the Royal Society of Tropical Medicine and Hygiene 81:322-326.

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

Thorsby E, Flam ST, Woldseth B, Dupuy BM, Sanchez-Mazas A, Fernandez-Vina MA. 2009. Further evidence of an Amerindian contribution to the Polynesian gene pool on Easter Island. Tissue Antigens 73:582–585.

Vogel, A. A. 1953. Papuans and Pygmies. Available from: New York: Roy Publishers

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A Note on Exact Tests of Hardy-Weinberg Equilibrium. Am. J. Hum. Genet. 76:887–893.

Williamson, R. W. 1912. The Mafulu: Mountain People of British New Guinea. San Francisco: Macmillan

S04 Reference panels, relatedness and runs of homozygosity

Oscar Lao, Anders Bergström, Anna-Sapfo Malaspinas

Whole genome sequence (WGS) reference panel

We assembled relevant modern genomes from previously published studies for subsequent analyses: 11 individuals from Meyer et al. (2012), 14 individuals from Prüfer et al. (2014), including two Aboriginal Australian genomes derived from cell lines from the European Collection of Cell Cultures (ECCAC, *https://www.phe-culturecollections.org.uk/collections/ecacc.aspx*), 14 "HGDP-Papuan" individuals from Raghavan et al., (2015), five South Asian individuals from the 1000 Genomes Project (*http://www.1000genomes.org/*) and four Eurasian individuals from Raghavan et al. (2014). These genomes were mostly processed as described in S03 with the following exceptions: they were mapped using bwa-0.6.2 aln and re-aligned using GATK-2.2-3. The alignments for the 14 individuals from Prüfer et al. 2014 were downloaded from *http://cdna.eva.mpg.de/neandertal/altai/ModernHumans/*, and an additional filter of mapping quality 30 was applied. The 14 "HGDP-Papuan" individuals were processed as described in Raghavan et al. (2015).

In addition to the modern genomes we also used the ancient Aboriginal genome from Rasmussen et al. (2011) and the archaic Denisovan and Neanderthal genomes from Meyer et al., (2012) and Prüfer et al. (2014). These genomes were processed as described in S03 except they were mapped using bwa-0.6.2 aln with the seed disabled, to increase mapping sensitivity for ancient data with high error rates (Schubert et al. 2012). We also used publically available read alignments from a 45,000 year old human from Ust'-Ishim in western Siberia, downloaded from the ENA database under study accession number ERP006169.

Depth of sequencing coverage and estimated heterozygosity for the WGS data from this study and the reference panel are shown in Table S04.1 and Figure S04.1, respectively.

Table S04.1 Overview of whole genome sequenced individuals used in the study (M: Male, F: Female, DoC: Depth of coverage).

ID	Reference	Sex	Region	Population	DoC	Note
HG01583	Abecasis et al. 2012	M	South Asia	Punjabi	38	-
HG03006	Abecasis et al. 2012	M	South Asia	Bengali	48	-
HG03642	Abecasis et al. 2012	F	South Asia	Tamil	37	-
HG03742	Abecasis et al. 2012	M	South Asia	Telugu	45	-
NA20845	Abecasis et al. 2012	M	South Asia	Gujarati	40	-
DNK02	Meyer et al. 2012	M	Africa	Dinka	24	-
HGDP00456	Meyer et al. 2012	M	Africa	Mbuti	20	-
HGDP00521	Meyer et al. 2012	M	Europe	French	23	-
HGDP00542	Meyer et al. 2012	M	Oceania	HGDP-Papuan	22	same as 13748_2
HGDP00665	Meyer et al. 2012	M	Europe	Sardinian	20	-

HGDP00778	Meyer et al. 2012	M	East Asia	Han	22	-
HGDP00927	Meyer et al. 2012	M	Africa	Yoruba	27	-
HGDP00998	Meyer et al. 2012	M	South America	Karitiana	21	-
HGDP01029	Meyer et al. 2012	M	Africa	San	27	-
HGDP01284	Meyer et al. 2012	M	Africa	Mandenka	21	-
HGDP01307	Meyer et al. 2012	M	East Asia	Dai	24	-
SS6004467	Prüfer et al. 2014	M	Africa	Dai	33	-
SS6004468	Prüfer et al. 2014	M	Europe	French	38	-
SS6004469	Prüfer et al. 2014	M	East Asia	Han	32	-
SS6004470	Prüfer et al. 2014	M	Africa	Mandenka	33	-
SS6004471	Prüfer et al. 2014	M	Africa	Mbuti	33	-
SS6004472	Prüfer et al. 2014	M	Oceania	HGDP-Papuan	39	-
SS6004473	Prüfer et al. 2014	M	Africa	San	33	-
SS6004474	Prüfer et al. 2014	M	Europe	Sardinian	35	-
SS6004475	Prüfer et al. 2014	M	Africa	Yoruba	36	-
SS6004476	Prüfer et al. 2014	M	South America	Karitiana	32	-
SS6004477	Prüfer et al. 2014	M	Oceania	ECCAC-Australian	37	-
SS6004478	Prüfer et al. 2014	F	Oceania	ECCAC-Australian	38	-
SS6004479	Prüfer et al. 2014	F	South America	Mixe	36	-
SS6004480	Prüfer et al. 2014	M	Africa	Dinka	32	-
13733_8	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	17	-
13748_1	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	19	-
13748_2	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	19	same as HGDP00542
13748_3	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	19	-
13748_4	Raghavan et al. 2015	F	Oceania	HGDP-Papuan	19	-
13748_5	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	18	-
13748_6	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	18	same as SS6004472
13748_7	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	19	-
13748_8	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	18	-
13784_1	Raghavan et al. 2015	F	Oceania	HGDP-Papuan	13	-
13784_2	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	18	-
13784_3	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	18	-
13784_4	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	14	-

13784_5	Raghavan et al. 2015	M	Oceania	HGDP-Papuan	14	-
BDV01	This study	F	Australia	Australian	78	Data access
BDV02	This study	M	Australia	Australian	75	Data access
BDV04	This study	F	Australia	Australian	70	Data access
BDV05	This study	M	Australia	Australian	72	Data access
BDV06	This study	F	Australia	Australian	70	Data access
BDV07	This study	F	Australia	Australian	70	Data access
BDV08	This study	M	Australia	Australian	70	Data access
BDV09	This study	F	Australia	Australian	74	Data access
BDV10	This study	M	Australia	Australian	72	Data access
CAI01	This study	M	Australia	Australian	84	Data access
CAI02	This study	M	Australia	Australian	74	Data access
CAI03	This study	F	Australia	Australian	77	Data access
CAI04	This study	F	Australia	Australian	71	Data access
CAI05	This study	M	Australia	Australian	80	Data access
CAI06	This study	M	Australia	Australian	78	Data access
CAI07	This study	M	Australia	Australian	71	Data access
CAI08	This study	M	Australia	Australian	70	Data access
CAI09	This study	M	Australia	Australian	79	Data access
CAI10	This study	M	Australia	Australian	73	Data access
ENY01	This study	M	Australia	Australian	69	Data access
ENY02	This study	F	Australia	Australian	79	Data access
ENY03	This study	F	Australia	Australian	83	Data access
ENY04	This study	F	Australia	Australian	83	Data access
ENY05	This study	F	Australia	Australian	78	Data access
ENY06	This study	F	Australia	Australian	70	Data access
ENY07	This study	M	Australia	Australian	73	Data access
ENY08	This study	M	Australia	Australian	71	Data access
NA04932	This study	M	Oceania	NewZealand	57	Data access
NGA01	This study	F	Australia	Australian	74	Data access
NGA02	This study	F	Australia	Australian	52	Data access
NGA03	This study	F	Australia	Australian	73	Data access
NGA04	This study	M	Australia	Australian	75	Data access
NGA05	This study	F	Australia	Australian	56	Data access
NGA06	This study	F	Australia	Australian	63	Data access
P2077	This study	M	Oceania	RapaNui	59	Data access
PIL01	This study	M	Australia	Australian	58	Data access
PIL02	This study	M	Australia	Australian	61	Data access
PIL03	This study	F	Australia	Australian	56	Data access

PIL04	This study	F	Australia	Australian	64	Data access
PIL05	This study	M	Australia	Australian	68	Data access
PIL06	This study	M	Australia	Australian	59	Data access
PIL07	This study	F	Australia	Australian	63	Data access
PIL08	This study	M	Australia	Australian	72	Data access
PIL09	This study	M	Australia	Australian	58	Data access
PIL10	This study	F	Australia	Australian	61	Data access
PIL11	This study	M	Australia	Australian	57	Data access
PIL12	This study	M	Australia	Australian	63	Data access
RIV01	This study	F	Australia	Australian	73	Data access
RIV02	This study	F	Australia	Australian	62	Data access
RIV03	This study	F	Australia	Australian	69	Data access
RIV04	This study	M	Australia	Australian	62	Data access
RIV05	This study	F	Australia	Australian	72	Data access
RIV06	This study	M	Australia	Australian	66	Data access
RIV07	This study	M	Australia	Australian	70	Data access
RIV08	This study	F	Australia	Australian	66	Data access
WCD01	This study	M	Australia	Australian	62	Data access
WCD02	This study	M	Australia	Australian	59	Data access
WCD03	This study	M	Australia	Australian	61	Data access
WCD04	This study	M	Australia	Australian	52	Data access
WCD05	This study	M	Australia	Australian	60	Data access
WCD06	This study	M	Australia	Australian	58	Data access
WCD07	This study	F	Australia	Australian	61	Data access
WCD08	This study	F	Australia	Australian	64	Data access
WCD09	This study	M	Australia	Australian	59	Data access
WCD10	This study	F	Australia	Australian	63	Data access
WCD11	This study	M	Australia	Australian	57	Data access
WCD12	This study	M	Australia	Australian	59	Data access
WCD13	This study	M	Australia	Australian	67	Data access
WON01	This study	M	Australia	Australian	71	Data access
WON02	This study	F	Australia	Australian	101	Data access
WON03	This study	F	Australia	Australian	65	Data access
WON04	This study	F	Australia	Australian	58	Data access
WON05	This study	M	Australia	Australian	56	Data access
WON06	This study	F	Australia	Australian	60	Data access
WON07	This study	F	Australia	Australian	57	Data access
WON08	This study	F	Australia	Australian	52	Data access
WON09	This study	M	Australia	Australian	20	Data access

WON10	This study	M	Australia	Australian	50	Data access
WON11	This study	F	Australia	Australian	58	Data access
WPA01	This study	F	Australia	Australian	51	Data access
WPA02	This study	M	Australia	Australian	50	Data access
WPA03	This study	M	Australia	Australian	51	Data access
WPA04	This study	F	Australia	Australian	52	Data access
WPA05	This study	F	Australia	Australian	56	Data access
WPA06	This study	M	Australia	Australian	53	Data access
BUN01	This study	M	Oceania	Papuan	46	Data access
BUN02	This study	M	Oceania	Papuan	50	Data access
BUN03	This study	M	Oceania	Papuan	53	Data access
BUN04	This study	M	Oceania	Papuan	41	Data access
BUN05	This study	M	Oceania	Papuan	53	Data access
KUN01	This study	M	Oceania	Papuan	42	Data access
KUN02	This study	M	Oceania	Papuan	40	Data access
KUN03	This study	M	Oceania	Papuan	38	Data access
KUN04	This study	M	Oceania	Papuan	45	Data access
KUN05	This study	M	Oceania	Papuan	52	Data access
MEN01	This study	M	Oceania	Papuan	39	Data access
MEN02	This study	M	Oceania	Papuan	39	Data access
MEN03	This study	M	Oceania	Papuan	52	Data access
MEN04	This study	M	Oceania	Papuan	43	Data access
MEN05	This study	M	Oceania	Papuan	45	Data access
TAR01	This study	M	Oceania	Papuan	41	Data access
TAR02	This study	M	Oceania	Papuan	53	Data access
TAR03	This study	M	Oceania	Papuan	48	Data access
TAR04	This study	M	Oceania	Papuan	46	Data access
TAR05	This study	M	Oceania	Papuan	44	Data access
MAR01	This study	M	Oceania	Papuan	46	Data access
MAR02	This study	M	Oceania	Papuan	45	Data access
MAR03	This study	M	Oceania	Papuan	39	Data access
MAR04	This study	M	Oceania	Papuan	41	Data access
MAR05	This study	M	Oceania	Papuan	44	Data access
AusAboriginal	Rasmussen et al. 2011	M	Australia	Australian	6	Ancient, 6X
Avar	Raghavan et al. 2014	M	Caucasus	Avar	13	Data access
Manny	Raghavan et al. 2014	F	India	South Indian	16	Data access
Mari	Raghavan et al. 2014	M	Europe	Mari	12	Data access, high error rate

Tadjik	Raghavan et al. 2014	M	Central Asia	Tajik	16	Data access
DenisovaPinky	Meyer et al. 2012	F	Siberia	Denisova	24.3	Archaic
AltaiNea	Prüfer et al. 2014	F	Siberia	Neanderthal	40.8	Archaic

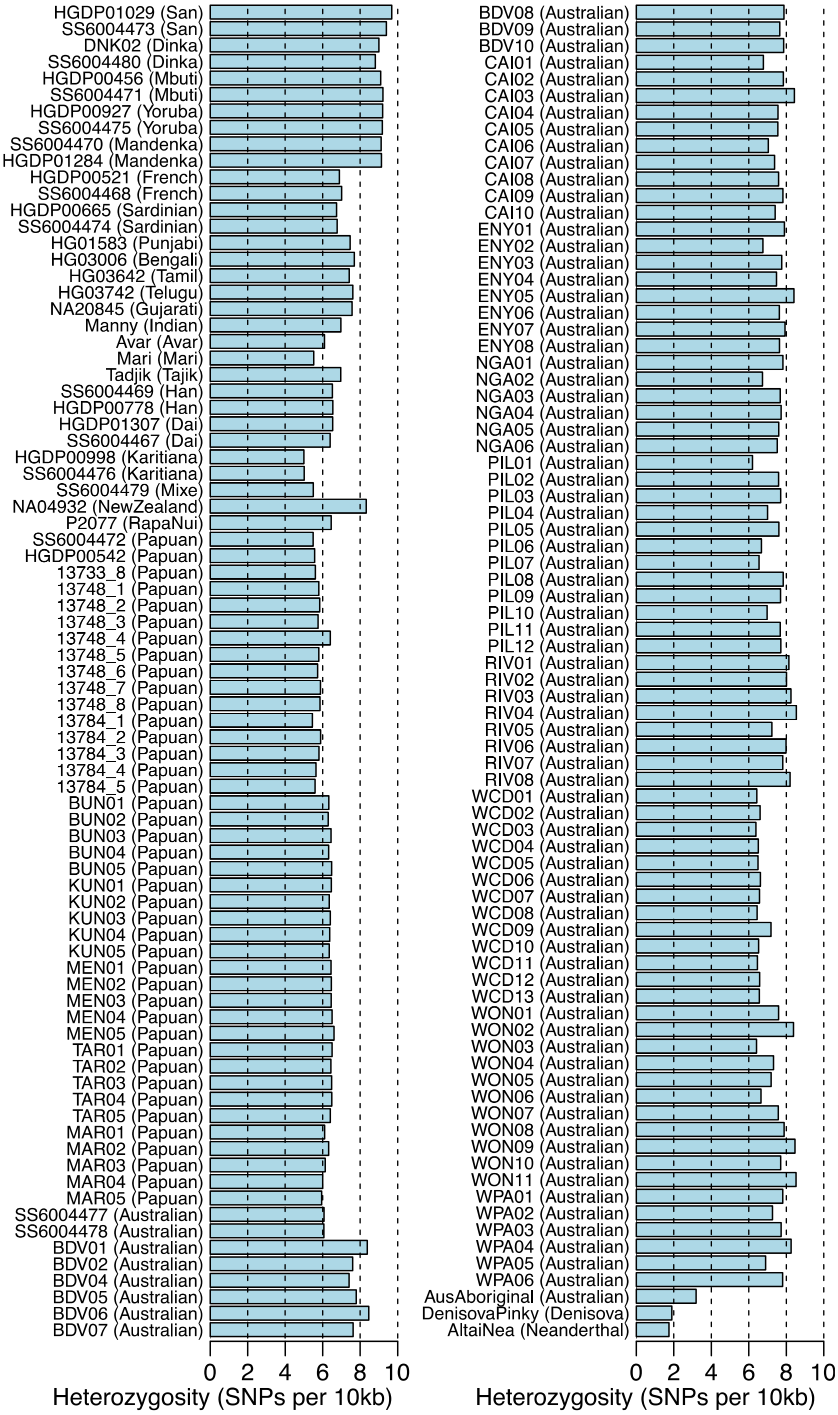


Figure S04.1 Overview of genome-wide heterozygosity as estimated directly from counts of the heterozygous and homozygous genotype calls made from whole-genome sequencing data (i.e., after the filtering described in S03). We note that the low heterozygosity in the ancient Australian Aboriginal genome (AusAboriginal) is an artifact of low sequencing coverage rather than reflecting genuinely low heterozygosity.

SNP array reference panel

In order to describe the genetic diversity of the 83 whole genome sequenced Australian samples on a worldwide level, we assembled SNP population data from previously published SNP microarray datasets. Given the limited number of shared SNPs among all the datasets, different subsets of those datasets were considered depending on the analyses and questions being addressed. Table S04.2 lists all the datasets that were considered.

SNP merging & cleaning

The merging procedure considered all the SNPs with known rsID and physical position (hg19) present in at least one of the databases. rsID's were recovered from the annotation file of each microarray platform. hg19 chromosome and physical positions were recovered from the snp123.txt file from the UCSC browser.

SNPs present in only one database were excluded. SNPs comprising A-T and C-G alleles were excluded from the merging to prevent flipping-strand problems. SNPs shared among at least two databases were merged after flipping when required and matching the alleles among databases.

Further SNP filtering, including checking for deviations to HWE in each population and global minor allele frequency (MAF) cutoffs, were performed on each subset of ascertained populations. We excluded SNPs out of HWE in at least one of the ascertained populations when the p-value was lower than $5*10^{-6}$ using the exact p-value as defined in Wigginton et al. (2005) or with a MAF < 0.01 considering all the individuals of the ascertained populations. The number of SNPs after cleaning differed depending on the analyses (see S05 and S13).

Individual merging & cleaning

After SNP merging, we checked for potential duplicate individuals genotyped in more than one database. To do so, for each pair of individuals from different databases, an Identity By State (IBS) distance was computed for the shared SNPs between the two databases. For a given proxy individual, individuals showing a smaller IBS distance than expected given the genetic variation present within each population were considered as potential duplicates among the two datasets. In most cases, the duplicate individual had the same population and individual identification label. In very few cases, we observed that the individuals shared the same individual identification label but a different acronym name for the population (for example, BantuSA in KRAUSE2014 and BantuSouthAfrica in HGDP). In those few cases, we "manually" curated the databases by changing the name of the population in one of the databases to match the name from the other database (for the previous example, BantuSA was set to BantuSouthAfrica). Duplicate individuals were removed by merging their genotypes into a single individual. When inconsistent, shared SNPs in duplicate individuals were set to missing.

Further identification of hidden relatedness within each population was performed by means of the KING software (Manichaikul et al., 2010), accounting for up to 3rd degree of relatedness. For each population, the set of unrelated individuals was obtained by applying a greedy algorithm. In a first step, the individuals within each population were sorted in descending order by the number of connections of relatedness to the other individuals of the population. In a second step, each individual from the sorted list was iteratively included in the final dataset if he/she was not related to any previously accepted individual in the list.

Table S04.2 SNP microarray datasets considered in this study

Label	**Snp array**	**Pops**	**Inds**	**SNPs**	**Reference**
HGDP	Illumina_650K	54	937	644,088	Li et al., 2008)
KRAUSE2014	Affymetrix_HumanOrigins	177	1426	523,871	(Lazaridis et al., 2014)
STONEKING2011	Affymetrix	14	115	868,068	(Reich et al., 2011)
STONEKING2013	Affymetrix	14	135	749,209	(Pugach et al., 2013)
STONEKING2015	Affymetrix_HumanOrigins	17	101	523,871	(Qin and Stoneking, 2015)
THISSTUDY	Illumina_HumanOmniExpressExomeBeadChips	12	113	917,356	NA
1000GENOMESAFFY	Affymetrix 6.0	26	3450	2,432,070	*http://www.1000genomes.org/*
1000GENOMES	Illumina Omni	21	2318	2,432,559	*http://www.1000genomes.org/*

Relatedness among Aboriginal Australian individuals

KING identified several first, second and third degree relationships between the sampled Aboriginal Australians. This allowed to the reconstruction of complex pedigree relationships among individuals (Figure S04.2).

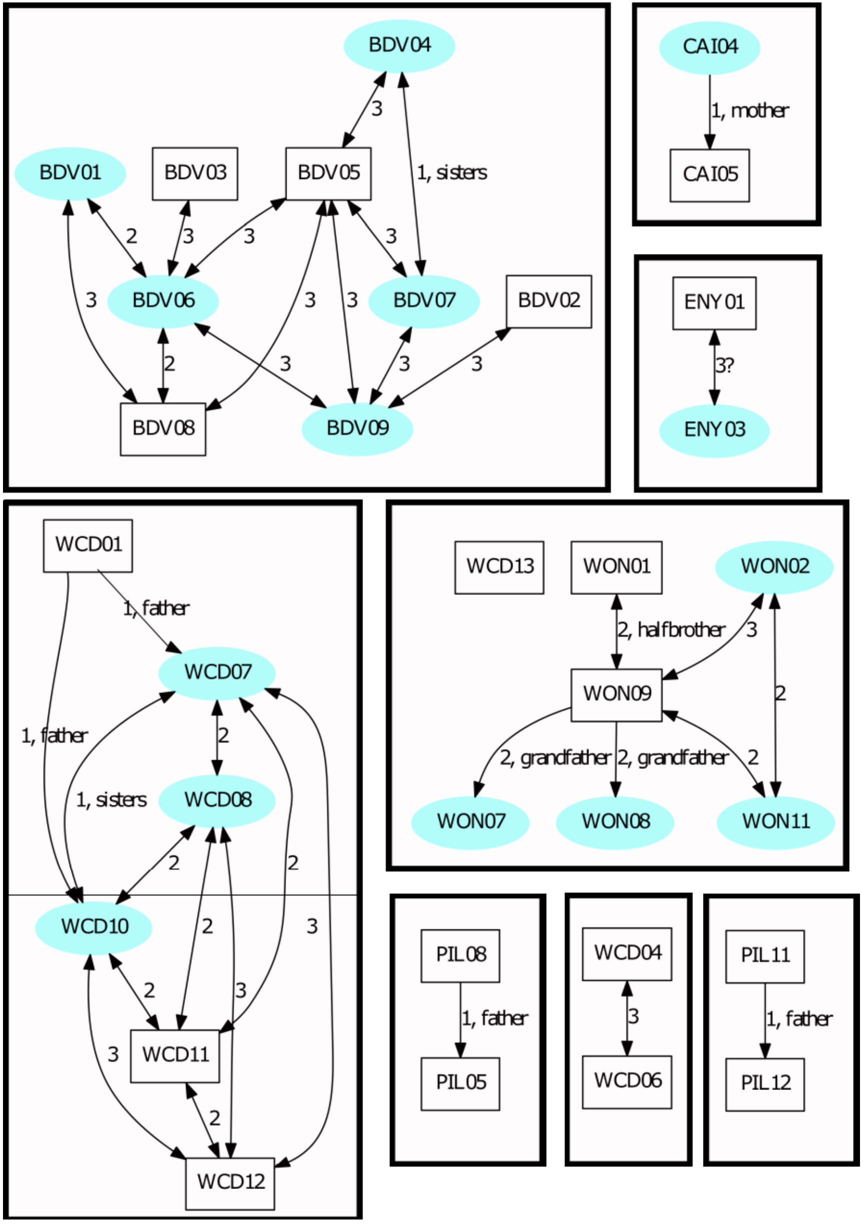


Figure S04.2 Relatedness among Aboriginal Australians. Graphical representation of the kinship relationships (first degree, second degree, third degree) identified by KING among Aboriginal Australian males (square) and females (circle).

The application of the algorithm for iteratively including non-related individuals identified 69 Aboriginal Australian samples which were used in subsequent analyses requiring unrelated samples.

Runs of homozygosity (ROHs)

We studied the spectrum of ROHs in the Aboriginal Australian populations and compared it with the one observed in other worldwide populations.

Methods

Quantification of the ROHs in the Aboriginal Australian populations was performed on the WGS data. Comparison with other worldwide populations was performed in the KRAUSE2014, STONEKING2015 and WGS datasets. Since we were interested in identifying long ROHs due to recent inbreeding, we excluded SNPs in strong LD using the command *--indep -50 -5* from Plink in both databases; furthermore, we only considered populations with at least five individuals. The merged worldwide dataset comprised 203,834 SNPs, 190 populations and 2,261 individuals. The WGS dataset comprised 540,194 LD pruned SNPs.

The individual estimation of ROHs adapted the protocol suggested by Pemberton et al. (2012). Briefly, we performed a genomic sliding window approach over the genome of each individual using a window size of 60 SNPs; within each window, we computed the LOD of autozygosity as defined in for the genotypes in window *k* (Broman and Weber, 1999):

$$LOD(j,k) = \sum_{i=j}^{k} log_{10}\left(\frac{\Pr(g_i|\text{autozygous at } i)}{\Pr(g_i|\text{not autozygous at } i)}\right)$$

Observed genotype	Probability that the segment is	
	Autozygous	Not Autozygous
AA	$(1-\varepsilon)p_A + \varepsilon p_A^2$	p_A^2
AB	$2\varepsilon p_A p_B$	$2p_A p_B$

We assumed a genotyping error rate ε of 0.001 (Pemberton et al., 2012). Following (Pemberton et al., 2012), missing genotypes were not taken into account in the estimation of the LOD of autozygosity. We modified the protocol proposed by (Pemberton et al., 2012) by using a set of unlinked markers rather than using the whole set of markers (Pemberton et al., 2012). This allowed us to estimate the expected null distribution of LOD of autozygosity under HWE within each population. We sampled the alleles at each locus under the assumption of random mating and computed the LOD values in the generated new individual genome. As a conservative measure of LOD of autozygosity, we set the autozygosity window cut off at the maximum value observed in the sampled genome. We further modified Pemberton's protocol by allowing a sliding jump of 10 SNPs between windows. By doing this, we reduced the amount of autocorrelation between successive windows, which we observed can affect the LOD of autozygosity distribution and enhance the periodicity effect observed in Figure 1 of (Pemberton et al., 2012). In order to get the ROHs, we concatenated adjacent windows showing LOD of autozygosity departing from the expected values under HWE.

ROHs fragments were classified by categories according to the observed frequency among individuals. We defined 11 ROHs categories: 0.146Mb, 0.246Mb, 0.371Mb, 0.533Mb, 0.753Mb, 1.087Mb, 1.647Mb, 2.735Mb, 5.857Mb, 10.0Mb and >10.0Mb.

The mean ROH length in population p was estimated by:

$$ROH(p) = \frac{\sum_{c=1}^{11} L_c * \frac{\sum_{i=1}^{n_p} F_{i,c}}{n_p}}{11}$$

where L_c is the fragment length of category c, $F_{i,c}$ is the number of ROH fragments in individual i from population p that belong to category c and n_p is the sample size of population p.

The observed distribution of average ROHs among populations was assumed to be a mixture of normal distributions. The maximum likelihood algorithm implemented in Mclust (Fraley et al., 2012) was used to identify the minimum number of normal distributions required to produce the observed data and to classify the populations according to their average ROH length.

“Admixture diversity” within each individual was estimated by computing the Entropy (*H*) based on ancestry proportions estimated in (S05):

$$H = \sum_{k=1}^{K} -f_k \log(f_k)$$

Results

We observe large differences in the distribution of ROH length among the Aboriginal Australian individuals and populations (Figure S04.3A). On average, WCD individuals tend to show longer ROH tracts compared to individuals from other Australian populations.

We observe a negative correlation between admixture diversity in the genome of each individual estimated by means of entropy and the mean of the length of ROHs observed (R = -0.6, p-value < 0.00005). This result suggests that the variability in ROH tracts length between Australian individuals could be mainly due to recent admixture with Europeans, East Asians and Papuans (S13).

Within a worldwide context, the average ROH length of each population ranged from 0.063Mb in the East Asian Xibo population to 28.93Mb in the Polynesian RenBel population; the population distribution showed multimodality (Figure S04.4) suggesting a mixture of ROHs patterns among worldwide populations.

The observed distribution of mean ROHs per population can be described by the mixture of four different normal distributions with unequal variance according to Mclust (log.likelihood = -476.65, BIC = -1011.739). Cluster four contains populations showing extreme ROHs patterns, such as RenBel or Onge. Cluster three includes populations showing an intermediate/large mean ROH length (>5Mb), whereas cluster one and two is defined by populations with low ROH length (Figure S04.5).

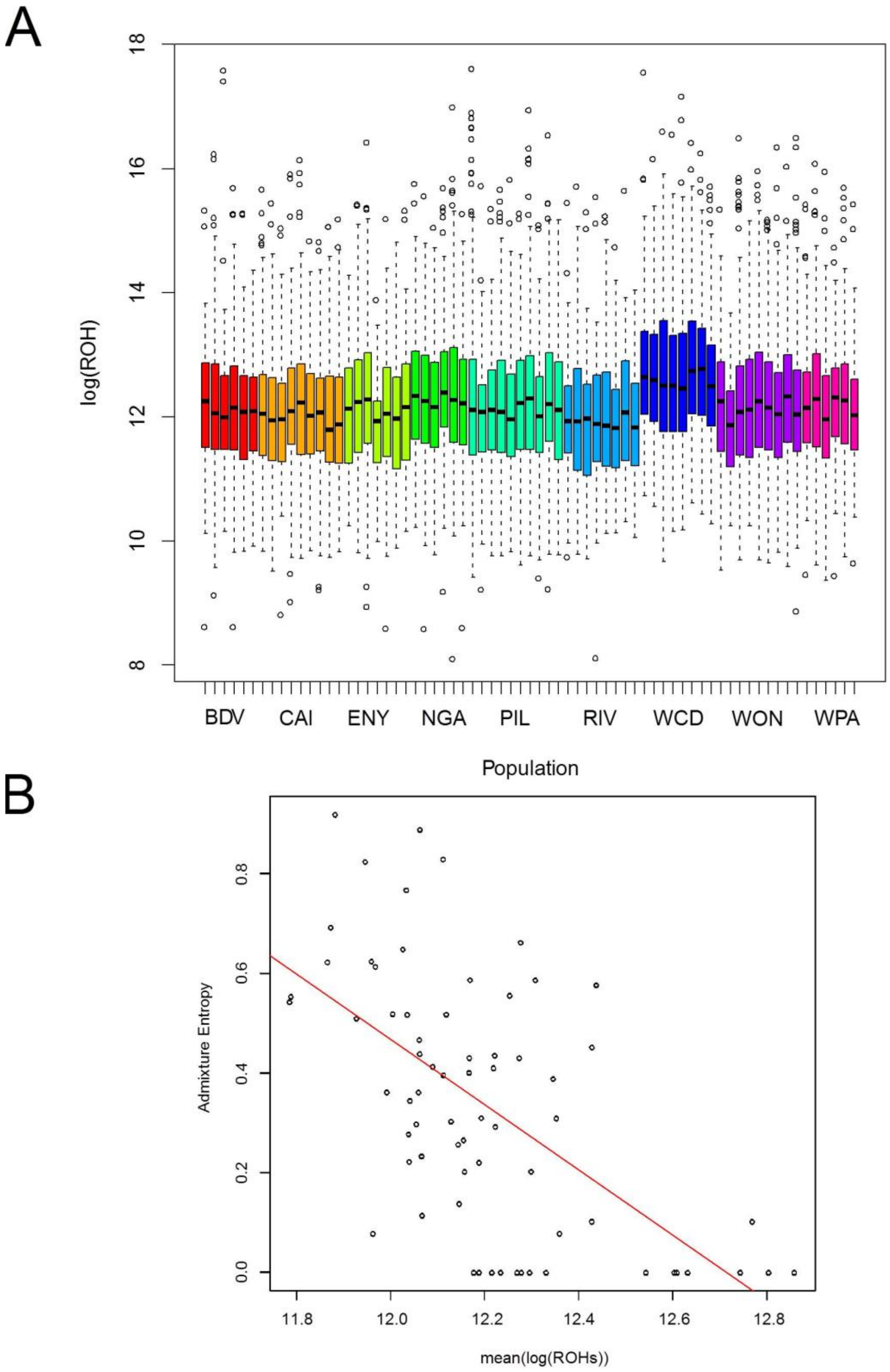


Figure S04.3 ROHs in the genome of Aboriginal Australian individuals. A) Boxplot of the distribution of ROHs in each Aboriginal Australian individual. B) Plot between the mean of the ROH distribution per individual and the amount of admixture heterogeneity, computed by means of Entropy.

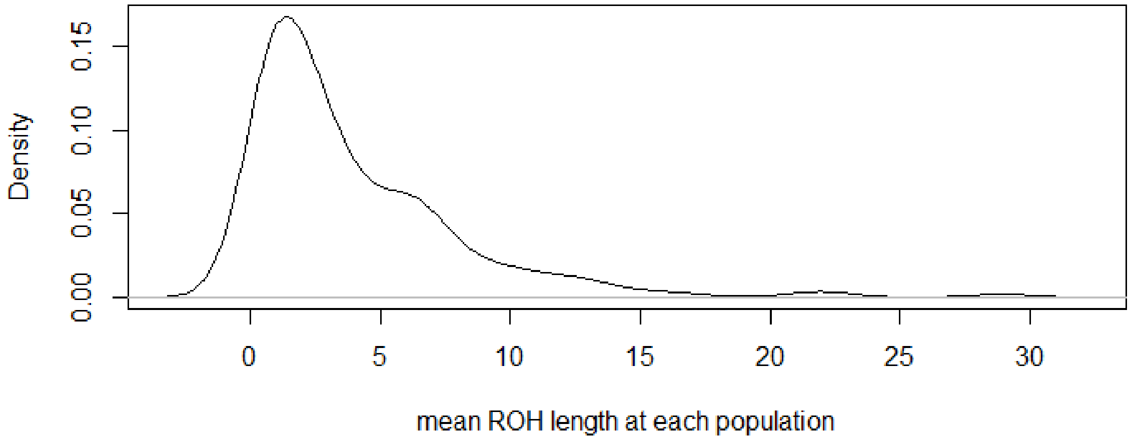


Figure S04.4 Distribution of mean ROHs (in Mb) per population on worldwide human populations.

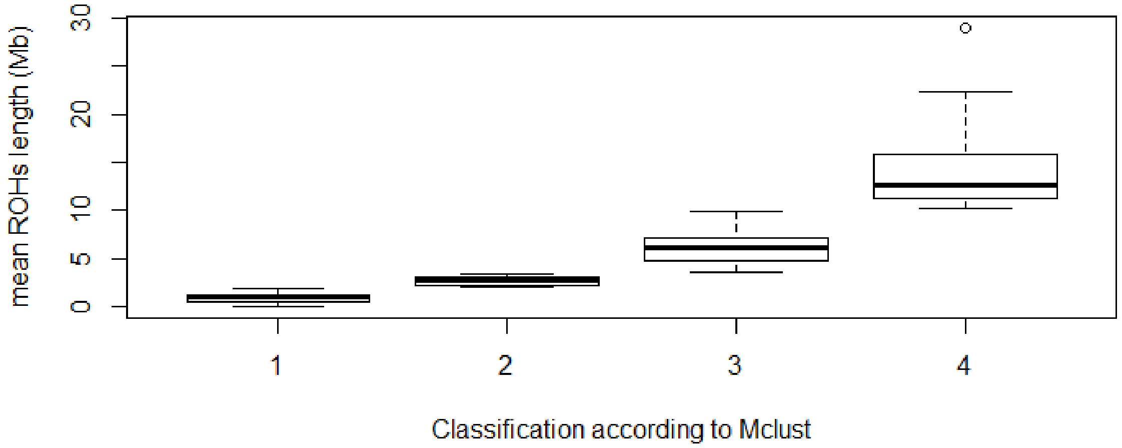


Figure S04.5 Boxplot with the four clusters identified by Mclust and the correspondence with the mean ROHs length (in Mb) of the populations.

The Australian populations showed a heterogeneous pattern of ROHs length. Australians Arnhem Land showed the largest mean proportion of ROH genome, with 8.844Mb, followed by WCD (6.14Mb), WON (3.5Mb) and PIL (3.1Mb). On the other extreme, CAI and WPA showed the lowest levels of mean genomic ROHs (0.99Mb and 1.36Mb respectively). At a worldwide level, the Australian populations were classified into the four different categories according to their mean ROHs length. Arnhem Land and WCD were classified into cluster three, which also includes Bougainville, HGDP-Papuan (KRAUSE2014), Papuan Central Province and Papuan Highlands (STONEKING2015), Jewish populations from KRAUSE2014 and most of the Solomon Islands populations from STONEKING2015, among others. ENY and WON were classified by Mclust into cluster two and the remaining Australian populations were classified into cluster one. The observed pattern of mean ROHs is concordant with the observation that Arnhem Land and WCD are the populations showing the lowest amount of recent admixture with Europeans and East Asians compared to other

Aboriginal Australian populations, yet it could also suggest traditionally small effective population sizes in Arnhem Land and WCD. The recent admixture in all the other populations increases the level of genetic variation and distorts the patterns of ROHs towards lower values.

S04 References

Broman, K.W., Weber, J.L., 1999. Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am. J. Hum. Genet. 65, 1493–1500.

Fraley, C., Raftery, A., Murphy, B., Scrucca, L., 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. Department of Statistics, University of Washington Technical Report No. 597.

Lazaridis, I., Patterson, N., Mittnik, A., et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513, 409–413.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., Myers, R.M., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100–1104.

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., Chen, W.-M., 2010. Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873.

Meyer, M., Kircher, M., Gansauge, M.-T., et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338, 222–226.

Pemberton, T.J., Absher, D., Feldman, M.W., Myers, R.M., Rosenberg, N.A., Li, J.Z., 2012. Genomic patterns of homozygosity in worldwide human populations. Am. J. Hum. Genet. 91, 275–292.

Prüfer, K., Racimo, F., Patterson, N., et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49.

Pugach, I., Delfin, F., Gunnarsdóttir, E., Kayser, M., Stoneking, M., 2013. Genome-wide data substantiate Holocene gene flow from India to Australia. Proc. Natl. Acad. Sci. U. S. A. 110, 1803–1808.

Qin, P., Stoneking, M., 2015. Denisovan Ancestry in East Eurasian and Native American Populations. Mol. Biol. Evol. 32 (10): 2665-2674.

Raghavan, M., Steinrücken, M., Harris, K., et al. 2015. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science. 349:aab3884.

Rasmussen, M., Guo, X., Wang, Y., et al. 2011. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334, 94–98.

Reich, D., Patterson, N., Kircher, M., et al. 2011. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am. J. Hum. Genet. 89, 516–528.

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J.F., AL-Rasheid, K.A., Willerslev, E., Krogh, A., Orlando, L., 2012. Improving ancient DNA read mapping against modern reference genomes. BMC Genomics 13, 178.

Wigginton, J.E., Cutler, D.J., Abecasis, G.R., 2005. A note on exact tests of Hardy-Weinberg equilibrium. Am. J. Hum. Genet. 76, 887–893.

S05 Linkage disequilibrium (LD) and population structure within Australia

Oscar Lao, Anna-Sapfo Malaspinas, Anders Bergström, Sankar Subramanian, Irina Pugach, Jade Y Cheng, Rasmus Nielsen

LD decay across worldwide populations

We investigated the patterns of linkage disequilibrium (LD) decay of the Aboriginal Australian populations and compared them with the ones observed in other worldwide populations.

Methods

The analysis of LD in the Aboriginal Australian samples and comparison with other worldwide populations was conducted on KRAUSE2014, STONEKING2015 (S03) and the whole genome sequence data. The merged dataset comprised 514,177 SNPs, 250 populations and 2,389 individuals.

The LD between a pair of SNPs was estimated with the HR statistic (Sabatti and Risch 2002), which is based on genotypic frequencies and does not require phased data.

In order to avoid biases due to differences in sample size and allele frequencies between populations (Rosenberg and Blum 2007), the HR statistic was computed from five randomly sampled individuals per population and only between pairs of SNPs with identical minimum allelic frequency.

We focused on a single parameter, *b*, which quantifies the rate of LD decay with genomic distance for a given population. For each population we estimated *b* by first estimating the mean HR statistic between pairs of SNPs and classifying the pair into bins of genomic distances d from 2.5kb up to 100kb. Next, we modeled *b* using an exponential function with intercept at 1 and a background LD of *c*:

$$HR(d) = (1 - c)e^{-b*\mathrm{d}} + c \qquad (1)$$

Using this model, we finally estimated *b* and *c* using non-linear least square regression by applying the *nls* function of the R stats package, with starting values at *c* = 0.1, b = 0.1, and using the Gauss-Newton algorithm.

In order to visualize dependence of the rate of LD decay on geography and to assess the estimated value of *b* for the Aboriginal Australian population in the worldwide context, two analyses were conducted. First, we created a density map based on the *b* estimates for all the populations using MapViewer 7.1.1767 (Golden) with the inverse of power algorithm for point interpolation. Second, a linear regression between the *b* value estimated for each population and the logarithm of the geodesic distance to Addis Ababa following main suggested migratory routes (Ramachandran et al. 2005) was computed.

Results and discussion

In agreement with previous results based on other estimators of genetic diversity such as heterozygosity (Ramachandran et al. 2005; Pemberton et al. 2013), the rate of the LD decay estimated by *b* decays with the logarithm of the distance to Addis Ababa (using main human migration paths, adjusted-R-squared: 0.3309, p-value: < 2.2e-16). Note that, given the observed recent genetic admixture with European, East Asian and Papuan for most of the Aboriginal Australian populations (see below) one could have expected lower *b* values for the admixed populations (Loh et al. 2013) (although see (Moltke et al. 2015)). Nevertheless, the Aboriginal Australian populations do not show any strong residual deviations in the linear regression suggestive of outlier points. One possible explanation is the large error in the estimation of LD due to low sample size used for these analyses (ten chromosomes per population).

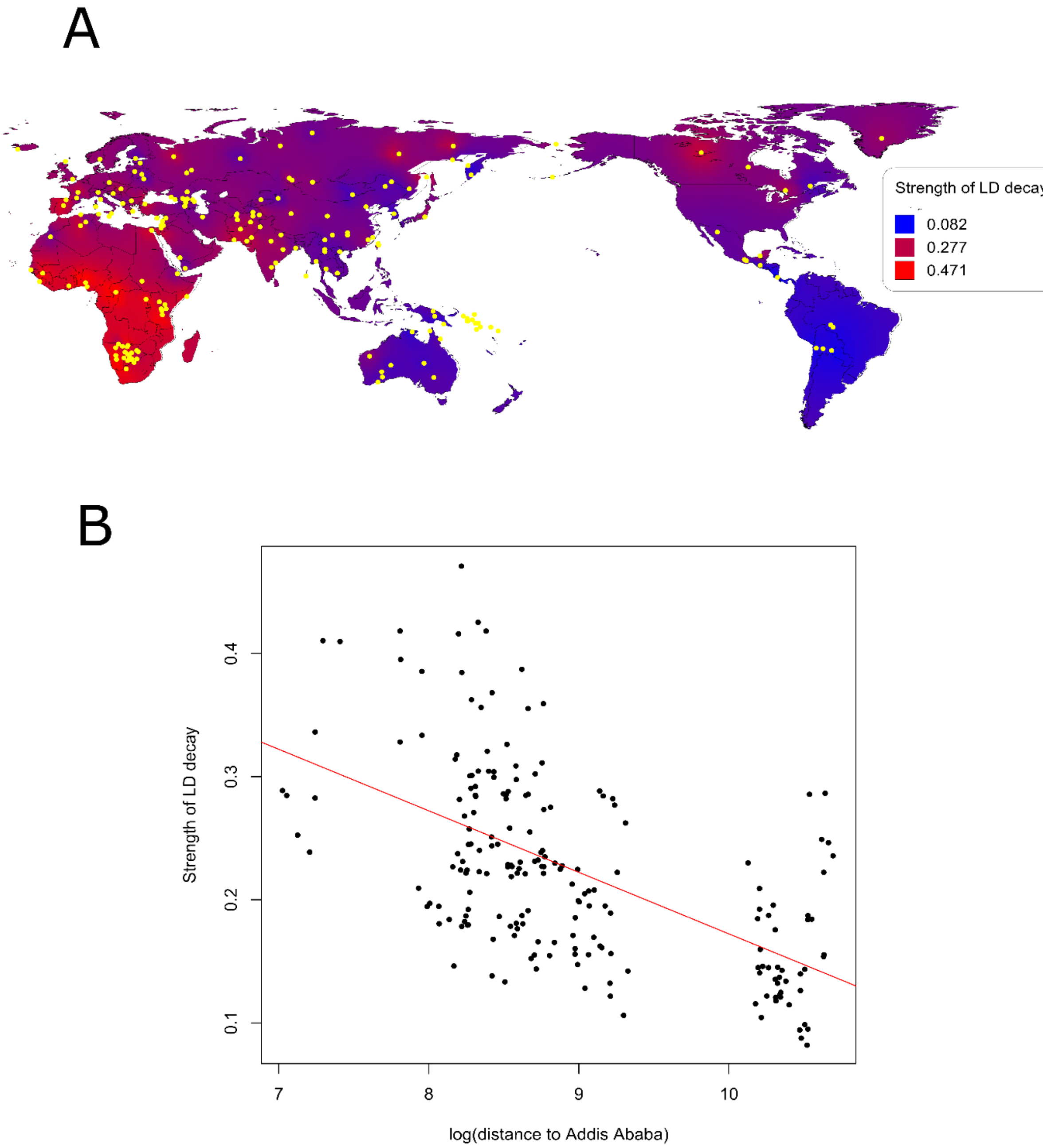


Figure S05.1 Rate of LD decay estimated by *b* with genomic distance for worldwide populations. A) Density plot of *b* in the different populations. Each dot represents a sampled population. B) Linear regression of *b* as a function of the log(distance to Addis Ababa) following the main routes of human migration.

Description of population substructure at the individual level

The presence of population substructure within the Aboriginal Australian individuals and their relationship with Eurasian populations was analyzed by means of different commonly applied algorithms for describing global ancestry (Wollstein and Lao 2015).

Databases

Two different databases were considered. In the first (DB1), we ascertained Eurasian and Oceanian populations from KRAUSE2014, STONEKING2015, STONEKING2013, STONEKING2011 and Papuans from THISSTUDY and merged them with WGS data from the 69 unrelated Aboriginal Australian individuals and the 25 WGS Papuan individuals (S04). Linkage Disequilibrium (LD) in the dataset was pruned with Plink (Purcell et al. 2007) using the option *--indep -50 -5*. After LD pruning, the number of considered SNPs was 54,971.

The second dataset (DB2) comprised WGS data from 1000 Genomes (British, Indian Telugu, Southern Han Chinese), HGDP Papuans, 69 Aboriginal Australians and 25 Papuans generated in this study. LD pruning was applied with the option --*indep -50 -5* in Plink. The number of individuals was 191; after LD pruning, the total number of SNPs was 566,359.

Methods: MDS and sNMF

The following individuals were ascertained from DB1, Australia: Arnhem Land, ECCAC, BDV, CAI, ENY, NGA, PIL, RIV, WCD, WON, WPA; East Asia: Cambodian, Dai, Han, Japanese, Naxi; Europe: English, French, Sardinian, Scottish, Spanish; India: Vishwabrahmin, Dravidian, Punjabi, Guaharati; New Guinea: HGDP Papuan (KRAUSE2014), Papuan: Central Province, Eastern Highlands, Gulf Province, Highlands (STONEKING2013 and STONEKING2015), PMO, KOI, KOS, BUN, KUN, MEN, TAR, MAR. An identical by state (IBS) distance between each pair of individuals was computed. We performed classical multidimensional scaling (MDS) on the generated distance matrix in R, including a constant (Cailliez 1983) to ensure positive eigenvalues.

In order to estimate ancestry proportions in the Aboriginal Australian individuals, we ran sNMF independently on two datasets. The first dataset considered a subset of DB1 including populations from (*i*) the Australian continent and surrounding regions; (*ii*) Europe (Orcadian, Scottish, English and Norwegian), India (Dravidian, Guaharati, Vishwabrahmin and Onge) and (*iii*) East Asia (Han-Chinese). The second dataset comprised all the populations from DB2.

For each dataset, sNMF (Frichot et al. 2014) was ran increasing K from 1 to 10. A cross-validation entropy (CV) statistic was estimated at each K and for each dataset the best K was identified as the one that minimized CV. For the ascertained K at each dataset, five different runs were performed using different starting seeds for the algorithm. The different runs were merged using CLUMPP (Jakobsson and Rosenberg 2007) using the greedy algorithm with default parameters.

Results: MDS

The first dimension of the MDS based on a subset of populations from DB1 explains 13.27% of the variation; it distinguishes Papuan populations from the rest. Individuals from Aboriginal Australian populations are distributed mainly between the axis defined by Papuan individuals and European individuals, suggesting recent admixture with the latter ones. Furthermore, some individuals from CAI and WPA show genetic affinities with populations from East Asia. The second dimension explains 5.35% of variation and distributes the individuals following a West to East longitudinal axis (Figure 2b**)**. In addition, we observed in the MDS that some Aboriginal Australian individuals were placed close to individuals of Indian ancestry. This result could corroborate previous studies suggesting a historical connection between the Indian subcontinent and the Aboriginal Australians (Pugach et al. 2013). However, given the observed patterns of recent admixture with Europeans and East Asians, and the intermediate position of Indian populations in the first dimension of the respective MDS, this could also be expected for Aboriginal Australians heavily admixed with European and East Asian populations.

Results: sNMF

In order to better describe the genetic variation in the geographic area comprising the Oceanian continent and geographically related populations, we ran sNMF with a set of populations derived from the DB1 database, including Oceania, Melanesia, Polynesia, Indonesia, Taiwan and South China. Furthermore, we included Indian and European individuals. The best number of ancestral components (*K*), as determined by cross-entropy, was seven. The main ancestry components matched Europe, India, Indonesia/South China/Taiwan, Papua, Australia, Melanesia and Polynesia. Aboriginal Australians appear as a mixture of mainly European and Australian ancestry components. Nevertheless, individuals from particular populations such as CAI, WPA and RIV, show additional Papuan and Indonesia/South China/Taiwan ancestry. A residual Melanesian ancestry background is observed in all the individuals (Figure 2a).

The sNMF analysis using WGS data and individuals from Europe, India, Han-Chinese, Papua and Australia identified five ancestral components. In agreement with the MDS, we observed that the Aboriginal Australians shared ancestry with Europeans, Papuans and East Asians (Figure S05.2). The ancestry proportions vary among individuals and populations, thus suggesting differences in the tempo and strength of admixture among the Aboriginal Australian groups (see also S14). In contrast, we quantified a similar Indian background (average = 3% per individual among all populations) for Aboriginal Australian individuals, with the exception of WCD where this component is lower and close to 1.1%.

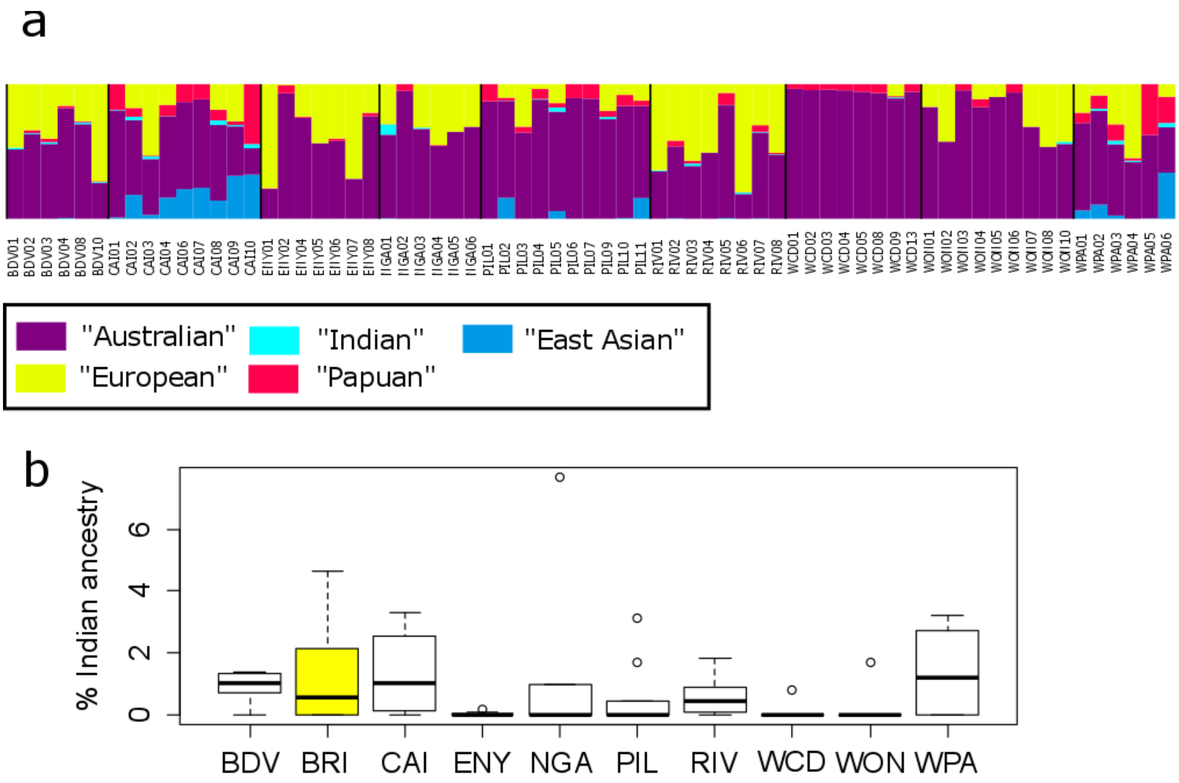


Figure S05.2 a) sNMF analysis based on *K*=5 components in DB2. Aboriginal Australians are mostly a mixture of four components. b) Boxplot of the percentage of Indian ancestry among the Aboriginal Australian populations and the British (BRI), highlighted in yellow.

Results: Indian gene flow in Aboriginal Australians

sNMF analyses

We noticed in the sNMF analysis that the Indian ancestry component is present in similar amounts in the proxy parental European population as in the Aboriginal Australian groups (Figure S05.2b). Thus, the observed Indian component in the Aboriginal Australians in each AA population could result from the recent admixture with European populations (Figure S05.2). In order to test this hypothesis, we performed a linear mixed model between the two ancestry components, treating the Aboriginal Australian populations as random effects. In this way, we model the linear regression between the Indian component and the European component by assuming that all populations share the same fixed effects for the slope and independent term of the linear regression, but also acknowledging that in addition each Aboriginal Australian population can have a different slope (random effect). We used the *lmer* function from the lme4 R package (Bates et al.) to implement the linear mixed model. The *R* command used was:

```
fm1 <- lmer(Indian ~ European + (European|Population),data = data)
```

Where *Indian* refers to each individual's Indian proportion of the genome, *European* is the corresponding European proportion and *Population* is the population of each individual.

The estimated fixed effects (that is, independent of which Aboriginal Australian population we are considering) for this model were:

*Indian_Ancestry = 0.0051 + European_Ancestry*0.09*

To estimate whether the inclusion of the “European ancestry” significantly improved the likelihood of the model, we run an ANOVA comparing this model with the nested model that considered only the intercept (so Indian and European ancestry were modeled as independent). We observed a Chi-square statistic of 49.88 (ANOVA p-value < 1e-11), thus suggesting that the Indian component estimated by sNMF in the Aboriginal Australians statistically significantly correlates with the European component; therefore, the Indian component we observe in the Aboriginal Australian samples is most likely due to the recent European admixture rather than ancient Indian admixture (Figure S05.3).

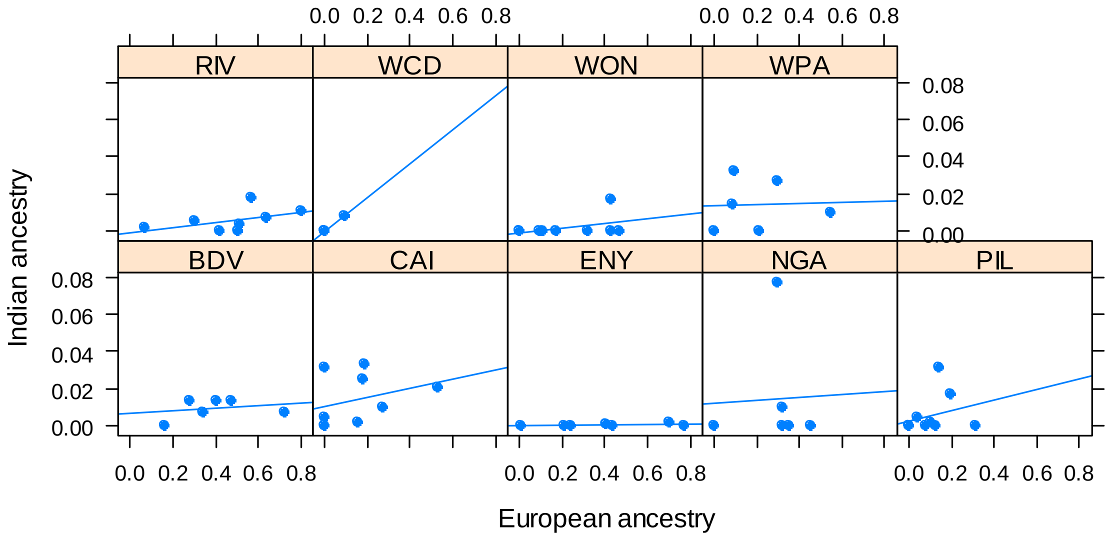


Figure S05.3 Plots of the percentage of Indian ancestry as a function of European ancestry for each Aboriginal Australian population.

f_3 and D-statistics

To further test the Indian gene flow hypothesis, we conducted formal tests of admixture using the f_3 and D-statistics on whole-genome sequencing data, as computed using ADMIXTOOLS version 3.0 ((Patterson et al. 2012) *https://github.com/DReichLab/AdmixTools*). In tests of the form f_3*(Single Aboriginal Australian; Source A, Source B)* we considered all possible pairs of source populations A and B from our set of generated and publically available sequenced genomes. A negative value for any such f_3 statistic would indicate admixture. We did not obtain a negative value in any test for 21 Aboriginal Australian individuals, irrespective of the source populations used. Furthermore, we performed D tests of the form *D(Yoruba, C; Aboriginal Australian, Papuan)*, using one genome per population and where C is any of French, Han, Gujarati, Telugu, Tamil, Punjabi or Bengali. A negative value of D would indicate gene flow from C to the Aboriginal Australian genome. Similarly to the f_3 results, we did not obtain a significantly negative value in any test for 23 Aboriginal Australian individuals, at a threshold of Z < -3 (Figure S05.4). At a threshold of Z < -2, the number is 18 individuals. To test whether the absence of admixture signatures could be caused by low power in single-sample tests, we pooled together the apparently non-admixed individuals. Neither the f_3 nor the D tests on these groups gave evidence for admixture, suggesting that the lack of admixture signatures is not due to low power. Thus, these formal tests of admixture show that there are several Aboriginal Australian individuals in our dataset that do not display signs of any external gene flow since the separation from Papuans. For the individuals that are admixed, the trends in the D-statistics show that it is European and East Asian, and not South Asian, ancestry that is driving the admixture signals (Figure S05.4).

This is consistent with the MDS and sNMF results above which show that the admixture signatures are explained by recent gene flow from European and East Asian populations or a combination of these. In summary, we find no support for the hypothesis of ancient gene flow from the Indian subcontinent to Australia.

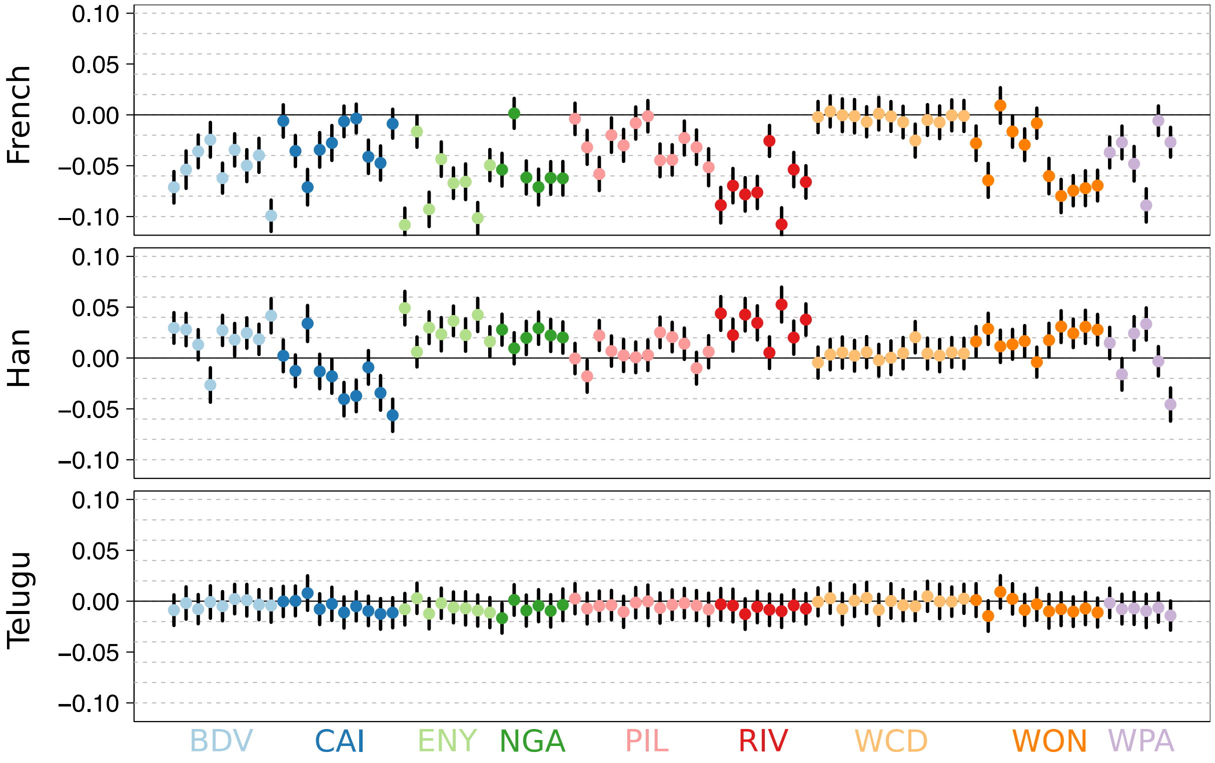


Figure S05.4 Tests for the presence of non-indigenous ancestry in Australian Aboriginal genomes using the D-statistic. Tests of the form *D(Yoruba, C; Aboriginal Australian, Papuan)* were carried out for different choices of C; here results are shown for French, Han and Indian Telugu. Vertical lines correspond to ±3 standard errors.

Description of population substructure at the population level

The description of population substructure among the Aboriginal Australian populations was estimated by means of a Neighbor-Joining (NJ) analysis (Saitou and Nei 1987) implemented in PHYLIP (Felsenstein) using a population pairwise F_{ST} (Weir and Cockerham 1984) distance matrix on European and East Asian masked tracts.

Methods

Weir and Cockerham' F_{ST} was computed between each pair of Aboriginal Australian populations using the masked data on European and East Asian ancestry (S06). Given the limited number of genotypes observed for certain SNPs, we decided to exclude SNPs based on minimum allele frequency (MAF) and minimum number of observed genotypes per population. In particular, we excluded SNPs that had a missing number of genotypes >80% in at least one of the populations and that had a MAF over all the populations <0.05. An unrooted NJ-tree was generated with PHYLIP. Statistical robustness of each branch was estimated by means of a bootstrap analysis (Felsenstein 1985). In particular, 1000 SNP datasets were generated by sampling with repetition at random the same number of markers as in the observed data. For each dataset, a F_{ST} distance matrix was computed between each pair of populations and a NJ-tree was generated. Branch robustness was computed as the number of times that the NJ-tree based on sampled SNPs showed the same branch as the tree obtained with the observed data.

Results

The obtained NJ-Tree is shown in Figure S05.5. We observe that the tree partitions reflect the geographic location of each population (S13).

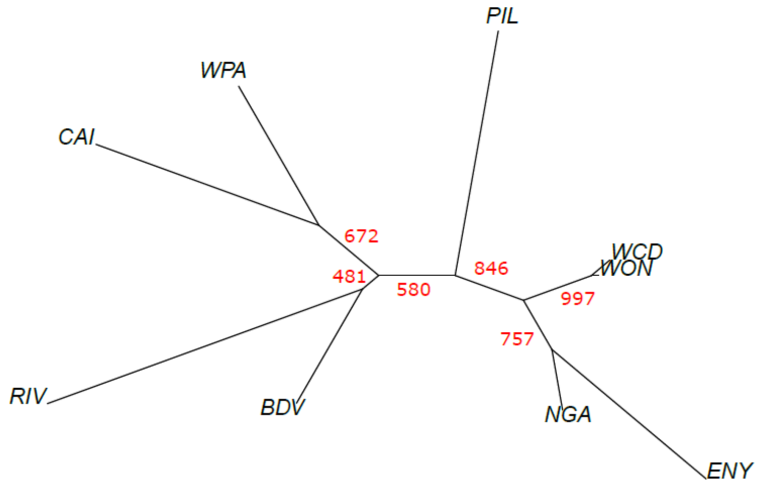


Figure S05.5 NJ-Tree based on Weir and Cockerham' F_{ST} distances between each pair of populations. The numbers in red indicate the number of times that the bootstrap procedure produced an identical branch as observed with the real data (out of the 1000 bootstrap replicates).

A new method combining admixture and covariance analysis

Method

Admixture Analysis

We estimate individual ancestries from multi-locus genotype datasets. We apply a model-based inference approach. The statistical model is used in existing software such as STRUCTURE (Pritchard et al. 2000), FRAPPE (Tang et al. 2005), ADMIXTURE (Alexander et al. 2009) or SPA (Yang et al. 2012). For the parameter inference, we develop the quadratic programming with active set, QPAS.

Likelihood Model

Under the assumption of Hardy Weinberg Equilibrium (HWE), the likelihood of assigning an observed genotype *g* in individual *i* at locus *j* to population *k* is function of the allelic frequency f_{kj} of the locus at *k* and the fraction of the genome of the individual q_{ik} that comes from that population. We thus consider the likelihood of the ancestral population proportions vector *Q* and their vector of allele frequencies *F* (i.e., the likelihood function $P_1(Q,F)$). In particular, if we denote *K* as the number of ancestry components, *I* as the number of individuals, and *J* as the number of polymorphic sites among the *I* individuals, then the probability of observing the genotypes is:

$$\ln[P_1(Q,F)] = \sum_{i}^{I}\sum_{j}^{J}\{g_{ij}\cdot\ln[\sum_{k}^{K} q_{ik}\cdot f_{kj}] + (2-g_{ij})\cdot\ln[\sum_{k}^{K} q_{ik}\cdot(1-f_{kj})]\}$$

Optimization

To estimate *Q* and *F*, we followed Newton's method. In general, we can approximate a function *F* with its second order Taylor expansion F_T. Each Newton update attempts to find the Δx such that the derivative of F_T with respect to Δx is zero. In our case, we need to satisfy a sequence of constraints while searching for Δx. Specifically, $\forall\Delta q_{ik}, q_{ik}+\Delta q_{ik}\in[0,1]$, $\forall\Delta f_{kj}, f_{kj}+\Delta f_{kj}\in[0,1]$, and $\forall\Delta q_{ik}\}$, $\sum_k\Delta q_{ik}=0$ because $\sum_k q_{ik}=1$.

To solve this inequality- and equality-constraint quadratic optimization problem, first we derive the first and second differentials for $\ln[P_1(Q,F)]$ with respect to values in *Q* and *F*, separately. Then we incorporate the active set algorithm (Murty 1988). A constraint is called active when its equality is satisfied and inactive when its strict inequality is satisfied (Nocedal and Wright 2006). An equality constraint is always active. To solve the equality problem defined by the active set and compute the Lagrange multipliers of the active set, we use the Karush-Kuhn-Tucker (KKT) approach (Karush 1939,Kuhn and Tucker 1951), a nonlinear programming generalization of the Lagrange multiplier method.

The active set algorithm operates by solving for equality quadratic subproblems. In each iteration it tries to find a better solution by walking along the active constraints. It deviates from the bounds when the Lagrange multipliers signal a better solution toward the feasible region. The maximum iterations is the number of inequality constraints. In the worst case, the algorithm walks along each inequality constraint once. We have $2K$+1 constraints for updating Q_i, each row in *Q*, and $2K$ constraints for updating F_j, each column in *F*, so for each update the cost is $\Theta(IK^2\cdot(1+2K)+JK^2\cdot 2K)$. We call this method QPAS.

This algorithm allows for improved optimization over methods such as ADMIXTURE. Indeed, we consistently find solutions with higher likelihood values than ADMIXTURE. It also allows for the estimation of a covariance matrix among populations, for which there are two applications that we use in this study. The first application is the estimation of population trees relating the history of the ancestry components (see Extended Data Figure 1 and below). The second is a new method for detecting selection that is described in detail in S16.

Covariance Analysis

We estimate the variances and covariances of the ancestral populations from the inferred allele frequencies and the multi-locus genotypes. We model the joint distribution of allele frequencies across all populations as a multivariate Gaussian (Pickrell and Pritchard 2012). For the parameter inference, we apply a black-box type of optimizer, the Nelder-Mead optimization, NM.

Likelihood Model

In our likelihood model, the variance of the multivariate Gaussian distribution is a product of two factors (Pickrell and Pritchard 2012). The first term $\mu_j(1–\mu_j)$ is site-specific, where μ_j is a vector of the allele frequency at site *j* at the ancestral populations. The second term Ω is constant across sites. Ω captures population variances and covariances. It is a symmetric positive-definite matrix.

$$P_2(f_j|\Omega,\mu_j) \sim N[\mu_j,\mu_j(1-\mu_j)\Omega]$$

Because of the symmetry of the Gaussian distribution, the system is under-determined. One unrooted tree corresponds to multiple different covariance matrices, i.e. rooted trees for which covariance matrices all induce the same probability distribution on the allele frequencies. To address this, we root the tree by defining one of the ancestral populations as the root. This corresponds to calculating the conditional probability of the data given the value observed in one of the populations, which can be arbitrarily chosen. We use the first population as the “root population”. Allele frequencies at other loci are replaced by the difference f_j' of the original values f_j, and the corresponding frequency in the first population f_{j0}. Ω′ is symmetric and of size (*K*–1)×(*K*–1).

$$\begin{aligned}
\ln[P_2(F)] &= \ln\{\prod_j^J[\frac{1}{\sqrt{|2\pi c_j\Omega'|}}\exp(-\frac{1}{2}\cdot f'^T_j\cdot(c_j\Omega')^{-1}\cdot f'_j)]\} \\
&= -\frac{1}{2}\cdot\sum_j^J\{K\cdot\ln(2\pi c_j)+\ln[\det(\Omega')]+\frac{1}{c_j}\cdot f'^T_j\cdot\Omega'^{-1}\cdot f'_j\} \\
\text{where } c_j &= \mu_j(1-\mu_j) \\
f'_j &= f_j - f_{j_0}
\end{aligned}$$

Optimization

We treat this likelihood function maximization as a black-box optimization problem and –as said earlier - consider a gradient-less optimization method: NM. NM is an iterative process that continually refines a simplex, which is a polytope of *D*+1 vertices in *D* dimensions (Nelder and Mead 1965). This process continues until the simplex collapses beyond a predetermined size, a maximum length of time expires, or a maximum number of iterations is reached. The amount of effect possible actions can have on the simplex is controlled by supplying to the algorithm coefficients for reflection, expansion, contraction, and shrinkage.

Data

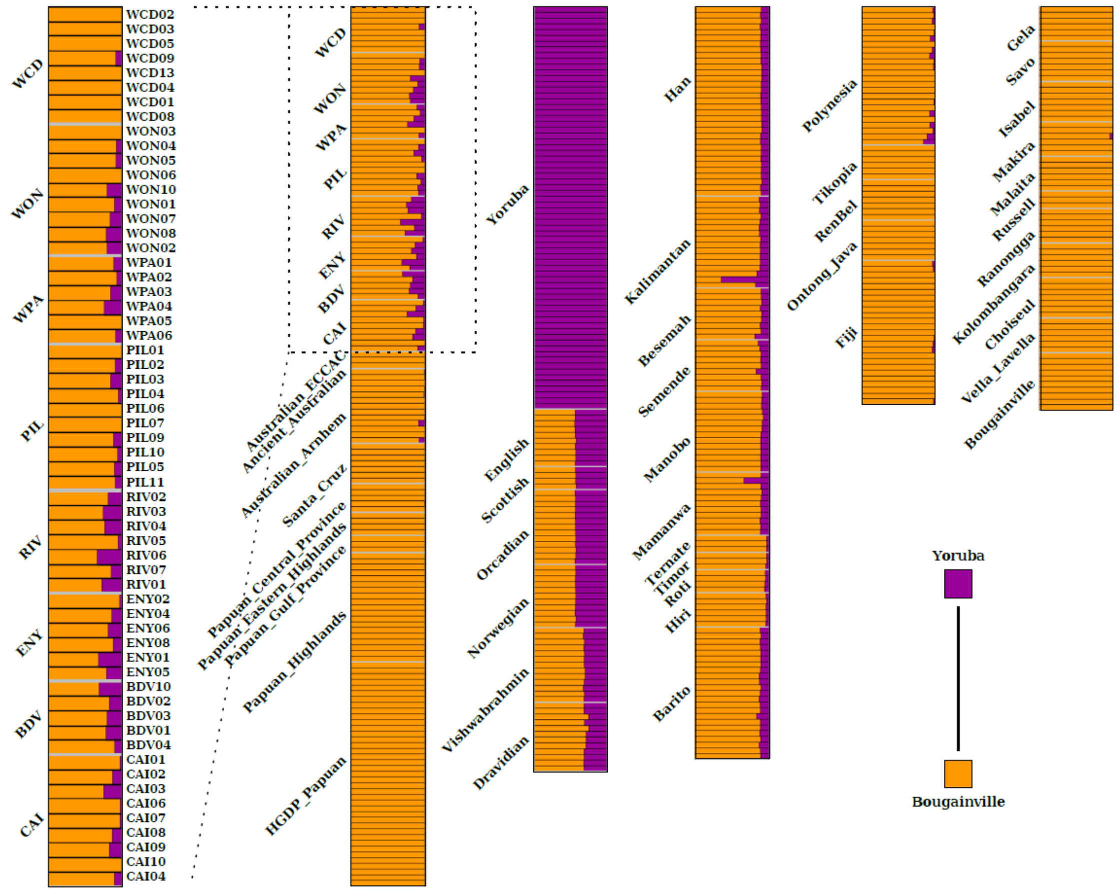
We ascertained proxy populations from KRAUSE2014, STONEKING2015, STONEKING2013, STONEKING2011 (which are genotyped in the Affymetrix platform and show a relatively large number of shared SNPs (S04)) and WGS data from the 69 unrelated

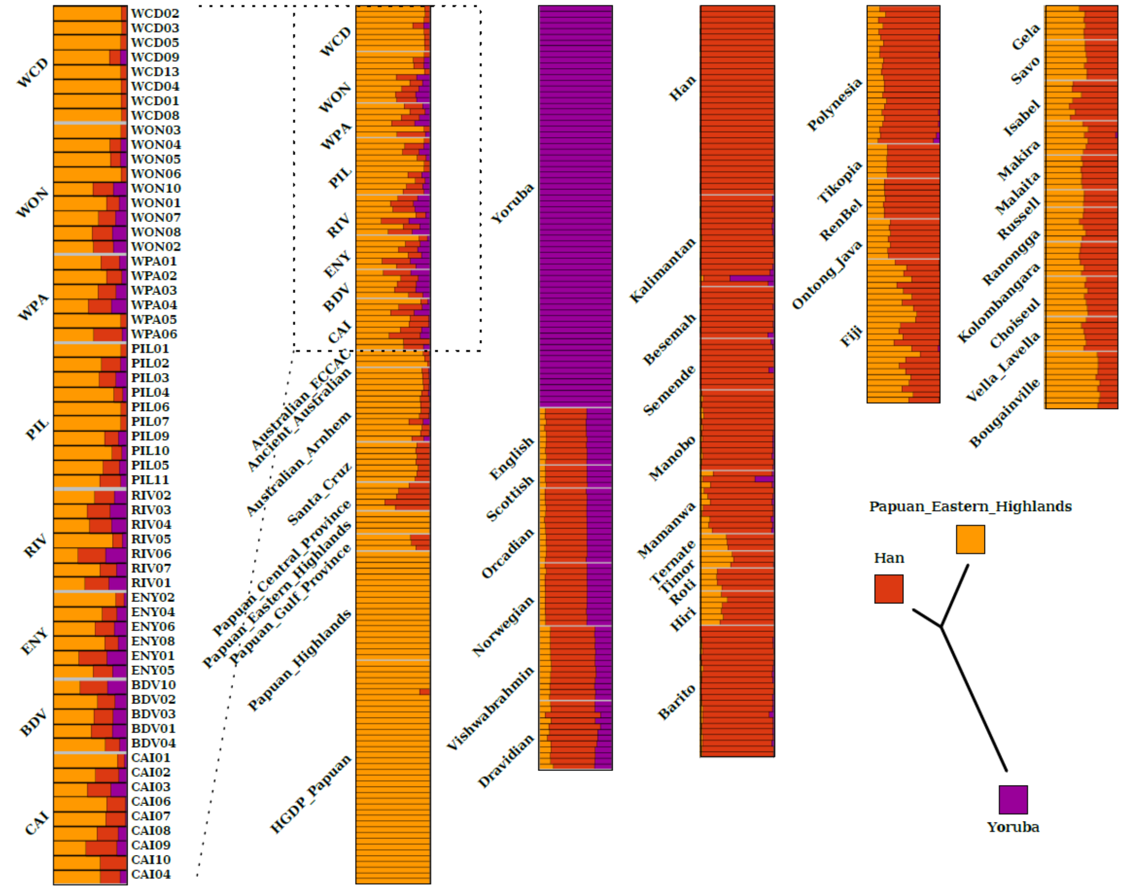
Aboriginal Australian individuals showing more than 50% of Aboriginal Australian ancestry (S06), covering Eurasian, Oceanian, Melanesian, Polynesian and African ancestry: Africa: Yoruba, Australia: Arnhem Land, ECCAC, BDV, CAI, ENY, NGA, PIL, RIV, WCD, WON, WPA; East Asia: Han; South-East Asia: Kalimatan, Besemah, Semende, Manobo, Mamanwa, Ternate, Timor, Roti, Hiri, Barito; Europe: English, Scottish, Orcadian, Norwegian; India: Vishwabrahmin, Dravidian; New Guinea: HGDP Papuan (KRAUSE2014), Central Province, Eastern Highlands, Gulf Province, Highlands (STONEKING2015) and Highlands (STONEKING2013); Melanesia: Bougainville, Vella Lavella, Choiseul, Kolombangara, Ranonga, Russell, Malaita, Makira, Isabel, Savo and Gela.
Linkage Disequilibrium (LD) in the dataset was pruned with Plink (Purcell et al. 2007) using the option --*indep -50 -5*. After LD pruning, the number of considered SNPs was 124,518.

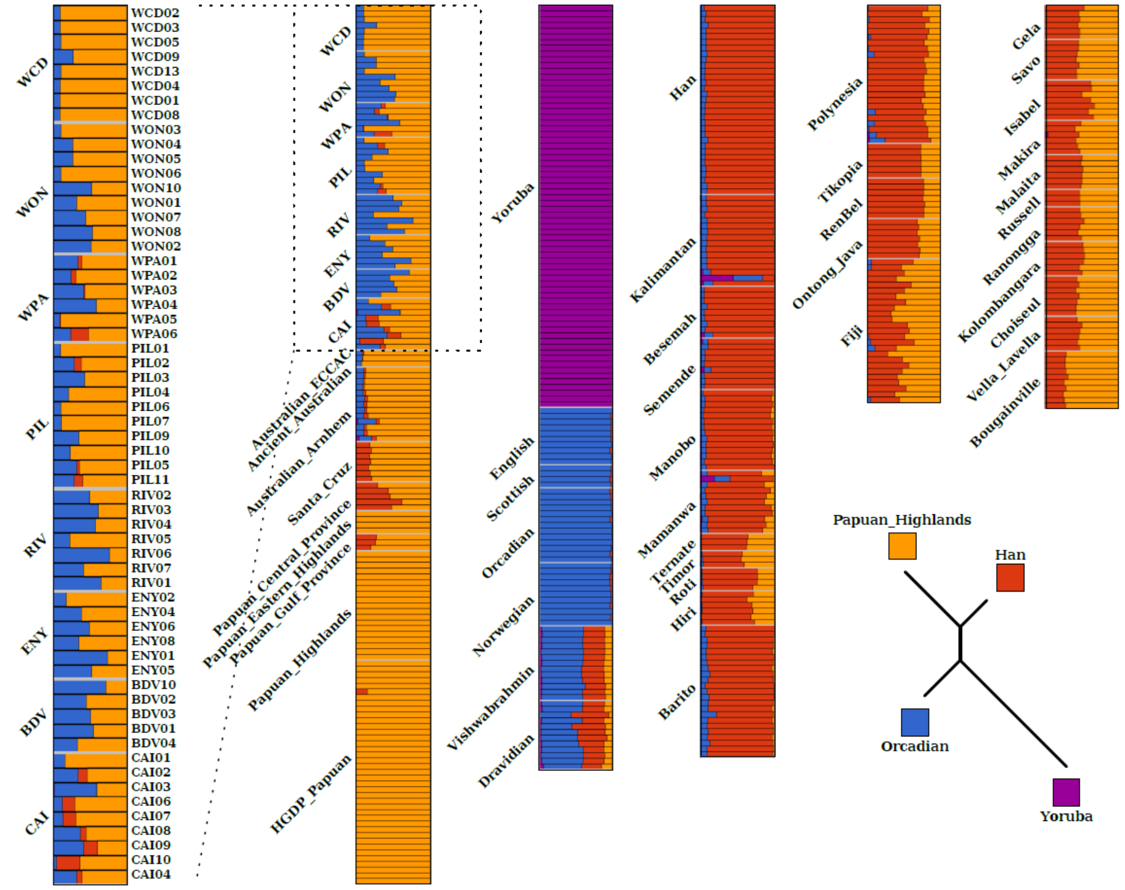
We ran our newly developed algorithm for *K* from 1 to 10. For each K value we ran the optimization ten times with different starting points and selected the result with the higher likelihood.

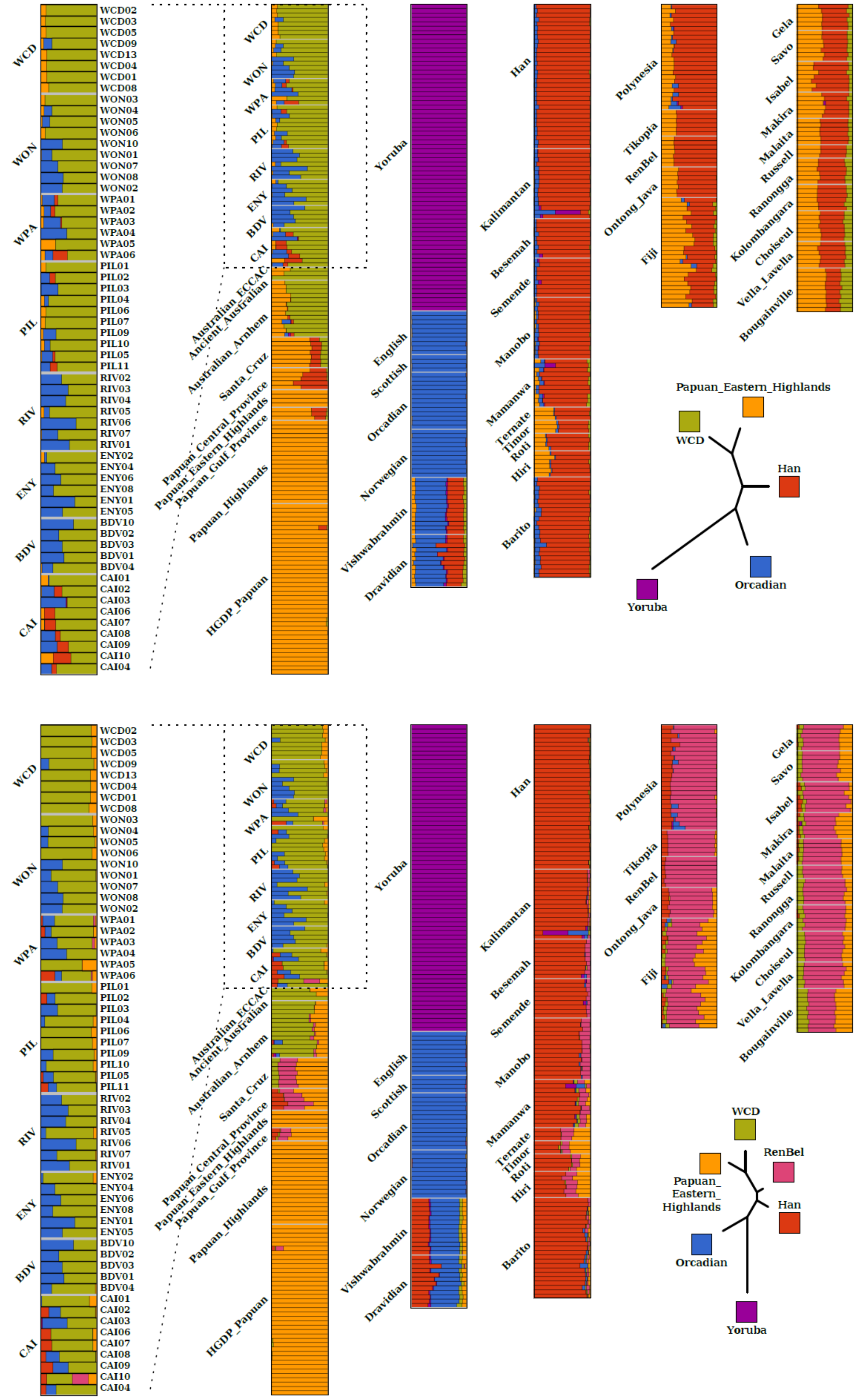
Results

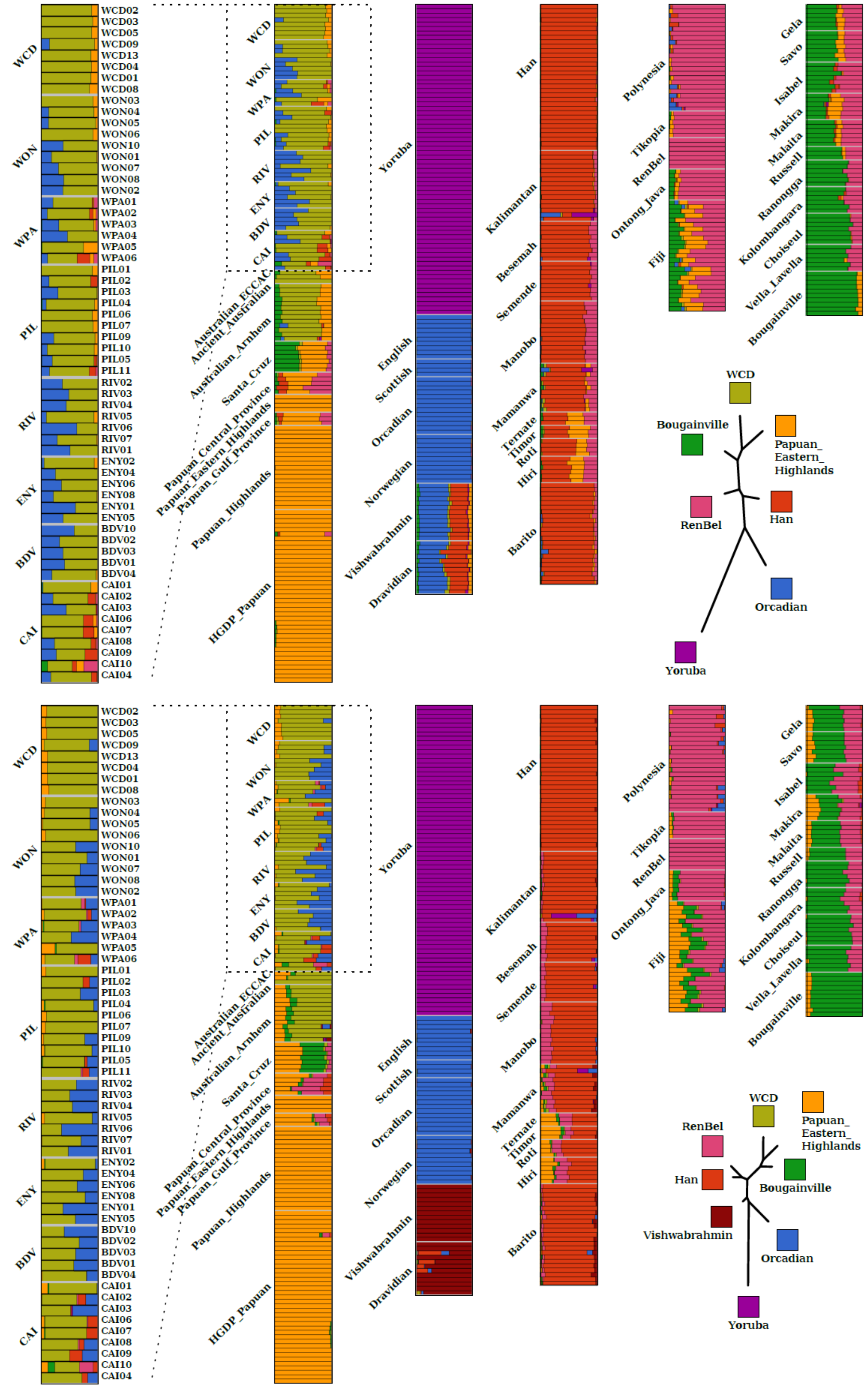
The results (the proportions of each ancestral component for each individual and the trees relating the ancestral components) are shown in Figure S05.6 and Extended Data Figure 1.

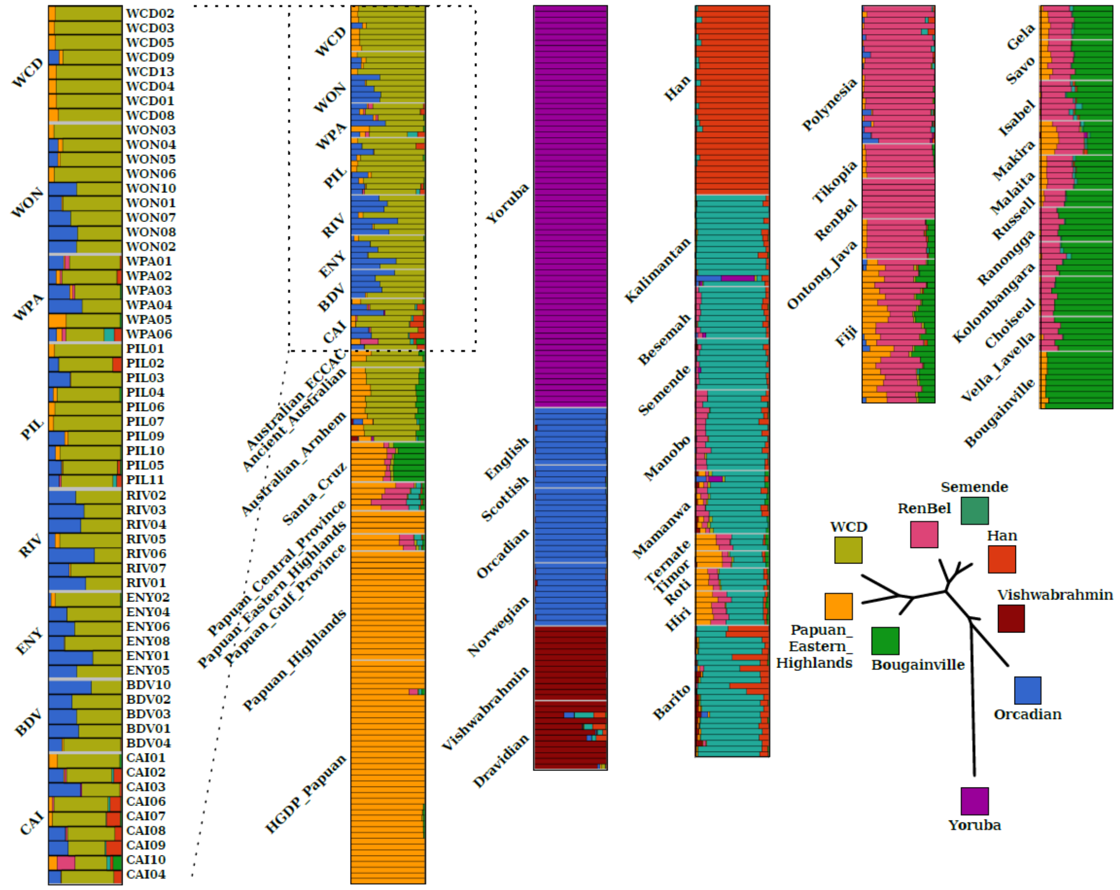



WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
WCD02
WCD03
WCD05
WCD09
WCD13
WCD04
WCD01
WCD08
WON03
WON04
WON05
WON06
WON10
WON01
WON07
WON08
WON02
WPA01
WPA02
WPA03
WPA04
WPA05
WPA06
PIL01
PIL02
PIL03
PIL04
PIL06
PIL07
PIL09
PIL10
PIL05
PIL11
RIV02
RIV03
RIV04
RIV05
RIV06
RIV07
RIV01
ENY02
ENY04
ENY06
ENY08
ENY01
ENY05
BDV10
BDV02
BDV03
BDV01
BDV04
CAI01
CAI02
CAI03
CAI06
CAI07
CAI08
CAI09
CAI10
CAI04
WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
Australian_ECCAC
Ancient_Australian
Australian_Arnhem
Santa_Cruz
Papuan_Central_Province
Papuan_Eastern_Highlands
Papuan_Gulf_Province
Papuan_Highlands
HGDP_Papuan
Yoruba
English
Scottish
Orcadian
Norwegian
Vishwabrahmin
Dravidian
Han
Kalimantan
Besemah
Semende
Manobo
Mamanwa
Ternate
Timor
Roti
Hiri
Barito
Polynesia
Tikopia
RenBel
Ontong_Java
Fiji
Gela
Savo
Isabel
Makira
Malaita
Russell
Ranongga
Kolombangara
Choiseul
Vella_Lavella
Bougainville
Papuan_Eastern_Highlands
Han
Yoruba


WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
WCD02
WCD03
WCD05
WCD09
WCD13
WCD04
WCD01
WCD08
WON03
WON04
WON05
WON06
WON10
WON01
WON07
WON08
WON02
WPA01
WPA02
WPA03
WPA04
WPA05
WPA06
PIL01
PIL02
PIL03
PIL04
PIL06
PIL07
PIL09
PIL10
PIL05
PIL11
RIV02
RIV03
RIV04
RIV05
RIV06
RIV07
RIV01
ENY02
ENY04
ENY06
ENY08
ENY01
ENY05
BDV10
BDV02
BDV03
BDV01
BDV04
CAI01
CAI02
CAI03
CAI06
CAI07
CAI08
CAI09
CAI10
CAI04
WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
Australian_ECCAC
Ancient_Australian
Australian_Arnhem
Santa_Cruz
Papuan_Central_Province
Papuan_Eastern_Highlands
Papuan_Gulf_Province
Papuan_Highlands
HGDP_Papuan
Yoruba
English
Scottish
Orcadian
Norwegian
Vishwabrahmin
Dravidian
Han
Kalimantan
Besemah
Semende
Manobo
Mamanwa
Ternate
Timor
Roti
Hiri
Barito
Polynesia
Tikopia
RenBel
Ontong_Java
Fiji
Gela
Savo
Isabel
Makira
Malaita
Russell
Ranongga
Kolombangara
Choiseul
Vella_Lavella
Bougainville
Papuan_Highlands
Han
Orcadian
Yoruba


WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
WCD02
WCD03
WCD05
WCD09
WCD13
WCD04
WCD01
WCD08
WON03
WON04
WON05
WON06
WON10
WON01
WON07
WON08
WON02
WPA01
WPA02
WPA03
WPA04
WPA05
WPA06
PIL01
PIL02
PIL03
PIL04
PIL06
PIL07
PIL09
PIL10
PIL05
PIL11
RIV02
RIV03
RIV04
RIV05
RIV06
RIV07
RIV01
ENY02
ENY04
ENY06
ENY08
ENY01
ENY05
BDV10
BDV02
BDV03
BDV01
BDV04
CAI01
CAI02
CAI03
CAI06
CAI07
CAI08
CAI09
CAI10
CAI04
Australian_ECCAC
Ancient_Australian
Australian_Arnhem
Santa_Cruz
Papuan_Central_Province
Papuan_Eastern_Highlands
Papuan_Gulf_Province
Papuan_Highlands
HGDP_Papuan
Yoruba
English
Scottish
Orcadian
Norwegian
Vishwabrahmin
Dravidian
Han
Kalimantan
Besemah
Semende
Manobo
Mamanwa
Ternate
Timor
Roti
Hiri
Barito
Polynesia
Tikopia
RenBel
Ontong_Java
Fiji
Gela
Savo
Isabel
Makira
Malaita
Russell
Ranongga
Kolombangara
Choiseul
Vella_Lavella
Bougainville
Papuan_Eastern_Highlands
WCD
Han
Orcadian
Yoruba
WCD
RenBel
Papuan_
Eastern_
Highlands
Han
Orcadian
Yoruba


WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
WCD02
WCD03
WCD05
WCD09
WCD13
WCD04
WCD01
WCD08
WON03
WON04
WON05
WON06
WON10
WON01
WON07
WON08
WON02
WPA01
WPA02
WPA03
WPA04
WPA05
WPA06
PIL01
PIL02
PIL03
PIL04
PIL06
PIL07
PIL09
PIL10
PIL05
PIL11
RIV02
RIV03
RIV04
RIV05
RIV06
RIV07
RIV01
ENY02
ENY04
ENY06
ENY08
ENY01
ENY05
BDV10
BDV02
BDV03
BDV01
BDV04
CAI01
CAI02
CAI03
CAI06
CAI07
CAI08
CAI09
CAI10
CAI04
Australian_ECCAC
Ancient_Australian
Australian_Arnhem
Santa_Cruz
Papuan_Central_Province
Papuan_Eastern_Highlands
Papuan_Gulf_Province
Papuan_Highlands
HGDP_Papuan
Yoruba
English
Scottish
Orcadian
Norwegian
Vishwabrahmin
Dravidian
Han
Kalimantan
Besemah
Semende
Manobo
Mamanwa
Ternate
Timor
Roti
Hiri
Barito
Polynesia
Tikopia
RenBel
Ontong_Java
Fiji
Gela
Savo
Isabel
Makira
Malaita
Russell
Ranongga
Kolombangara
Choiseul
Vella_Lavella
Bougainville
WCD
Bougainville
Papuan_Eastern_Highlands
RenBel
Han
Orcadian
Yoruba
WCD
RenBel
Papuan_Eastern_Highlands
Han
Bougainville
Vishwabrahmin
Orcadian
Yoruba


WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
WCD02
WCD03
WCD05
WCD09
WCD13
WCD04
WCD01
WCD08
WON03
WON04
WON05
WON06
WON10
WON01
WON07
WON08
WON02
WPA01
WPA02
WPA03
WPA04
WPA05
WPA06
PIL01
PIL02
PIL03
PIL04
PIL06
PIL07
PIL09
PIL10
PIL05
PIL11
RIV02
RIV03
RIV04
RIV05
RIV06
RIV07
RIV01
ENY02
ENY04
ENY06
ENY08
ENY01
ENY05
BDV10
BDV02
BDV03
BDV01
BDV04
CAI01
CAI02
CAI03
CAI06
CAI07
CAI08
CAI09
CAI10
CAI04
WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
Australian_ECCAC
Ancient_Australian
Australian_Arnhem
Santa_Cruz
Papuan_Central_Province
Papuan_Eastern_Highlands
Papuan_Gulf_Province
Papuan_Highlands
HGDP_Papuan
Yoruba
English
Scottish
Orcadian
Norwegian
Vishwabrahmin
Dravidian
Han
Kalimantan
Besemah
Semende
Manobo
Mamanwa
Ternate
Timor
Roti
Hiri
Barito
Polynesia
Tikopia
RenBel
Ontong_Java
Fiji
Gela
Savo
Isabel
Makira
Malaita
Russell
Ranongga
Kolombangara
Choiseul
Vella_Lavella
Bougainville
Semende
RenBel
WCD
Han
Vishwabrahmin
Papuan_
Eastern_
Highlands
Bougainville
Orcadian
Yoruba

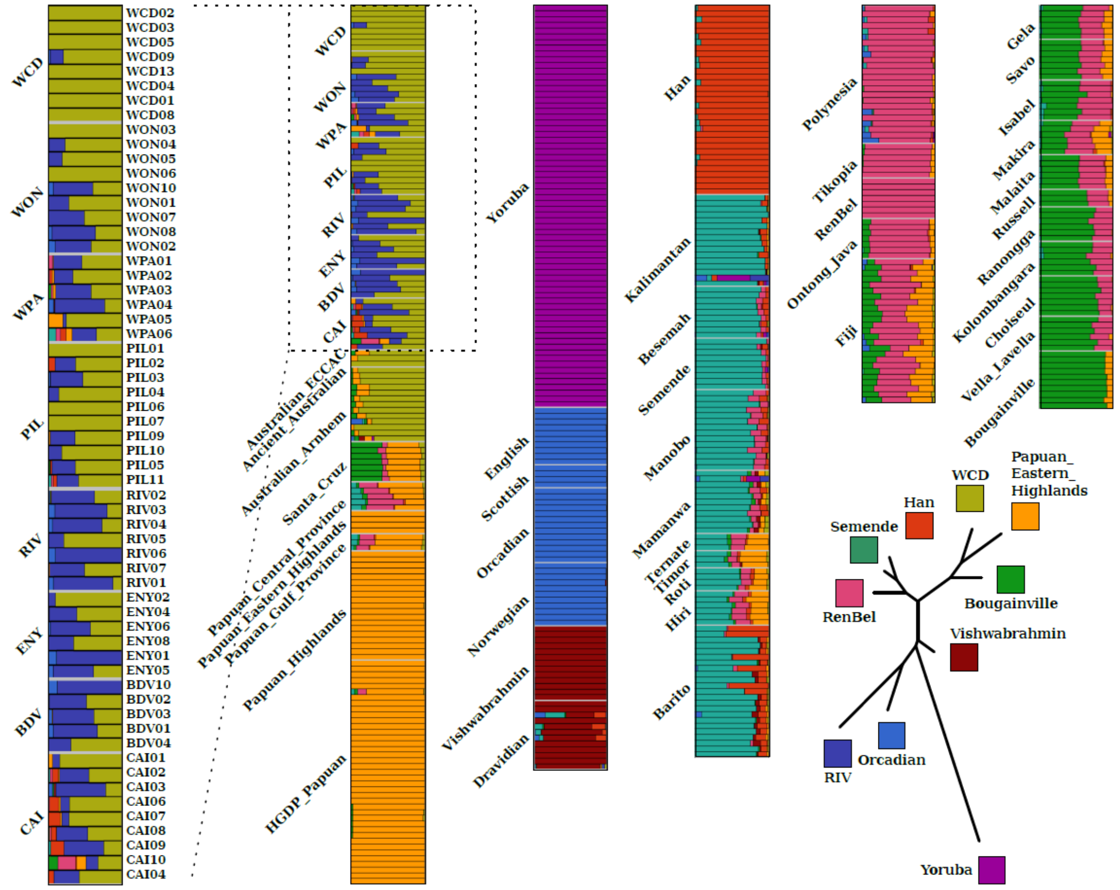

WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
WCD02
WCD03
WCD05
WCD09
WCD13
WCD04
WCD01
WCD08
WON03
WON04
WON05
WON06
WON10
WON01
WON07
WON08
WON02
WPA01
WPA02
WPA03
WPA04
WPA05
WPA06
PIL01
PIL02
PIL03
PIL04
PIL06
PIL07
PIL09
PIL10
PIL05
PIL11
RIV02
RIV03
RIV04
RIV05
RIV06
RIV07
RIV01
ENY02
ENY04
ENY06
ENY08
ENY01
ENY05
BDV10
BDV02
BDV03
BDV01
BDV04
CAI01
CAI02
CAI03
CAI06
CAI07
CAI08
CAI09
CAI10
CAI04
WCD
WON
WPA
PIL
RIV
ENY
BDV
CAI
Australian_ECCAC
Ancient_Australian
Australian_Arnhem
Santa_Cruz
Papuan_Central_Province
Papuan_Eastern_Highlands
Papuan_Gulf_Province
Papuan_Highlands
HGDP_Papuan
Yoruba
English
Scottish
Orcadian
Norwegian
Vishwabrahmin
Dravidian
Han
Kalimantan
Besemah
Semende
Manobo
Mamanwa
Ternate
Timor
Roti
Hiri
Barito
Polynesia
Tikopia
RenBel
Ontong_Java
Fiji
Gela
Savo
Isabel
Makira
Malaita
Russell
Ranongga
Kolombangara
Choiseul
Vella_Lavella
Bougainville
WCD
Papuan_
Eastern_
Highlands
Han
Semende
Bougainville
RenBel
Vishwabrahmin
RIV
Orcadian
Yoruba

Figure S05.6 Per individual admixture proportions of ancestral components as computed with our new method applied to data including Aboriginal Australians, New Guineans, Europeans, Africans, Melanesians and Polynesians. The genome of each individual is depicted as a bar and is colored according to the estimated genome-wide proportions of ancestry components. The results are shown for K=2 (top) up to K=10 (bottom). An unrooted tree showing the relationships between the identified ancestral components is also estimated by our method. Each ancestry has been labelled with the name of the population showing the highest fraction of that ancestral component.

S05 References

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

Bates D, Maechler M, Bolker B, Walker S. lme4: Linear mixed-effects models using Eigen and S4. Available from: http://CRAN.R-project.org/package=lme4

Cailliez F. 1983. The analytical solution of the additive constant problem. Psychometrika 48:305–308.

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39:783–791.

Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196:973–983.

Golden S. MapViewer.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinforma. Oxf. Engl. 23:1801–1806.

Karush W. 1939. Minima of Functions of Several Variables with Inequalities as Side Constraints. M.Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois. Dept. of Mathematics, Univ. of Chicago, Chicago, Illinois

Kuhn H, Tucker A. 1951. Nonlinear programming. In: Nonlinear programming. Proceedings of 2nd Berkeley Symposium. Berkeley: University of California Press. p. 481–492.

Loh P-R, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, Berger B. 2013. Inferring admixture histories of human populations using linkage disequilibrium. Genetics 193:1233–1254.

Moltke I, Fumagalli M, Korneliussen TS, Crawford JE, Bjerregaard P, Jørgensen ME, Grarup N, Gulløv HC, Linneberg A, Pedersen O, et al. 2015. Uncovering the genetic history of the present-day Greenlandic population. Am. J. Hum. Genet. 96:54–69.

Murty KG. 1988. Linear complementarity, linear and nonlinear programming. Berlin: Heldermann Verlag

Nelder R, Mead JR. 1965. A simplex method for function minimization. Comput. J. 7:308–313.

Nocedal J, Wright SJ. 2006. Numerical Optimization. Springer

Patterson NJ, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient Admixture in Human History. Genetics:genetics.112.145037.

Pemberton TJ, DeGiorgio M, Rosenberg NA. 2013. Population structure in a comprehensive genomic data set on human microsatellite variation. G3 Bethesda Md 3:891–907.

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

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945–959.

Pugach I, Delfin F, Gunnarsdóttir E, Kayser M, Stoneking M. 2013. Genome-wide data substantiate Holocene gene flow from India to Australia. Proc. Natl. Acad. Sci. U. S. A. 110:1803–1808.

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

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc. Natl. Acad. Sci. U. S. A. 102:15942–15947.

Rosenberg NA, Blum MGB. 2007. Sampling properties of homozygosity-based statistics for linkage disequilibrium. Math. Biosci. 208:33–47.

Sabatti C, Risch N. 2002. Homozygosity and linkage disequilibrium. Genetics 160:1707–1719.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28:289–301.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358–1370.

Wollstein A, Lao O. 2015. Detecting individual ancestry in the human genome. Investig. Genet. 6:7.

Yang W-Y, Novembre J, Eskin E, Halperin E. 2012. A model-based approach for analysis of spatial structure in genetic data. Nat. Genet. 44:725–731.

S06 Local ancestry

Georgios Athanasiadis, Thomas Mailund, Mikkel H Schierup

Method

We used RFMix (Maples et al., 2013) for local ancestry inference on Aboriginal Australian chromosomes and follow-up analysis of total genomic admixture proportions and population structure. RFMix takes into account linkage disequilibrium (LD) among studied markers and identifies ancestry tracts originating in each of the admixing populations.

We organized the available data into *target* and *reference* populations (representing possible sources of admixture). The target population (N = 58) included all unrelated Aboriginal Australian samples, except for the WCD individuals, who were used as a reference population, CAI10, who had declared to be originally from Papua New Guinea and WPA05, who was found to be an outlier (Extended Data Figure 2). The reference populations included the 1000 Genomes Project (1000GP) Europeans represented by Brits/Scots (N=27), the 1000GP East Asians represented by Southern Han Chinese (N=29), the HGDP-Papuans (N = 13, excluding 13748_4 presumably partially admixed with East Asians), and the Aboriginal Australians represented by WCD (N=8, including individual WCD09 - presumably admixed with Europeans (S05)).

Note that we used phased data for this analysis. The Aboriginal Australian and HGDP-Papuan genome sequence data were phased with IMPUTE2 (S03), while we downloaded the publically available 1000GP phased data (*https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html*).

RFMix requires a unique SNP set across all reference and target populations with no missing values. After filtering for 100% data completeness, we acquired a set of 12,136,410 autosomal SNPs.

RFMix analysis was repeated three times, assuming the admixture took place g=10, 30 or 100 generations ago to assess if results were robust to model assumptions. Having found that the results were indeed robust across the three trials, we present the results for g = 10 in what follows.

Results and discussion

Total genomic admixture proportions

Ancestry tract length was calculated by subtracting initial from final genomic position (in cM) of each tract as reported by RFMix. Tract length calculations were restricted within chromosome arms, thus omitting centromeres. We then summed over all ancestry tracts from each ancestry and for each individual, and produced total genomic admixture (ancestry) proportions (Figure S06.1). We found that the European signal was present across all populations, while only some populations (CAI, WPA and PIL) carried an obvious East Asian signal (mean values: 20.14%, 14.68% and 5.86%, respectively). These results are in agreement with the global ancestry analysis performed with sNMF (Frichot et al., 2014) (S05).

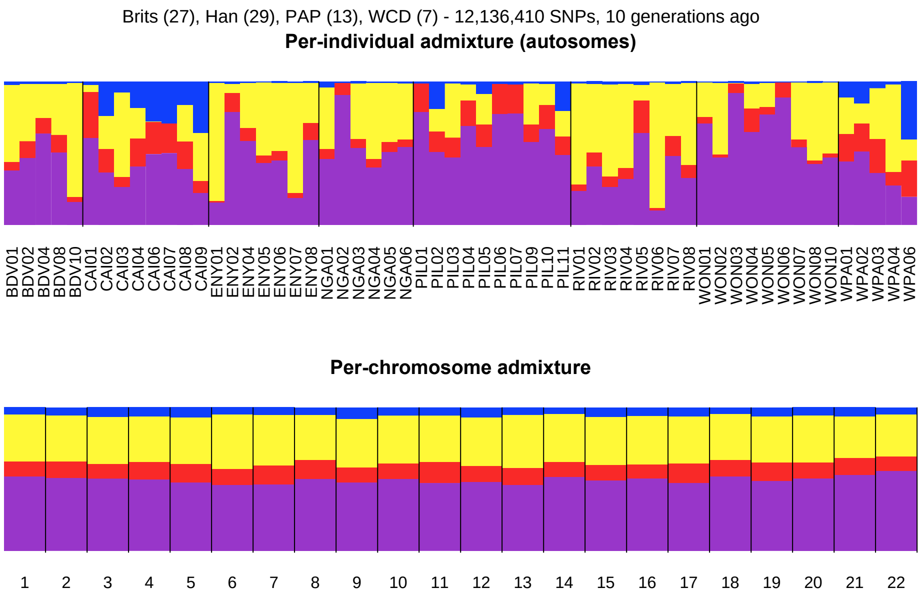


Figure S06.1 Per-individual (upper panel) and per-chromosome (lower panel) total genomic admixture proportions in the Aboriginal Australian samples. Colour code: purple = Aboriginal Australian; red = Papuan; yellow = European; blue = East Asian.

Distribution of ancestry tracts

Length distribution of European, East Asian and Papuan tracts were produced by binning the tracts into 1 Mb-long fragments. The resulting frequencies are shown in Figure S06.2. We found that the European tracts (mean length = 34.4 cM; median length = 25.8 cM; variance = 1029.4 cM^2) were predominant over the East Asian tracts (mean length = 9.4 cM; median length = 3.9 cM; variance = 201.9 cM^2) and the Papuan ones (mean length = 8.2; median length = 5.9 cM; variance = 55.0 cM^2) both in frequency and length.

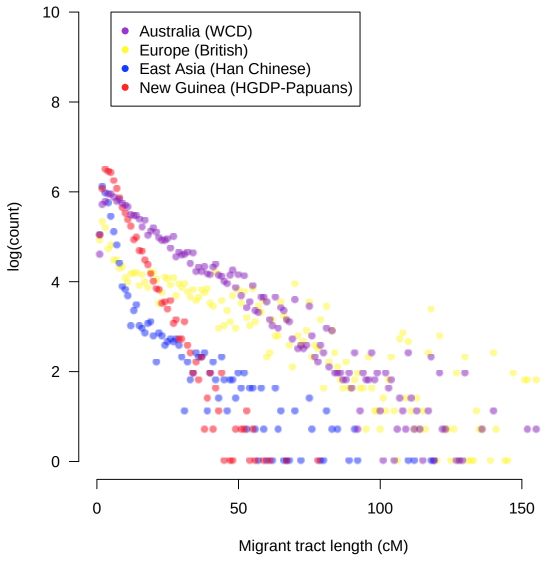


Figure S06.2 Ancestry tract length distribution for the 58 unrelated non-WCD Australian samples.

To further investigate the genetic influence of Papuans on Aboriginal Australia, we ran follow-up analyses focusing on the Aboriginal Australian and Papuan fraction of ancestry tracts.

We first wanted to make sure that Papuan tract detection was trustworthy rather than confounded by the relatively close relationship between Papuans and Aboriginal Australians. Figure S06.3 shows the relationship between Papuan genomic admixture proportion, as inferred by RFMix, and (i) median length; (ii) maximum length; and (iii) number of Papuan tracts. Note that, the Papuan admixture proportions are relative to the total Sahul component, i.e they were divided by the sum of both Papuan and Aboriginal Australian proportions. In all cases, correlation was strong and significant (data not shown) and samples from different locations tended to cluster in a geographically meaningful manner.

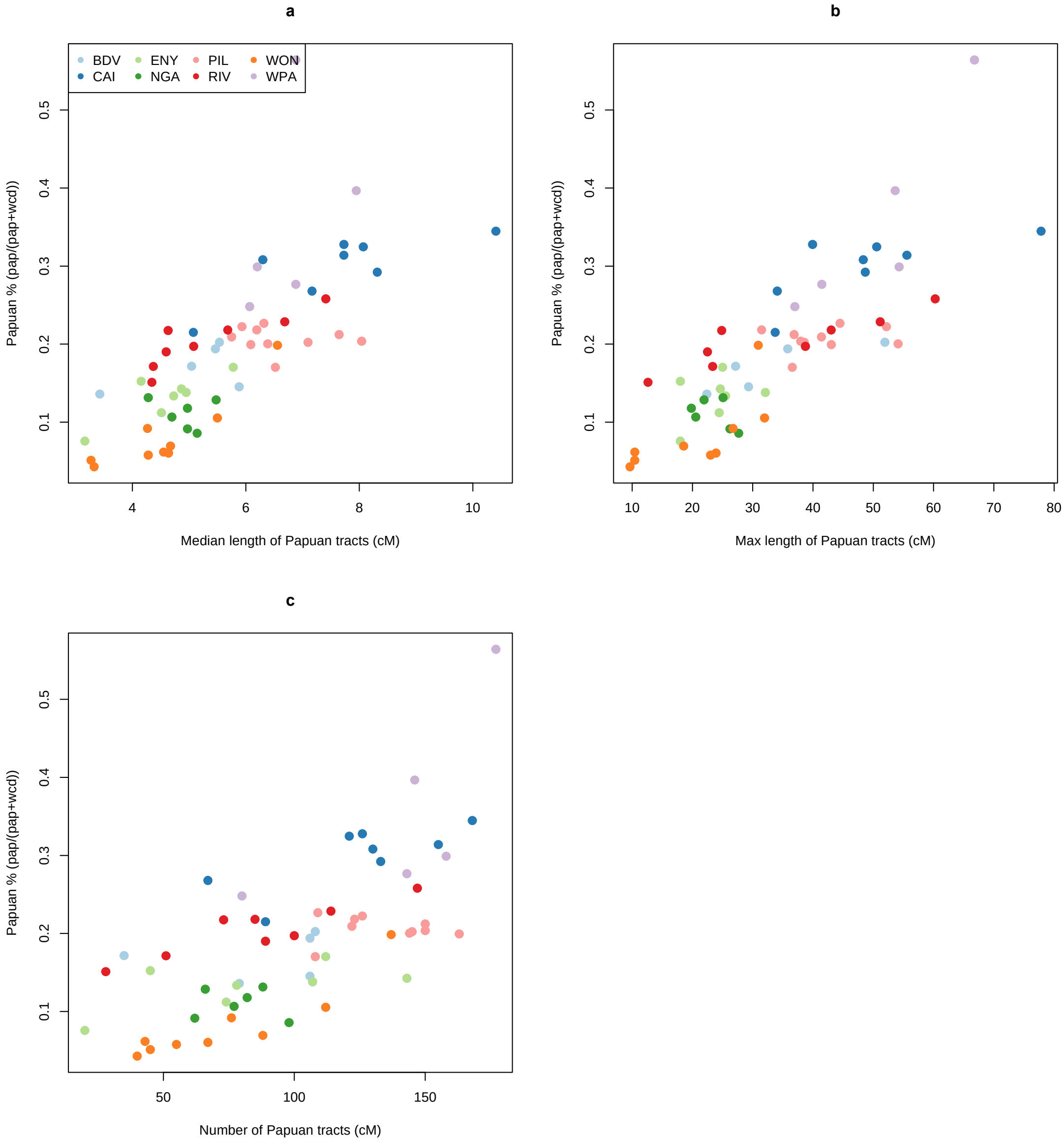


Figure S06.3 Per-individual scatter plots of (a) median length; (b) maximum length; and (c) number of Papuan tracts vs. Papuan ancestry within Sahul context (i.e Papuan ancestry over Papuan and Aboriginal Australian ancestry proportions).

We then checked whether Papuan tract statistics for each Aboriginal Australian sample presented significant correlation with distance from a putative distal point, e.g., West Central Desert. Figure S06.4 shows the relationship between geographic distance from WCD and (i) median Papuan tract length; (ii) its variance; and (iii) its coefficient of variation. In all three cases, we observe significant correlation with geography (correlation coefficient *r*: 0.536, 0.637 and 0.637, respectively; p-val << 0.0001), with ancestry tracts looking “younger” (higher median length and variance/coefficient of variation thereof) the more we approach Papua New Guinea and “older” (lower median length and variance/coefficient of variation thereof) the more we approach WCD. We interpret this trend as suggestive of continuous gene flow from Papua New Guinea into Aboriginal Australians, whereby we expect the Papuan tracts to be both fewer and shorter the further we move from the source.

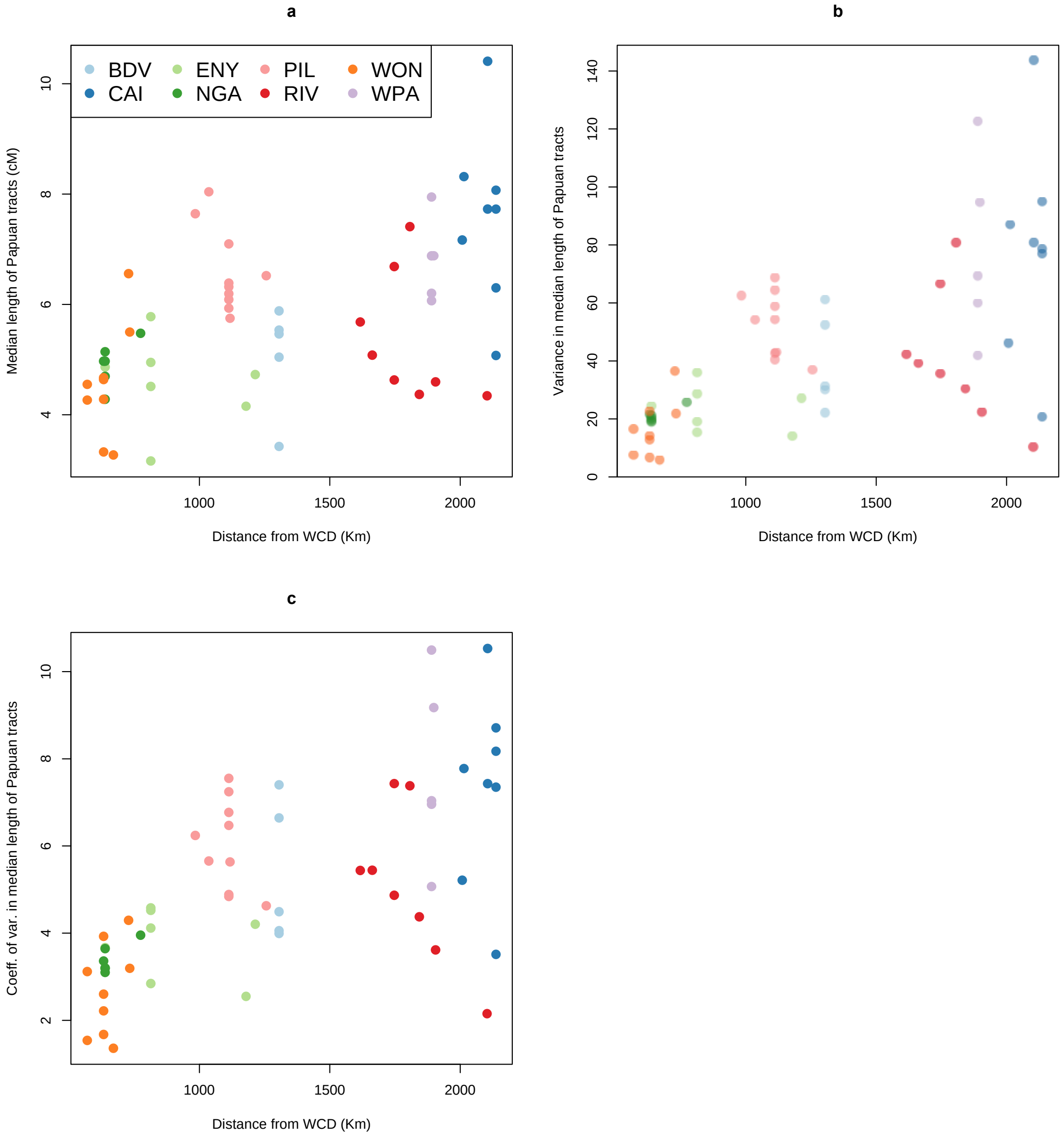


Figure S06.4 Per-individual scatter plots of (a) median Papuan tract length; (b) its variance; and (c) its coefficient of variation vs. geographic distance from the WCD population**.**

Masking the non-Aboriginal tracts

We revisited standard population genetic analysis (such as multidimensional scaling and ADMIXTURE) by focusing only on the Aboriginal fraction of the Australian genomes. In particular, we leveraged the information from the local ancestry inference by keeping only loci with both alleles likely to be of Aboriginal Australian ancestry. Because local ancestry profiles vary across individuals, fully Aboriginal loci showed high variance in number and location. We used PLINK (Chang et al., 2015; Purcell et al., 2007) to set non-Aboriginal loci to missing. We then created an "LD-thinned" (window size = 50 SNPs; step size = 5 SNPs; r^2 threshold = 0.5) dataset to reduce the LD among loci considered and further filtered the resulting masked dataset to exclude individuals with too many missing genotypes. For all analyses with masked data and in order to keep the larger number of individuals possible, we set maximum per-individual genotype missingness to 90%, thus losing 11 out of 66 samples. We also applied per-locus filters to the remaining data (MAF > 0.05 and per-locus genotype missingness < 50%). The final dataset included 32,105 SNPs (genotype completeness = 59.6%). See Figure S06.5 for the result of the multidimensional scaling (MDS) for all Aboriginal Australian samples (target and WCD, i.e., 66 samples) produced by PLINK.

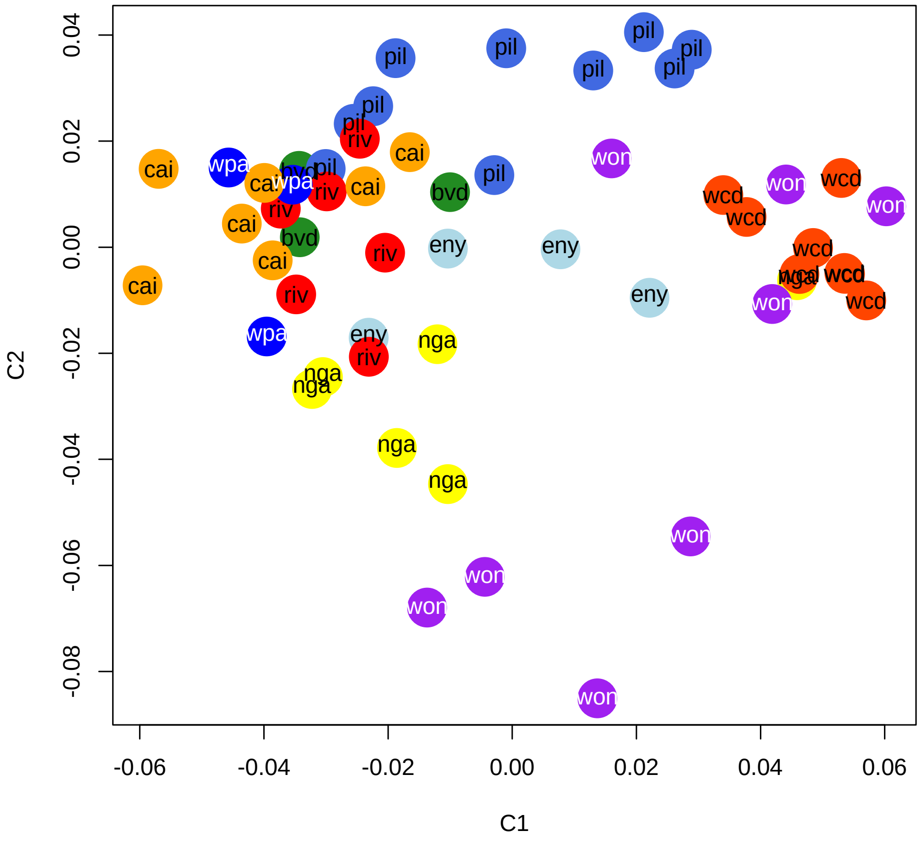


Figure S06.5 MDS based on the ancestry masked data (32,105 SNPs) for Aboriginal Australians including WCD (see also S13).

Finally, ancestral component analysis was performed on the masked data using two different methods: sNMF and ADMIXTURE (Alexander et al., 2009). sNMF was used on the masked Aboriginal Australian data alone. In particular, we ran sNMF for K = {1, 2, …, 10}, calculating the cross validation error at each run. As seen in Table S06.1, the smallest error corresponds to K = 1, suggesting that the Aboriginal fraction of the Australian samples is best described as one homogeneous cluster.

Table S06.1 Cross validation error for runs of sNMF assuming K = {1, 2, …, 10}

K ancestral components	Cross validation error
1	0.786
2	0.805
3	0.850
4	0.867
5	0.916
6	0.941
7	0.984
8	1.024
9	1.072
10	1.142

We used the same masked dataset together with data from four reference populations (the 1000GP Brits, the 1000GP Han Chinese, the 1000GP Telugu and the HGDP Papuans) to calculate admixture proportions assuming K = 5 for direct comparisons with the equivalent unmasked analysis (S05). This analysis allows us in principle to re-address the question of Indian admixture in the Aboriginal Australian component after removing the effect of European, Asian and Papuan admixture. For this purpose, we ran ADMIXTURE five times and the results were combined using the CLUMPP permutation algorithm (Jakobsson and Rosenberg, 2007) to produce Figure S06.6.

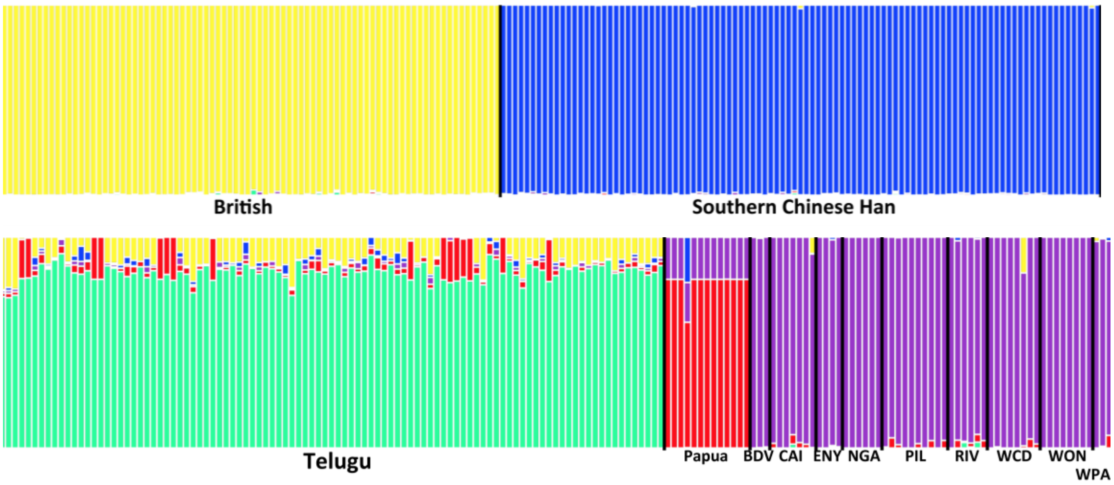


Figure S06.6 Admixture proportions of reference (Europe, India, East Asia, Papua New Guinea) and target populations (Aboriginal Australians).

We see that WON, NGA, and ENY, as well as BDV, had virtually 100% membership to the Aboriginal component. Most individuals from the remaining populations, including WCD, still showed low proportions of Papuan admixture (median proportion = 2.17%), but given the masking of the data for Papuan admixture, we interpret this as noise. In some RIV and CAI individuals, Indian admixture was also present in very low proportions (less than 3%), within the range of the artifactual Papuan admixture. Therefore, we interpret this signal as noise. The result for WCD09 is as expected given that this individual appears to have European admixture (S05) (note that the WCD samples were not masked but rather added in the analysis *a posteriori*).

S06 References

Alexander, D.H., Novembre, J., Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664. doi:10.1101/gr.094052.109

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4. doi:10.1186/s13742-015-0047-8

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., François, O., 2014. Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics 196, 973–983. doi:10.1534/genetics.113.160572

Jakobsson, M., Rosenberg, N.A., 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinforma. Oxf. Engl. 23, 1801–1806. doi:10.1093/bioinformatics/btm233

Maples, B.K., Gravel, S., Kenny, E.E., Bustamante, C.D., 2013. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. Am. J. Hum. Genet. 93, 278–288. doi:10.1016/j.ajhg.2013.06.020

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. doi:10.1086/519795

S07 Demographic inferences

Vitor C Sousa, Isabel Alves, Isabelle Dupanloup, Laurent Excoffier

Statistical framework

Likelihood inference based on the Site Frequency Spectrum

We inferred demographic scenarios using the site frequency spectrum (SFS) (Nielsen 2000; Adams, Hudson 2004) by approximating the likelihood of a given model with coalescent simulations (Nielsen 2000). All computations were done with an extension of the fastsimcoal2 simulation software (Excoffier et al. 2013). Coalescent simulations are performed under specific parameters values θ of a given model to estimate the expected entries of the SFS $\hat{p}_i$, and the likelihood is then obtained as

$$L_{full} = \Pr(X \mid \theta) \propto P_0^{L-S}(1-P_0)^S \prod_{i=1}^{n-1} \hat{p}_i^{m_i}$$

where $X = \{m_1, \ldots, m_{n-1}\}$ is the observed (multidimensional) SFS, *n* is the total number of entries in the SFS given by the product of $n = \prod_{i=1}^{d} (n_i + 1)$, n_i are the number of gene copies in the *i*-th deme, *S* is the number of polymorphic sites, *L* is the length of the observed sequence data, and P_0 is the probability of no mutation on the tree, obtained as $P_0 = e^{-\mu T}$ assuming a Poisson distribution of mutations occurring at rate μ, where *T* is the expected tree length. Note that if the data contains linked SNPs this is actually a composite likelihood estimator. As described in (Excoffier et al. 2013), the likelihood is maximized using a conditional maximization algorithm (ECM, Meng, Rubin 1993), which is an extension of the EM algorithm where each parameter of the model is maximized in turn, while keeping the other parameters fixed at their last estimated value. The maximization of each parameter is done using Brent’s algorithm (Brent 1973). We start with initial random parameter values, and perform a series of ECM optimization cycles until estimated values stabilize or until we have reached a predefined number of ECM cycles (50, unless specified otherwise).

We used a strategy where we begin by optimizing the full likelihood L_{full} for a given number of cycles (i.e. 10) and then optimize $L_{SFS} \propto \prod_{i=1}^{n-1} \hat{p}_i^{m_i}$ for the remaining 40 cycles. This strategy aims at maximizing the fit between the expected and the observed SFS. At the end of the run, a rescaling factor is computed as $RF = S_{obs} / S_{\exp}$, where S_{obs} and $S_{\exp}$ are the observed and expected numbers of polymorphic sites, respectively. $S_{\exp}$ is obtained as $S_{exp} = \mu\hat{T}$, with $\hat{T}$ being the expected tree length for the maximum-likelihood parameters. The final maximum-likelihood parameters are then rescaled by *RF* in order to produce a number of polymorphic sites equal to those observed: the effective population sizes *N*’s and divergence times *T*’s are multiplied by *RF*, whereas migration rates *m*’s are divided by *RF*, such that *Nm* parameters are left unchanged. Admixture rates are also left unchanged.

Demography of the Out of Africa settlement process

Data preparation and processing

We selected, for our analyses, seven Aboriginal Australian samples, from the Western Central Desert region (WCD01, WCD02, WCD03, WCD04, WCD05, WCD12, and WCD13), that seem to be the least admixed with European and East Asians (S05). In our demographic models, we also included two archaic humans (Denisova, Altai Neanderthal) and six modern humans, two from Africa, two from Europe and two from East Asia (Yoruba, Sardinians, Han Chinese, respectively) previously sequenced at high coverage (Meyer et al. 2012; Prufer et al. 2014).

We then considered autosomal SNPs found outside genic (as defined by Ensembl version 71, April 2013, Cunningham et al. 2015), and outside CpG islands (as defined on the UCSC platform, Rosenbloom et al. 2015). We kept only the SNPs found in regions that passed a set of minimal filters used for the analysis of the high coverage Denisovan and Altai Neanderthal genomes (map35_50%, Meyer et al. 2012; Prufer et al. 2014).

The ancestral state of the SNPs was inferred using the ancestral hg19 genome, which is used by the 1000G consortium (Abecasis et al. 2012), and which was inferred from the alignments of six primate genomes, and released in the Ensembl Compara 59 database (Flicek et al. 2011).

We generated a dataset of autosomal SNPs and generated the multidimensional SFS for these different datasets with the Arlequin software ver 3.5.2.2 (Excoffier and Lischer 2010). This dataset was generated by considering blocks of 1Mb concatenated regions on the autosomes that have been sequenced in all individuals (*i.e.*, no missing data), and that fall outside genic regions and CpG islands, but within *map35_50%* regions. We identified 985 such blocks on the autosomes. For the non-parametric block-bootstrap analysis, we resampled with replacement these blocks to generate 100 sets of 985 blocks. These sets have the same cumulative length (985 Mb) and carry very similar numbers of autosomal SNPs (between 4,288,078 and 4,353,529 SNPs).

Out of Africa models

We used a model-based approach to test whether the SFS data support: (i) a one wave Out of Africa scenario (*1OoA*), where all non-Africans, including Aboriginal Australians, descend from a single ancestral population that left Africa, or (ii) a two waves Out of Africa scenario (*2OoA*) with a first exit corresponding to an early dispersal leading to the colonization of Australia, and a second later exit leading to the colonization of Eurasia. Rather than comparing the likelihood of alternative models, we considered models that could encompass either *2OoA* or *1OoA* scenarios, depending on the combination of parameter values. We chose this approach because the likelihood function is a composite likelihood (due to presence of linked sites in the data) and so we cannot use classical model choice procedures such as likelihood ratios tests or Akaike Information Criterion (AIC; Varin 2008; Varin, Reid, Firth 2011) without extensive simulations. However, we can estimate the parameters that maximize the likelihood and, based on those, distinguish among alternative scenarios.

Given the evidence for admixture of Neanderthal and Denisovan with non-African modern human populations (Meyer et al. 2012; Prufer et al. 2014), we allowed for archaic admixture in our models, explicitly accounting for the shared ancestry of Neanderthals and Denisovans. We thus considered models including the following events (see Figure S07.1.A for a graphical representation): (i) two potential exits out of Africa represented by population splits from an unsampled “ghost” population related to Yoruba, which corresponds to the source population for the exits out of Africa (light blue bar); (ii) three potential bottlenecks, one in the ancestors of all non-Africans (black bar) and two others immediately after the split from the “ghost” population (light green bars); (iii) four potential pulses of Neanderthal admixture (orange arrows) occurring in four different periods: in the ancestors of all non-Africans, in the ancestors of Eurasians, in the ancestors of Aboriginal Australians, and in the ancestors of East Asians (Wall et al. 2013; Prufer et al. 2014; Vernot, Akey 2014). Furthermore, we assumed that admixture with Denisovans occurred only in the Aboriginal Australian branch (Meyer et al. 2012).

Importantly, under our parameterization scheme the *1OoA* and the *2OoA* can be seen as particular combinations of parameters. Indeed, we can recover the *1OoA* when (i) there is an ancestral bottleneck in all non-Africans (associated with the out of Africa founder event), (ii) there is a Neanderthal admixture pulse shared by the ancestors of all non-Africans, and (iii) the Aboriginal Australians and Eurasians splits from the “ghost” occur at the same time. In contrast, under a *2OoA* scenario we would expect (i) no ancestral bottleneck in the ancestors of all non-Africans and stronger continent-specific bottlenecks, (ii) no shared Neanderthal admixture in the ancestors of all non-Africans, and (iii) distinct split times from the “ghost” for the Aboriginal Australians and Eurasians, corresponding to the two independent expansions out of Africa. Note that the inferred Neanderthal admixture proportions can help us distinguish a *1OoA* from a *2OoA* scenario assuming that Neanderthal admixture happened shortly after the exit(s) from Africa (Sankararaman et al. 2012; Prufer et al. 2014; Vernot, Akey 2014). Furthermore, we assumed that the out of Africa events were associated with bottlenecks.

To investigate the relationship between Aboriginal Australians and East Asians, we considered two different population-tree topologies for all models: (i) East Asians and Europeans have a more recent common ancestor than Aboriginal Australians and East Asians (i.e. two expansion waves into Asia, Figure S07.1A/B); or (ii) Aboriginal Australians and East Asians descend from a common ancestor (i.e. a single expansion wave into Asia, Figure S07.1C).

We included the unsampled population (“Ghost”) in our model because it is unlikely that the Yoruba are the source of the exit out of Africa, and in order to take into account ancestral population structure within Africa, which, if ignored, could lead to spurious estimates of archaic admixture (Green et al. 2010; Eriksson, Manica 2012; Racimo et al. 2015). This “ghost” population can thus be seen as an East African population from where the expansions out of Africa took place. Note however that under the *1OoA*scenario, after the bottleneck, this “ghost” represents the ancestral population of all non-Africans that admixed with Neanderthal.

We modelled symmetric migrations between modern human populations under a stepping-stone model, by allowing gene flow between population from neighbouring geographical regions, i.e. between West Africa and the unsampled ghost population (East Africa/Middle East), between the unsampled ghost population and Europe, between Europe and East Asia, and between East Asia and Australia (grey arrows, Figure S07.1A). We also considered an extra migration parameter between the ancestors of Eurasians and Aboriginal Australians. Migration was assumed to only occur after the split of Eurasians from the unsampled ghost population. Moreover, we considered that the archaic populations that admixed with modern humans were two ghost populations related to the Denisovan and Neanderthal Altai samples (Denisovan-related, D.R., and Neanderthal-related, N.R.Figure S07.1B/C), respectively. This way of modelling admixture accounts for a likely genetic structure of Neanderthal and Denisovan populations, and reflects the fact that the populations that contributed to modern humans did not necessarily live in the Altai region and therefore that admixture events probably occurred at different places. All divergence times were estimated assuming a constant mutation rate of 1.25e-8/gen/site (Scally, Durbin 2012) and a generation time of 29 years (Fenner 2005).

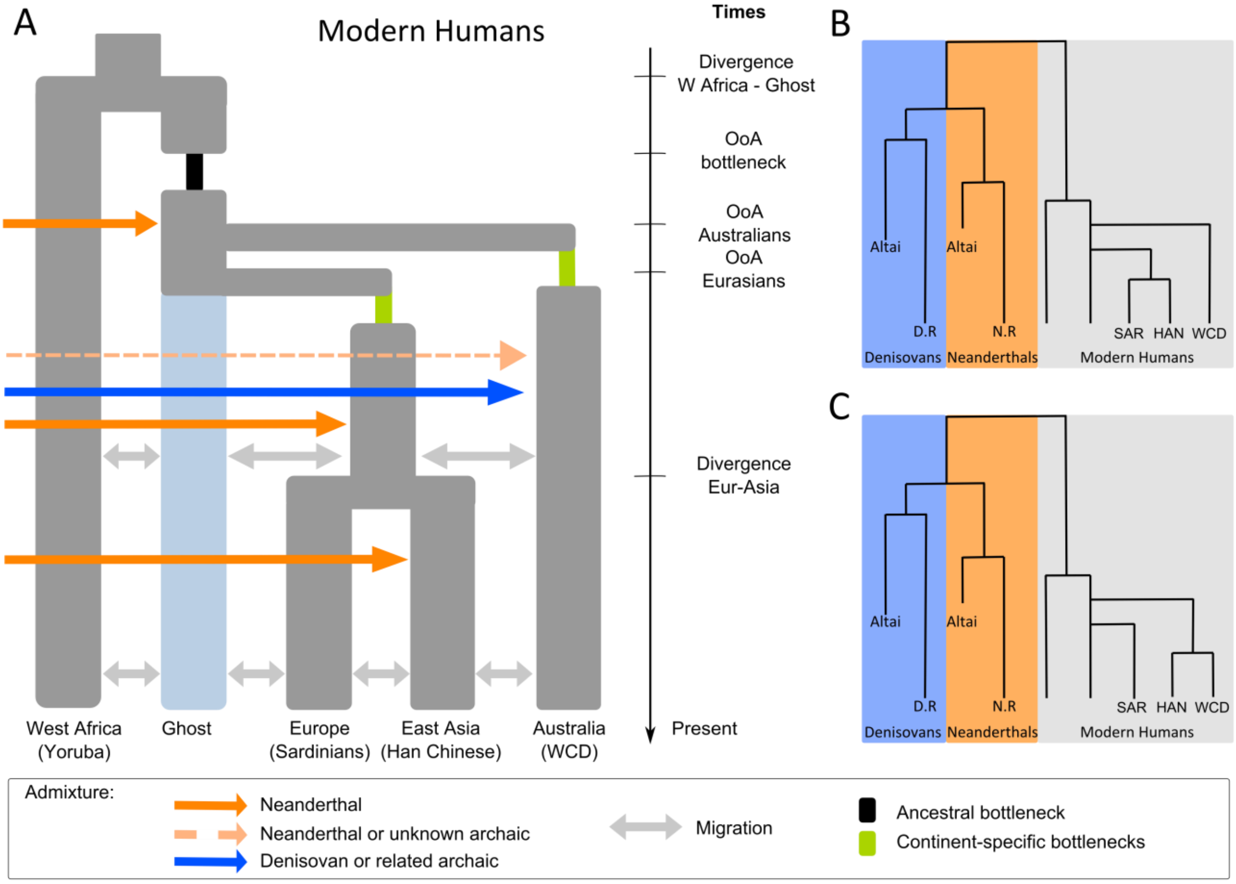


Figure S07.1 Schematic representation of the models tested in the present study. A) Representation of the model without the archaic samples but including the archaic admixture events (see panel B for the description of the full topology). We considered that modern humans admixed with unsampled populations related to Neanderthals (N.R.– Neanderthal-related) and Denisovan (D.R. – Denisovan-related), respectively. Admixture events between modern humans and N.R. and D.R. populations are represented by the orange and blue arrows, respectively. B) Representation of the population-tree topology including both archaic populations. Under this topology East Asians share a more recent common ancestor with Europeans than with Aboriginal Australians ((SAR,HAN);WCD). We explicitly modelled the relationship of Neanderthals and Denisovans. C) Same as B but with a more recent common ancestor shared between Aboriginal Australians and East Asians (SAR,(HAN,WCD)).

Maximum likelihood parameters

We estimated the set of parameters that maximize the likelihood for each model by specifying the following search ranges: (i) for all the population effective sizes we considered ranges between 100 and 100,000, but with an open upper bound that is extended if parameters get close to the boundary during the ECM optimization; (ii) for the bottlenecks we fixed a duration of 100 generations with effective sizes varying between 10 and 5,000 (sizes larger than ~2,500 correspond to mild or no bottleneck effect as the probability of coalescent events during that period becomes negligible); (iii) for the timing of events we set the divergence time between West Africans (Yoruba) and the "ghost" to vary between 1,000 and 5,000 generations, and all younger events were conditioned to be between 0 and that time; (iv) for the migration rates we assumed a range of 10^{-6} to 10^{-3} based on previous estimates (Gutenkunst et al. 2009; Rasmussen et al. 2011). For the parameters related to the archaic demography the parameter search range was based on the values reported in (Prufer et al. 2014), except that Neanderthal admixture proportions were allowed to vary between 0.0001% and 10%, and effective sizes of Denisovan and Neanderthals could vary between 1,000 and 10,000 (again with open upper bounds). The sampling times for the Altai Denisovan and Altai Neanderthal, as well as the split times of Denisovan and Neanderthal-related populations (D.R. and N.R. in Figure S07.1) were fixed to the values reported in Prufer et al. (2014). Namely, we set (i) the sampling time of Altai Denisovan and Altai Neanderthal to 2,330 and 2,957 generations ago, respectively; (ii) the divergence times between D.R. and Altai Denisovan, and between N.R. and Altai Neanderthal to 13,580 and 3,820 generations ago, respectively; (iii) the divergence time between the Denisovans and Neanderthals to 17,080 generations ago. We also took into account the inbreeding coefficient of 0.125 reported for the Altai Neanderthal individual (Prufer et al. 2014). The expected SFS used for the computation of the likelihood of a given set of parameter was obtained by performing 500,000 coalescent simulations. For the original dataset of 985 blocks of 1Mb, we performed 100 optimization runs starting from different initial conditions and selected the run leading to the highest likelihood to get parameter estimates.

Non parametric bootstrap analysis

We estimated confidence intervals for the model with maximum likelihood (i.e., the model depicted in Figure S07.1A and B with the topology where East Asians and Europeans are sister groups) by estimating parameters on 100 bootstrap datasets. All the settings for parameter estimation were the same as those used for the analyses of the original dataset, except that we performed only 10 optimization runs per bootstrap data set, due to computational constraints. To assess the effect of the lower number of runs for each bootstrap we performed 100 optimization runs for a subset of 10 bootstrap datasets. We did not find consistent differences, other than a slight increase in the variance, making this a conservative approach (data not shown). The 95% confidence intervals for each parameter were computed based on the percentile method (interval [Q0.025,Q0.975], where Qa is the a percentile of the bootstrap distribution; Davison, Hinkley 1997), as implemented in the R *boot* package.

Results

Models without archaic admixture

When considering models only with modern human populations the parameter estimates point to a *2OoA* scenario. Indeed, we find: (i) a difference of ~14 ky between the two exits out of Africa (Δt), and (ii) a strong bottleneck in the branch leading to Aboriginal Australians (N_{BOT} ~100, Figure S07.2). Overall, models where East Asians merge with Europeans more recently than with Aboriginal Australians have higher likelihoods (Table S07.1). Under this *2OoA* scenario, Aboriginal Australian ancestors would have exited Africa ~2,800 generations ago (~81 kya), followed by a second exit of Eurasian ancestors ~2,350 gen ago (~67 kya). We find evidence for high migration rates from the ancestors of Eurasians and Europeans into the "ghost" African population (2*Nm*~3), as well as between West Africa and the "ghost" (2*Nm*>4), consistent with a period of high gene flow after the exit out of Africa. Moreover, consistent with the estimates of Rasmussen et al. (2011), we find relatively large migration rates between the ancestors of Eurasians and Aboriginal Australians (2*Nm*>2).

Table S07.1 Log-likelihood values obtained for the models without archaic admixture comprising data from four modern human populations (Yoruba (YRI), Sardinians (SAR), Han Chinese (HAN) and WCD Aboriginal Australians), according to different topologies.

Model	Topology	Number of Parameters	Estimated Log10(Lhood)
Two waves into Asia	(SAR, HAN), WCD	20	-7,749,618.6
One wave into Asia	SAR, (HAN, WCD)	20	-7,752,331.6

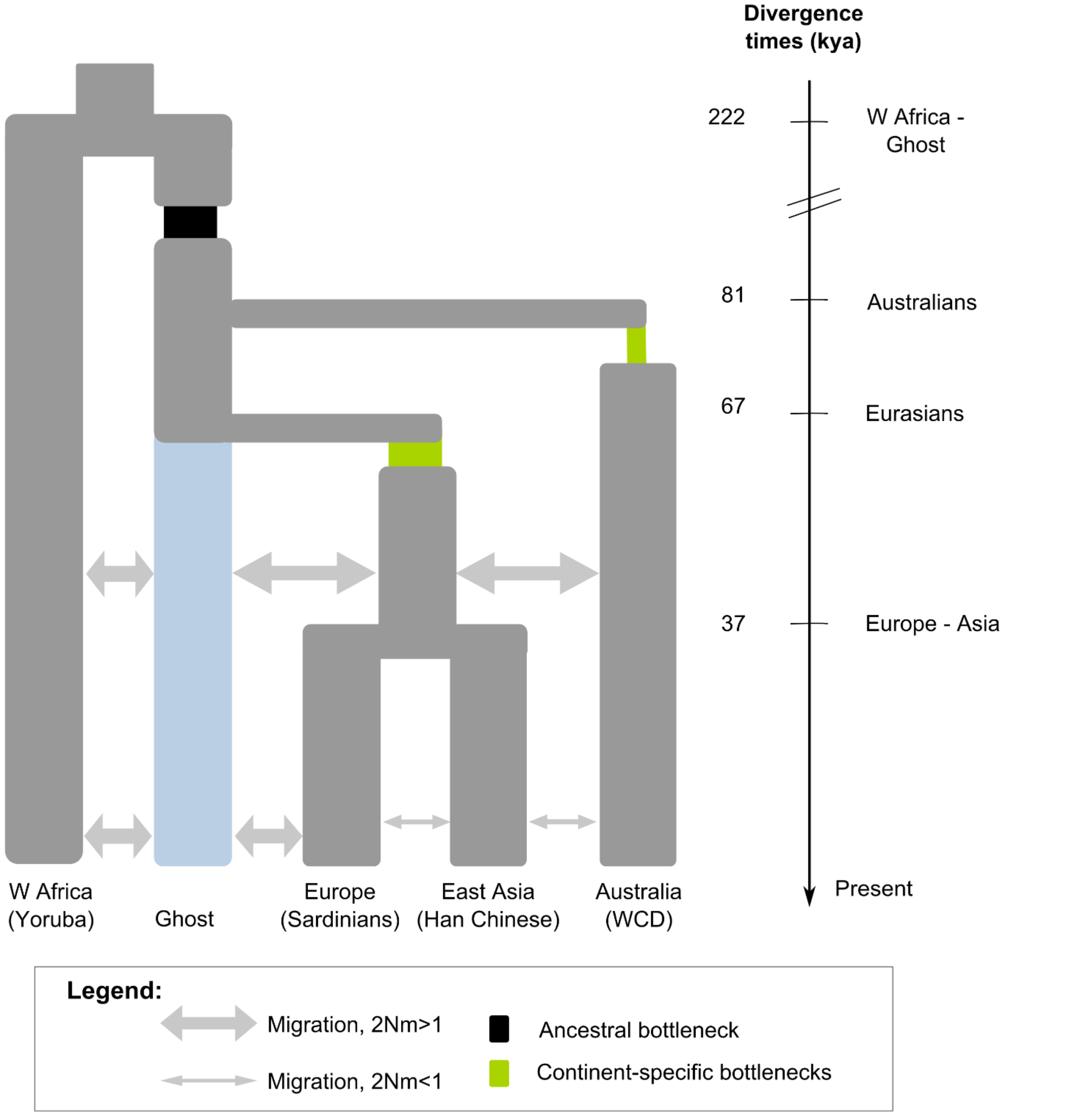


Figure S07.2 Schematic representation of the parameter estimates for a model including modern human samples only, and thus not accounting for any archaic admixture. The width of the bottlenecks is proportional to its intensity.

Models including admixture between archaics and modern humans

We find that models including archaic admixture have a much higher likelihood than models without archaic admixture, showing that including such events lead to a better fit to the data (Table S07.2). Interestingly, we no longer find support for a *2OoA* scenario when including Denisovan and Neanderthal-related admixture in the models, as the parameter estimates suggest that the two splits from the "ghost" occurred less than 1,000 years apart (Figure S07.3), with broadly overlapping 95% CI intervals (Figure S07.4A), and with most bootstraps supporting a difference smaller than ~320 generations (~9,300 y, Figure S07.4B). Moreover, the maximum likelihoods are similar under the full model with two separate splits occurring almost at the same time from the "ghost" and under a model where the two splits occur exactly at the same time (Figure S07.4C). Together with the similar divergence times, the *1OoA* scenario is further supported by a moderately strong bottleneck in the ancestral population of all non-Africans (N_e ~800, 95% CI 192 - 2221), and a major 2.4% Neanderthal contribution (95% CI 1.1%-3.5%) into all non-Africans. As seen for models including only modern humans, a topology where Europeans (Sardinians) are more closely related to East Asians (Han Chinese) has a higher likelihood than models where East Asians merge with Aboriginal Australians first (Table S07.2). These results are thus compatible with a two waves expansion into Asia from an ancestral population of all non-Africans (probably in the Near East), after a major pulse of admixture with Neanderthals. The confidence intervals for the parameters of the model are shown in Tables S07.3 and S07.5, and point estimates for the divergence times according to different mutation rates and generation times are shown in Table S07.4.

Note that these results do not depend on the inclusion of a "ghost" population. Indeed, we obtain similarly close differences in split times for a model without the ghost (approximately ~3,700y), even though it exhibited a lower likelihood than models with the "ghost" population (Table S07.2). Interestingly, the inferred divergence between Yoruba and the "ghost" population is relatively old (83-170 kya) and similar to estimates of population divergence within Africa (Veeramah et al. 2012). We also considered a model where the expansions out of Africa could start earlier than the split of west Africa and the "ghost" (up to 260 kya), thus allowing the ancestors of Aboriginal Australians and/or Eurasians to descend from old lineages. In that case we still find a divergence of Aboriginal Australians and Eurasians after the split of West Africans and the "ghost", with estimates very close to those reported in Table S07.3 (not shown).

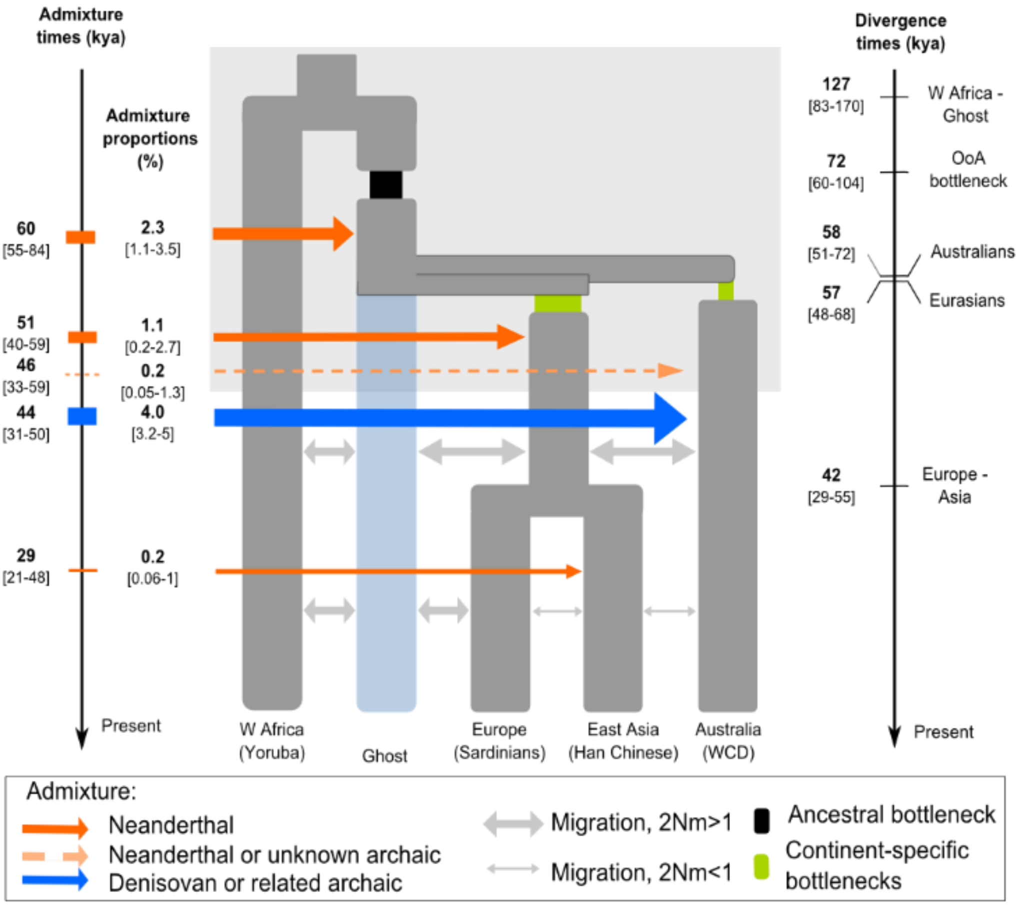


Figure S07.3 Schematic representation of the parameter estimates for a model with modern human and archaic samples and including archaic admixture. Even though the model involves two splits from the "ghost" population, which could represent the two exits out of Africa, under a *1OoA* scenario those events correspond to the divergence of Aboriginal Australians and Eurasians from the ancestral population of all non-Africans. Thus, after the ancestral bottleneck (assumed to be associated with the out of Africa event), the "ghost" corresponds to the ancestral population of all non-Africans, a population that admixed with a population related to Neanderthals, probably located in the Near East.

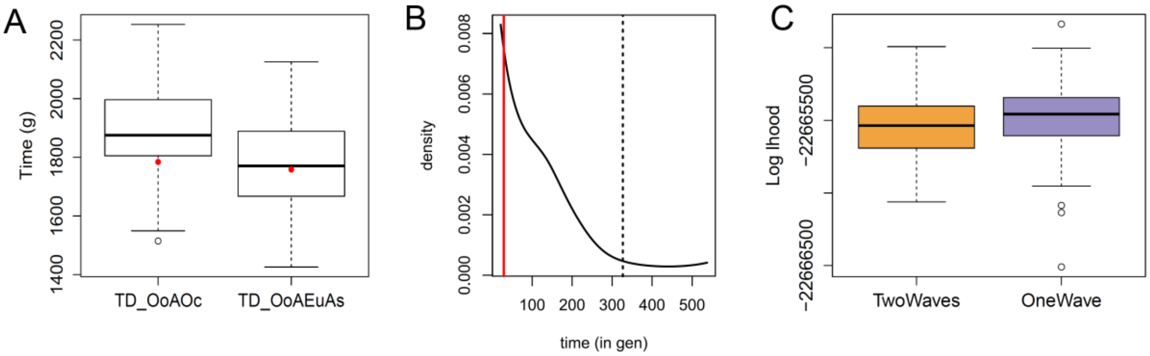


Figure S07.4 A) Distribution of the divergence times in generations obtained for the bootstrap replicates, for the Aboriginal Australian split from the ghost (TD_OoAOc) and the Eurasian split from the ghost (TD_OoAEuAs). Note that by definition we assumed that the split from the "ghost" leading to the colonisation of Australia (TD_OoAOc) had to be older than the one leading to the colonisation of Europe and East Asia (TD_OoAEuAs). Estimates obtained for the observed SFS (non-bootstrapped dataset) are shown in red. B) Distribution of the difference in divergence times (in generations) obtained for the bootstraps. The red vertical line corresponds to the estimates obtained with the observed SFS (non-bootstrapped dataset), and the dashed line corresponds to the 95% quantile of the bootstrap distribution, showing that most density is below ~320 generations. C) Comparison of the Log likelihood of full model with two splits from the "ghost" at different times (TwoWaves consistent with *2OoA*) and a nested model with two splits from the "ghost" occurring at the same time (OneWave consistent with *1OoA*). For the nested model, we used the parameters estimated for the full model, except that the two exits were both set to happen at the time of the oldest exit. Distributions were obtained from 100 expected SFS approximated with 10^6 coalescent simulations.

Table S07.2 Log-likelihood values obtained for the OoA models accounting for archaic admixture. The likelihood values are based on the parameter values that maximize the likelihood for each model, and were obtained as the mean of 100 expected SFS approximated with 10^6 coalescent simulations.

Model	**Population tree topology**	**No. of parameters**	**Log10(Likelihood)**
No archaic admixture	(SAR, HAN), WCD	25	-10,701,528
No archaic admixture	SAR, (HAN, WCD)	25	-10,706,239
Archaic Admixture	(SAR, HAN), WCD	35	-10,657,802
Archaic Admixture	SAR, (HAN, WCD)	35	-10,661,110
Archaic Admixture + possible old exit*	(SAR, HAN), WCD	35	-10,657,529
Archaic Admixture + possible old exit* + no "Ghost"**	(SAR, HAN), WCD	28	-10,660,912

*possible old exit – model where the OoA events could be older than the divergence West Africa/"Ghost"

**no "Ghost" – model without the "Ghost" population

Table S07.3 Point estimates and 95%confidence intervals for the parameters of the best OoA model with topology (SAR, HAN), WCD. Confidence intervals were calculated according to the percentile method. Point estimates correspond to the parameters inferred with the original data set. Times of divergence in years are obtained assuming a generation time of 29 years and a mutation rate of 1.25e-8/gen/site.

		95% Confidence intervals	
Parameter	**Point estimate**	**Lower limit**	**Upper limit**
Effective sizes(number of diploids)			
N_e ancestral archaics/humans	18,296	17,565	18,954
N_e Denisovan Altai	2,846	2,519	2,958
N_e unsampled Archaics (D.R and N.R)	7,419	6,094	9,476
N_e Neanderthal Altai	463	351	538
N_e West Africa (Yoruba)	27,122	21,673	51,874
N_e Ghost	4,769	2,091	43,744
N_e Europeans (Sardinians)	3,899	2,170	8,205
N_e East Asians (Han Chinese)	5,054	4,201	7,949
N_e Aboriginal Australians (WCD)	4,947	3,741	5,622
N_e ancestral Eurasians	7,264	9,114	49,747
N_e ancestral modern humans	23,275	19,905	25,533
N_e bottleneck Aboriginal Australians	136	71	263
N_e bottleneck Eurasians	1,305	108	2,170
N_e ancestral bottleneck of all non-Africans	781	192	2,221
Times of divergence			
Divergence modern humans/archaics	656,908	623,908	684,362
Divergence West Africa/“Ghost”	127,192	82,782	170,541
Time ancestral bottleneck non-Africans	72,041	60,320	103,541
Time for divergence Aboriginal Australians	57,944	51,117	72,089
Time for divergence Eurasians	57,100	47,864	67,862
Time for divergence Europeans/East Asians	41,997	28,906	54,722
Time adm. N.R. non-Africans	60,185	55,460	84,120
Time adm. N.R. Eurasians	50,864	39,634	59,304
Time adm. N.R. Asians	28,680	20,842	48,035
Time adm. N.R. Aboriginal Australians	45,862	33,007	59,222
Time adm. D.R. Aboriginal Australians	43,945	31,127	49,872
Admixture proportions			
Admixture N.R. non-Africans	0.02353	0.01132	0.03510
Admixture N.R. Eurasians	0.01148	0.00174	0.02695
Admixture N.R. East Asians	0.00228	0.00061	0.00966
Admixture N.R. Aboriginal Australians	0.00217	0.00052	0.01341
Admixture D.R. Aboriginal Australians	0.03976	0.03271	0.04996

Table S07.4 Comparison of the point estimates for the times of events in years of the best OoA model with topology ((SAR, HAN), WCD), obtained according to two time scales: (i) a higher mutation rate (1.4e-8 muts/gen/site) and a shorter generation time (25 years/gen); and (ii) a lower mutation rate (1.25e-8 mut/gen/site) and a longer generation time (29 years/gen).

	High mutation rate		**Low mutation rate**	
Parameters (Times)	**No. of generations**	**years (25y/gen)**	**No. of generations**	**years (29y/gen)**
Divergence modern Humans/Archaics	20,225	505,625	22,652	656,908
Divergence West Africans/“Ghost”	3,916	97,889	4,385	127,178
Bottleneck non-Africans	2,218	55,438	2,484	72,026
Divergence Aboriginal Australians	1,784	44,603	1,998	57,948
Divergence Eurasians	1,758	43,960	1,969	57,113
Divergence Europeans/East Asians	1,293	32,333	1,449	42,007
Admixture Neanderthal non-Africans	1,853	46,335	2,076	60,198
Admixture N.R. Eurasians	1,566	39,161	1,754	50,878
Admixture N.R. East Asians	883	22,067	989	28,669
Admixture N.R. Australians	1,412	35,301	1,582	45,864
Admixture D.R. Australians	1,353	33,817	1,515	43,935

Table S07.5 Point estimates and 95% confidence intervals for the 2Nm migration rates forward in time for the best OoA model with topology (SAR, HAN), WCD. Confidence intervals are calculated according to the percentile method. Point estimates correspond to the parameters inferred with the full data set.

		95% Confidence intervals	
Scaled migration rates (2Nm) forward in time	**Point estimate**	**Lower limit**	**Upper limit**
“Ghost” into West Africans	8.66	3.47	13.45
West Africans into “ghost”	1.52	0.21	13.27
Europe into “ghost”	3.76	1.48	53.98
“Ghost” into Europe	3.07	1.52	8.55
East Asia into Europe	0.22	0.01	0.37
Europe into east Asia	0.28	0.02	0.48
Australia into east Asia	0.52	0.02	1.08
East Asia into Australia	0.51	0.02	0.88
Ancestral Australians into Ancestral Eurasians	7.42	0.36	66.58
Ancestral Eurasians into ancestral Australians	5.05	0.06	8.16
“Ghost” into ancestral Eurasians	5.73	5.10	63.35

Assessing the fit of the SFS and other summary statistics

To visualize how well our model could reproduce the observed data we compared the marginal distribution of the observed and expected SFS (Figure S07.5). Overall, we have a very good fit of the expected to the observed marginal SFS, suggesting that our model and the corresponding parameter estimates capture relevant aspects of the data. We also looked in more detail at the joint SFS (6 dimensions, 6D) to find the entries that could not be well explained by our model. Overall, most entries of the SFS are well predicted/fitted. The entries of the SFS with the poorest fit are mostly those where most archaic and modern human samples are fixed for the derived allele, except for a few gene copies in modern humans (i.e. entries with "ancestral singletons"), which are systematically underestimated by our model. Note that these entries had relatively few SNPs and their impact on the parameter estimates should thus be limited. These are most likely cases of mis-assignment of the ancestral state. Furthermore, our model predicts less SNPs than the observed for entries where all samples are fixed for the ancestral allele, whereas Altai Denisovans are fixed derived and Aboriginal Australians have derived frequencies (allele counts) of one, two or three. In agreement with the results in S10**,** this could be due to admixture with an unknown archaic hominin related to Denisovans, as such events would increase the number of rare derived alleles in Aboriginal Australians shared with Denisovans. Finally, we find that entries related to the divergence of archaics and modern humans (e.g. fixed ancestral for Denisovan and Neanderthal and fixed derived in all modern humans, and vice-versa) tend to have a poorer fit, which is likely because we fixed some of the parameters related to archaics according to Prufer et al. (2014) (i.e., the sampling times and divergence times of Altai Neanderthal and Denisovans from the corresponding D.R. and N.R. population that contributed to modern humans) (Figure S07.6).

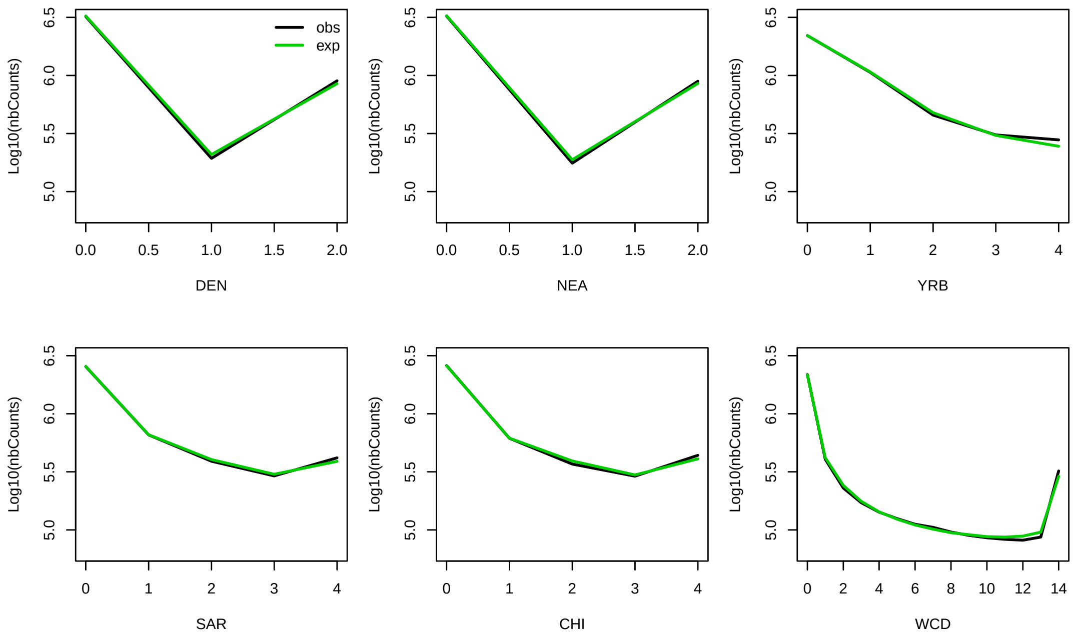


Figure S07.5 Comparison of the marginal observed and marginal expected SFS. The x-axis shows the derived allele frequencies (allele counts) and the y-axis shows the number of SNPs with a given frequency (in log10 scale). The expected SFS was obtained as the average of 100 simulated SFSs (each approximated with 10^6 coalescent simulations), according to the parameter estimates obtained with the original dataset.

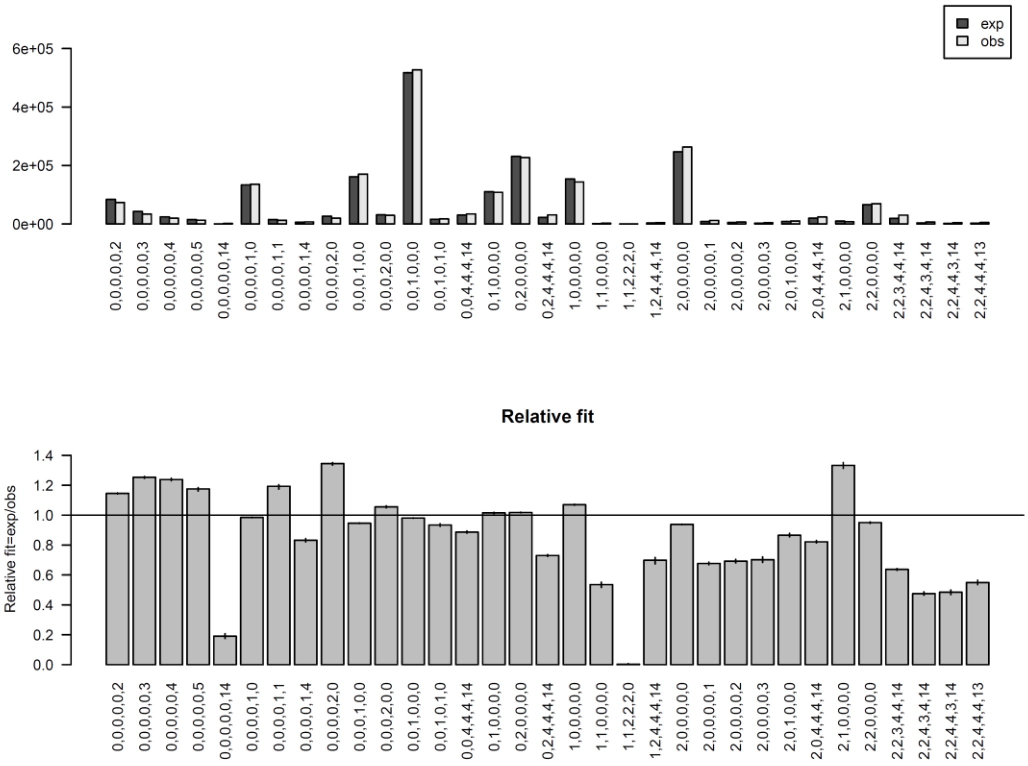


Figure S07.6 Comparison of the multidimensional joint observed and expected SFS for the 33 entries with the worst fit, out of the 16875 entries in the multidimensional joint SFS. We defined an arbitrary threshold to define entries with the worst fit, by selecting those that exhibit a difference between the expected and observed SFS larger than 500 units (i.e. $|(m_i\mathrm{Log}_{10}(p_i))\text{- }(m_i\mathrm{Log}_{10}(m_i/L)|>500$, where m_i is the observed counts at the i-th entry, p_i is the expected SFS at the i-th entry and L is the total number of polymorphic sites). Each column corresponds to one entry of the SFS, coded as *d,n,y,s,h,a* (from bottom to top) as the frequency of the derived allele in Altai Denisovans (*d*), Altai Neanderthal (*n*), Yoruba (*y*), Sardinians (*s*), Han Chinese (*h*) and WCD Aboriginal Australians (*a*). A) Comparison of the expected and observed counts (y-axis) for each entry. B) Comparison of the relative fit, defined as the relative number of SNP counts for a given entry (Relative fit= *#expSNPs/ #obsSNPs,* y-axis). Expected SFS obtained as the average of 100 simulated SFSs (approximated with 10^6 coalescent simulations), according to the parameter estimates obtained with the original dataset. Error bars correspond to the 0.01 and 0.99 quantiles of the 100 simulated SFSs.

We also computed the difference in the amount of derived sites shared among populations by means of the D-statistic (Durand et al. 2011; see S09 for further description of the D-statistic). These computations were performed on data simulated under the best model with topology ((SAR, HAN),WCD) for each of the 100 bootstrap replicates and then compared with the distribution of the D-statistics from the 100 observed bootstrap datasets. We first looked at the derived sites present in Africans (here represented by Yoruba) and computed the difference in the number of those sites that are shared with Eurasians and with Aboriginal Australians, i.e. a D-statistic of the form (H1=Aboriginal Australian, H2=European or East Asian, H3=West African, H4=chimp). Both observed and simulated data show larger derived allele sharing between Africans and Eurasians than between Africans and Aboriginal Australians (positive D-statistics, Figure S07.7A). This could be interpreted as evidence for a two waves out of Africa scenario, as Aboriginal Australians seem to be genetically more distant from Yoruba. However, the positive values of D-statistics from the observed bootstrap replicates are largely reproduced by our model, which accounts for archaic admixture and migration between modern human populations. Thus, we suggest that the larger allele sharing observed between Africans and Eurasians than between Africans and Aboriginal Australians are also compatible with a one wave out of Africa scenario followed by archaic admixture and migration among modern humans.

Moreover, we computed D-statistics by comparing the derived sites of African vs. non-African populations shared with the Neanderthals/Denisovans (Figure S07.7B). The distributions of the D-statistics related to Neanderthal (left-hand side of the graph) show the same trend in the observed and simulated datasets, with relatively lower values for Europeans, intermediate for East Asians and higher for Aboriginal Australians (with considerable overlap of simulated and observed values). For the D-statistics related to Denisovans (right-hand side of the graph), we find values close to zero in Europeans and East Asians, and values > 0.05 for the Aboriginal Australians, both for the observed and simulated values. These results indicate that our best model is able to reproduce qualitatively the general pattern observed for these D-statistics, even though they do not match exactly the observed values. One likely explanation for the difference is that we fixed some of the parameters related to the archaics according to previously reported values (Prufer et al. 2014).

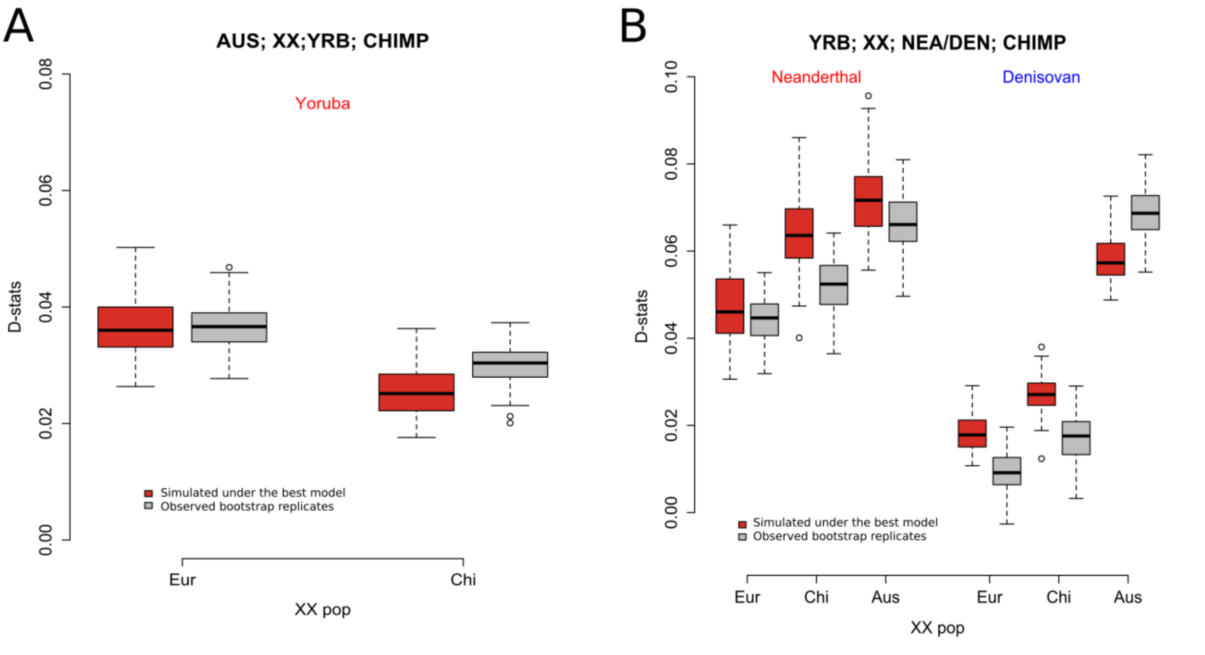


Figure S07.7 Comparison of the observed (grey) and expected (red) D-statistics, simulated according to the parameter estimates obtained across the 100 bootstrap replicates. The best model under which we obtained the expected D-stats has the following topology (SAR,HAN),WCD. In A) H1=Yoruba, H2= Sardinians/Han Chinese, H3=Neanderthal/Denisova, H4=chimp; B) H1=WCD, H2=Sardinians/Han Chinese, H3=Yoruba, H4=chimp.

Admixture with an unknown archaic population in Aboriginal Australians

There are three main lines of evidence that, taken together, might suggest that Aboriginal Australians have admixed with an unknown archaic population that could have inhabited southeast Asia and/or Sahul at the time of the colonization of that area by the ancestors of present day Aboriginal Australians. First, we estimate 4% admixture with a Denisovan-related population (D.R. in Figure S07.1B/C) that diverged ~400 kya from Altai Denisovan. Note that we fixed this divergence time to the values reported by Prufer et al. (2014). This old divergence for D.R. suggests that the population that admixed with Aboriginal Australians was very different from the Altai Denisovan sampled in Siberia. Thus, this D.R. population could actually represent an unknown archaic species/population living in southeast Asia and/or Sahul. Second, we estimate very similar times (~45 kya), for the ~0.2% Neanderthal and ~4% Denisovan introgression into Aboriginal Australians. Rather than resulting from two independent contributions, this could be due to a single admixture pulse from an unknown archaic sharing a common ancestry with Altai Denisovan and to a lesser extent with Altai Neanderthal. Third, by looking at the haplotype similarities between archaics and modern humans, we find that Aboriginal Australians have unique haplotypes that could have come from an unknown archaic, even though this would represent less than 0.1% admixture (S10).

To investigate the possibility that the D.R represents an unknown archaic related to Denisovan we tested the fit of two alternative models: a) single admixture pulse of ~4% from an unknown archaic (i.e., U.A. – unknown archaic population) that itself received input from Altai Denisovan; b) single admixture pulse of ~4% from D.R., considering that D.R. was admixed with an unknown archaic. We re-estimated parameters related to this unknown archaic admixture keeping all the other parameters fixed to the values in Table S07.3, including the admixture times and proportions. The N.R. pulse of admixture into Aboriginal Australians was removed from the two above models.

Results show that the two tested models involving admixture with an unknown archaic population reach likelihoods close to the value obtained under the full model. Slightly larger likelihoods are obtained for the model with direct admixture with an unknown archaic (U.A.). In that case we infer that the U.A. population would have received ~40% contribution from Altai Denisovan ~225 kya, and would have diverged from the ancestors of Denisovan and Neanderthal ~500 kya, close to the divergence of Denisovan and Neanderthal. For the model where D.R. could admix with an unknown archaic, we re-estimated the divergence with Altai Denisovan and infer a very similar split time (~400 kya) and a very limited contribution (~0.02%) from the unknown archaic. Furthermore, removing Neanderthal admixture into Aboriginal Australians leads to similar likelihood values, suggesting that the Neanderthal pulse inferred for the Aboriginal Australians is probably due to admixture with a population that is closer to Denisovans but also sharing a common ancestor with Neanderthals. These results suggest that we cannot discard the possibility that during the colonization of southeast Asia and/or Sahul the ancestors of present day Aboriginal Australians encountered an archaic population inhabiting that region.

Table S07.6 Likelihood of models where Aboriginal Australians are admixed with an unknown archaic. All parameters were set fixed to the ones of the full model, except parameters related to the archaic populations. Likelihood values were obtained as the mean of 100 expected SFS, each approximated with 10^6 simulations, according to the parameter values that maximized the likelihood.

Model	Estimated Log10(Lhood)
Full model with Denisovan (D.R.) and Neanderthal (N.R.) admixture	-10,657,802
Admixture of 4% with D.R. with contribution from unknown archaic. No Neanderthal admixture.	-10,657,837
Admixture of 4% with unknown archaic instead of D.R. Unknown archaic with contribution from Altai Denisovan. No Neanderthal admixture.	-10,657,817

Demography of the Australian settlement

Data preparation and processing

In addition to the seven WCD samples used for the Out of Africa analyses, we used six additional Aboriginal Australian individuals showing limited evidence of admixture with Europeans to test for alternative scenarios of the settlement of Australia (see Table S07.7). Four of the six additional individuals were sampled in two groups of Eastern Australia (Cairns area (CAI) and Northern Queensland (WPA)). The remaining two individuals were sampled in north-eastern Goldfields (WON), in south-western Australia. Also, we included two Papuan individuals from New Guinea (HGDP-Papuans), for which whole-genome data is publicly available (Prufer et al. 2014). To decrease the dimensionality of the SFS due to the very large number of populations we decided to perform parameter inference using the multidimensional SFS computed on a subset of six groups (6D-SFS) related to the settlement of Australia (Han Chinese, HGDP-Papuans, North-eastern Australians – CAI and WPA – and South-western Australians – WON and WCD). For our demographic inferences, we then used the 6D-SFS computed from the original dataset as well as from 100 datasets of autosomal SNPs (i.e., bootstrap replicates), both generated as described in the Out of Africa section.

Table S07.7 Australian samples used to test alternative scenarios of the settlement of Australia

Eastern Australia	Western Australia
CAI: CAI01, CAI06, CAI07	**WCD**: WCD01, WCD02, WCD03, WCD04, WCD05, WCD12, WCD13
WPA: WPA05	**WON**: WON03, WON06

Tested models

Based on geographic distribution of the samples, and due to the genetic differentiation detected between South-western and North-eastern Aboriginal Australians (Extended Data Figure 2, Extended Data Figure 7) and the different patterns of admixture with East Asians observed within Australia (Figure 2b, S05) we tested three alternative models for the settlement of Australia: (i) modern South-western and North-eastern Aboriginal Australians derive from a single source population that colonised the continent some time ago and then diversified into South-western and North-eastern Aboriginal Australians (*1waveAus*, (SW,NE),PNG; Figure S07.8A); (ii) modern South-western Aboriginal Australians derive from an ancestral population that first colonised Australia, whereas North-eastern Aboriginal Australians split more recently from the common ancestor with Papuans from New Guinea (*2waveAus*, (PNG,NE),SW; Figure S07.8B-top); iii) modern North-eastern Aboriginal Australians derive from an ancestral population that first colonised Australia, whereas South-western Australians derive from a more recent common ancestor with Papuans (*2waveAus*, (PNG,SW),NE; Figure S07.8B-bottom).

To better investigate the expansion leading to the settlement of Australia we modelled the relationship between populations within Australia using a hierarchical continent-island model (Slatkin, Voelm 1991; Hudson 1998), as these models have been shown to capture properties of recent spatial expansions (Excoffier 2004) and/or structured meta-populations (Wakeley, Aliacar 2001). Under this type of models, we assume that sampled populations of Aboriginal Australians represent "islands" that send migrants, backwards in time, to the corresponding north-eastern or south-western "continent" (Figure S07.8).

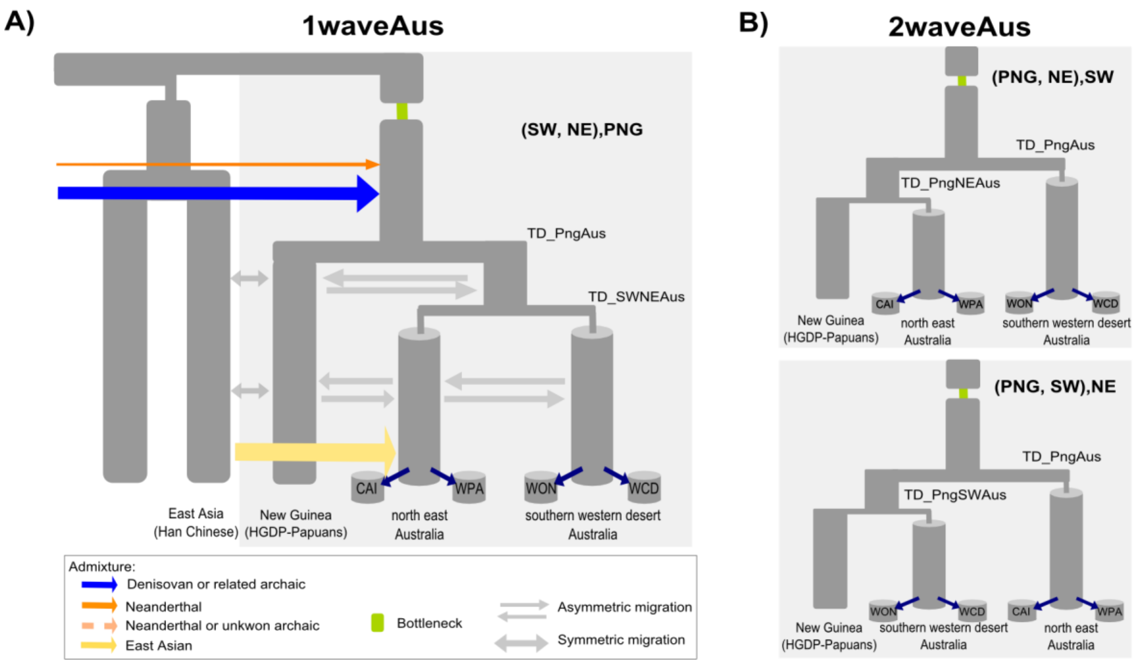


Figure S07.8 Schematic representation of the models tested for the settlement of Australia. A) *1waveAus*. B) *2waveAus* models but showing only the wave leading to the Australo-Papuans with topologies ((PNG, NE), SW) and ((PNG, SW), NE) displayed at the top and bottom pane, respectively.

To get a finer picture of the relationship of the ancestors of Australo-Papuans to Neanderthals and Denisovans, keeping the number of parameters as low as possible, we assumed that the background evolutionary model was the one inferred for the out of Africa. In other words, we modified our best out of Africa model to include Papuan samples and additional samples from Aboriginal Australians, keeping all parameters indirectly related to the evolutionary history of Australo-Papuans fixed to the inferred values in Table S07.3. As for the out of Africa models, migration was assumed to occur in a stepping-stone manner, with gene flow between East Asians and Papuans, Papuans and North-eastern Australians as well as between North-eastern Australians and South-western Australians. Nevertheless, in the two last cases, migration was assumed to be potentially asymmetrical. We also tested models where gene exchange between Papuans and North-eastern Australians was modelled as a single pulse of admixture. Given that some individuals appeared admixed with East Asians, we modelled the contact between East Asians and North-eastern Australians as a recent pulse of admixture (Figure S07.8, yellow arrow).

The parameter ranges used in our inference approach were the following: (i) for all the population effective sizes we considered ranges between 100 and 100,000 (with no fixed upper bound); (ii) for the bottleneck in the ancestors of Australo-Papuans we fixed a duration of 100 generations with effective sizes varying between 10 and 5,000 (sizes larger than 2,500 correspond to mild or no bottleneck effect) allowing it to happen somewhere between the split of Australo-Papuans from the African "ghost" population and the time of the first population split of Australo-Papuans (TD_PngAus, Figure S07.8A/B/C); (iii) for the timing of events, the divergence time between Papuans and any of the Australian "continents" (*2waveAus* models) or Papuans and the ancestors of all modern Aboriginal Australians (*1waveAus* models) was constrained to occur more recently than the divergence time between Europeans and East Asians (i.e., younger than 42 kya, Table S07.3); (iv) scaled migration rates (2*Nm*) were constrained to be between 0.0001 and 100. The times of Neanderthal and Denisovan admixture were allowed to take place during the same time interval as the bottleneck (see point ii), keeping the admixture proportions fixed to the point estimates reported in Table S07.3.

Non parametric bootstrap analysis

We performed a non-parametric block-bootstrap analysis by generating 100 datasets as described above for the out of Africa models.

Results

One wave into Australia

The comparison of alternative scenarios for the colonisation of Australia suggests that models with a single wave into Australia (*1waveAus*) have higher likelihoods (red boxes, Figure S07.9) than scenarios where the two Australian regions (northeast and southwest) would have been colonised by two independent migration events (grey and blue boxes, Figure S07.9). The likelihood distributions recovered for the alternative models display similar variances and are partially overlapping. Nevertheless, the average likelihood for models assuming one wave of colonisation into Australia (*1waveAus*) is larger than the average likelihood of *2waveAus* models, suggesting that *1waveAus* models are better supported by the data. In addition, we find more support for models where the contact between Papuans and NE Australians is modelled as continuous migration rather than as a single pulse of admixture. Indeed, the *1waveAus* and the *2waveAus* ((PNG, SW), NE), models with continuous gene flow exhibit higher likelihoods. Thus, our SFS-based approach seems to support a *1waveAus* model with continuous migration between Papuans and NE Australians. Interestingly, by looking at the maximum likelihood parameter estimates for the scaled migration rates (*2Nm*, Table S07.8), described below in further detail, we found that gene flow between Papuans and Northeast Australians has been quite reduced with *2Nm*<1. Note, however, that there is a large uncertainty on these estimates (95%CI: 0.0009-20.35), with the lower limit of the confidence interval being close to zero. Thus, to test whether continuous gene flow between these two populations is required to explain the SFS patterns we compared the likelihood distributions from *1waveAus* with and without continuous migration (the two leftmost boxplots, Figure S07.9). As one can see in Figure S07.9, the likelihood distribution of these two models are rather overlapping suggesting that one cannot completely reject a model without migration between Papuans and North-eastern Australians based on the SFS. However, given other lines of evidences for recent gene flow (see S05, S06), we performed the non-parametric bootstrap analysis for obtaining the 95% CI under this model.

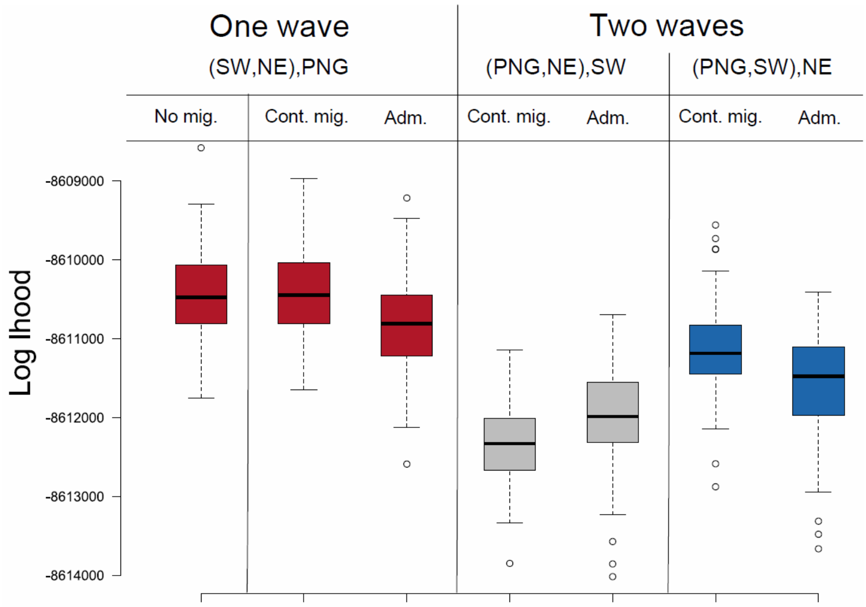


Figure S07.9 Likelihood (in log10) distributions for *1waveAus* and *2waveAus* scenarios. Distributions were obtained from 100 expected SFS approximated with 10^6 coalescent simulations, with the combination of parameters corresponding to the maximum likelihood under each model. "Cont. mig" stands for continuous migration, whereas "Adm" represents models where gene flow between Papuans (PNG) and NE Aboriginal Australians was modelled as a single pulse of admixture.

Table S07.8 Point estimates and 95% confidence intervals for the 2Nm migration rates forward in time. Confidence intervals are calculated according to the percentile method. Point estimates are the parameters inferred using the original data set.

Scaled migration rates (2Nm) Backwards in time	Point estimates	95% Confidence intervals	
		Lower limit	Upper limit
NE Austr. into SW Austr.	0.0031	0.0005	6.1587
SW Austr. into NE Austr.	0.0114	0.0005	11.2492
Papuans into NE Austr.	0.4088	0.0009	20.3506
NE Austr. into Papuans	0.0151	0.0005	1.7293
East Asia into Papuans	0.50	0.02	0.76
WCD into SW Austr. Continent	28.56	18.29	67.51
WON into SW Austr. Continent	43.04	18.03	78.01
CAI into NE Austr. Continent	47.40	18.42	73.27
WPA into NE Austr. Continent	15.97	6.64	24.09
Papuans into ancestors Austr.	0.07308	0.00003	1.44934

Parameter estimates

In Figure S07.10 we report the parameter estimates for the most relevant parameters in the model. We infer a strong bottleneck in the ancestors of Australo-Papuans, dating to approximately 50 kya (95%CI: 35-54 kya). According to our estimates, the divergence between Papuans and Aboriginal Australians seems to have occurred around 37 kya (95%CI: 25-40 kya), whereas the split between Southwestern and Northeastern Aboriginal Australians occurred ~6 ky later, approximately 31 kya (95%CI: 10-32 kya). We also find evidence for little gene flow from North-eastern to Southern-western Australians (*2Nm*<1). We estimate a very large effective size for the north-eastern population with (95%CI Ne: 117,000-748,000) and a smaller effective population size for the south-western group (95%CI Ne: 4,000-58,000). This suggests that North-eastern Aboriginal Australians are highly structured and/or exchange migrants with other populations, whereas South-western Aboriginal Australians are more isolated. All parameter estimates and corresponding 95% confidence intervals are reported in Table S07.9. Note that we estimate a very recent East Asian contribution of 18% into north-eastern Australia (95%CI: 9-29%), consistent with other analyses (S05), but we have limited power to estimate the timing of such recent events.

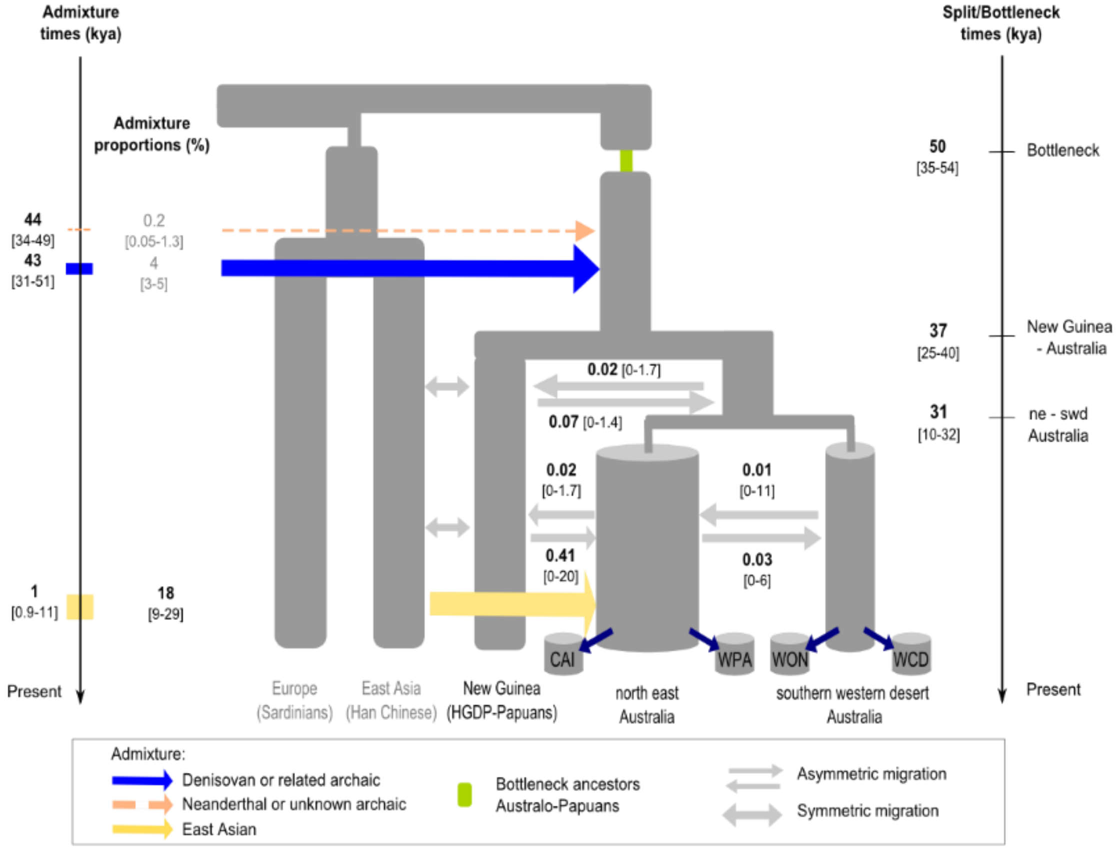


Figure S07.10 Estimates obtained for the best model for the settlement of Australia. Parameters values in black were estimated from the data, whereas all parameters not shown or shown in pale grey were fixed to the point estimates recovered under the best scenario for the Out of Africa (see details in the previous section). Note however that, while we estimated the times of admixture with the archaic populations, we kept the admixture proportions fixed to the ones obtained in the Out of Africa section. All admixture and migration events are represented forward in time.

Table S07.9 Parameter point estimates and 95%confidence intervals obtained under *1waveAus* scenario. Confidence intervals were calculated according to the percentile method. Point estimates correspond to the parameters inferred with the original data set. All other parameters were fixed to the ones estimated for the OoA model. Times of divergence in years obtained assuming a generation time of 29 years and a mutation rate of 1.25e-8/gen/site.

		95% Confidence intervals	
Parameter	**Point estimate**	**Lower limit**	**Upper limit**
Effective sizes(number of diploids)			
N_e Papuans	3,586	1,893	4,755
N_e North East Australia	34,214	65,847	418,621
N_e South West Australia	9,779	2,320	32,680
N_e ancestral Australia	6,116	2,919	44,980
N_e ancestral Australo-Papuans	1,633	817	45,306
N_e bottleneck Australo-Papuans	332	153	1,788
Times of divergence/admixture			
Bottleneck ancestors of Australo-Papuans	50,331	34,944	53,969
Time admixture Denisovan D. R.	42,983	30,814	50,815
Time admixture Neanderthal N.R.	44,305	34,196	49,332
Divergence Papuans/Aboriginal Australians	37,018	25,342	39,500
Divergence within Australia	31,422	10,150	32,421
Time admixture East Asians	1,445	990	11,233
Proportion of admixture (in %)			
Admixture East Asians	0.179	0.096	0.291

Assessing the fit of the SFS

To assess whether our model can reproduce the observed data we compared the marginal distribution of the observed and expected SFS (Figure S07.11). Globally, we found a very good fit between the expected and the observed marginal SFS, suggesting that the *1waveAus* model is able to reproduce the main patterns of the observed data. By further investigating the fit of all entries in the 6D-SFS, we find that most are well predicted under the *1waveAus* scenario (Figure S07.12). Nevertheless, as described in the Out of Africa section, we found a systematically poor fit in the joint SFS entries where all individuals, except one, are homozygous for the derived allele and the ancestral allele is present in a single copy in a heterozygous individual (i.e. “ancestral singletons”). As explained above these are most likely cases of mis-assignment of the ancestral state. Additionally, a poor fit is also found in the SFS entries corresponding to: 1) loci where the derived allele is fixed in Aboriginal Australians and private to these populations, and; 2) singletons and doubletons in Papuans shared with WPA. In these cases, the *1waveAus* scenario predicts less of those sites than observed (Figure S07.12 upper panel). Given the low number of sites in these entries, we expect this to have a very limited effect on our parameter estimates.

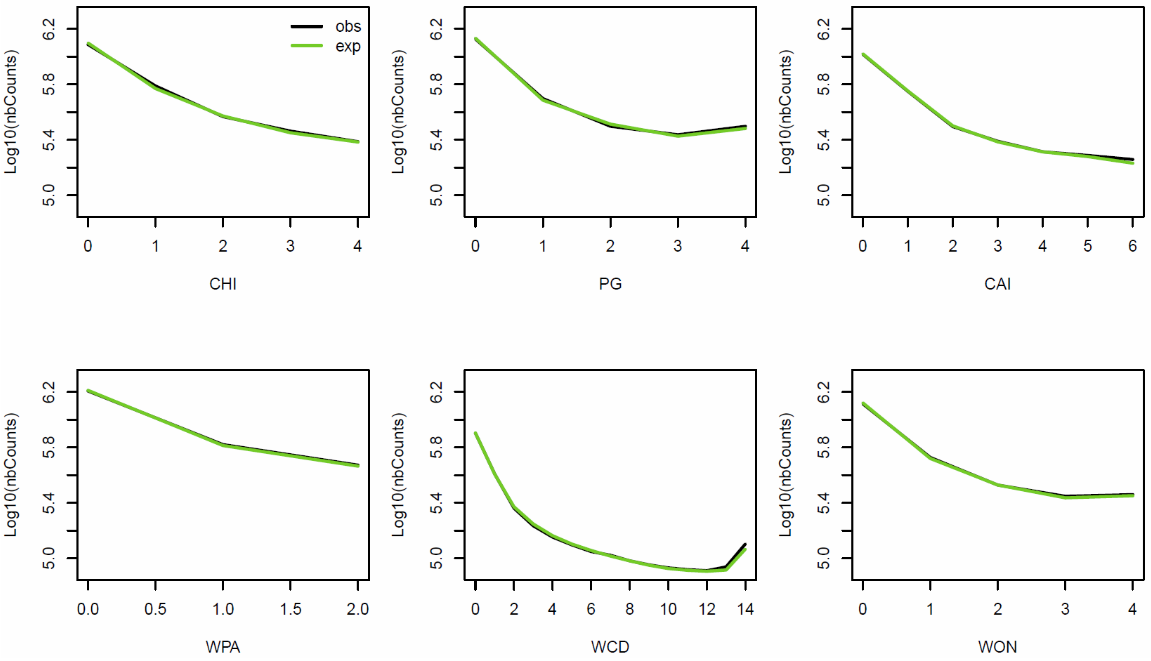


Figure S07.11 Comparison of the marginal observed and marginal expected SFS, simulated according to the parameter estimates obtained with the original dataset and under the *1waveAus* model. The x-axis shows the derived allele frequencies (allele counts) and the y-axis shows the number of SNPs with a given frequency (in log10 scale). Although our model for the settlement of Australia includes more samples than the ones shown here, we reduced the dimensionality of the SFS by computing it only on the six groups directly involved with the parameters to estimate (see Data preparation and processing for further details).

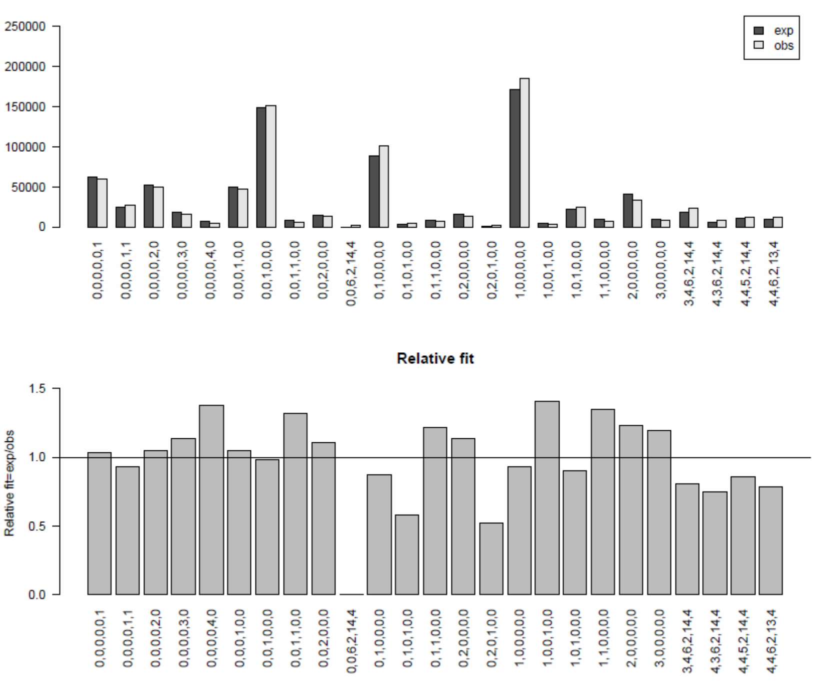


Figure S07.12 Comparison of the multidimensional joint observed and expected SFS for the 25 entries with the worst fit, out of the 39,375 entries in the 6D joint SFS (all the other entries had a reasonably good fit). We defined an arbitrary threshold to define entries with the largest discrepancy between the expected and observed counts, by selecting those that exhibit a difference between the expected and observed SFS larger than 500 units (i.e. $|(m_i\text{Log}_{10}(p_i))\text{-}\ (m_i\text{Log}_{10}(m_i/L)|>500$, where m_i is the observed counts at the i-th entry, p_i is the expected SFS at the i-th entry and L is the total number of polymorphic sites). Each column corresponds to one entry of the SFS, coded (*han*, *png*, *cai*, *wpa*, *wcd*, *won*) (from bottom to top) as the frequency of the derived allele in Han Chinese (*han*), HGDP-Papuans (*png*), CAI (*cai*), WPA (*wpa*), WCD (*wcd*) and WON (*won*) Aboriginal

Australians. A) Comparison of the expected and observed counts (y-axis). B) Comparison of the relative fit, defined as the relative number of SNP counts for a given entry (Relative fit= *#expSNPs*/ *#obsSNPs*, y-axis). Expected SFS obtained by averaging the 100 simulated SFSs (approximated with 10^6 coalescent simulations), according to the parameter estimates obtained with the original dataset.

Effect of data filtering on the SFS

The joint observed SFSs used for the demographic analyses were built based on the called genotypes (i.e., assuming that the genotype with higher likelihood was correct), after applying several quality filters for mapping, coverage and genotype calls (see data preparation sections above and S03 and S04).This was done given that all the individuals included in the demographic analyses exhibited high mean coverage (larger than ~20x, see S04, Table S04.1), suggesting that we could use the called genotypes with confidence. Even though it has been shown in simulation studies that biases on the SFS can arise and are particularly problematic for low depth of coverage datasets (<10x, Nielsen et al. 2012), such biases are minimized for depths of coverage higher than 12x (Kim et al. 2011; Fumagalli 2013; Han, Sinsheimer, Novembre 2014). For instance, Crawford and Lazarro (2012) found no biases in the sampled SFS and no influence on parameter estimates when inferring population expansion parameters with sites having ≥12x coverage.

To assess the quality of genotype calls, which could influence the SFS, we calculated the rate of concordance with the array genotype data available for the Aboriginal Australian samples across the spectrum of minor allele counts. We also calculated a “minor allele concordance” metric, which is the concordance restricted to those samples where sites are heterozygotes or homozygotes for the minor allele in the array genotype, such that we could assess specifically our ability to correctly call genotypes for rare variants. Whereas we observe a drop in minor allele concordance for the lowest minor allele counts, concordance is very high in absolute terms across all counts, with values > 98.5% (see Fig. S07.13). This high concordance reflects the high depth of coverage in our data (mean coverage >50x for Aboriginal Australians). Another contributing factor is probably that genotypes were called independently per individual (i.e. single-sample calling), such that allele frequency information did not influence the calls, as opposed to the multi-sample calling approach which is typically used with lower-coverage data (e.g. in the 1000 Genomes Project) and is associated with reduced sensitivity for rare variants.

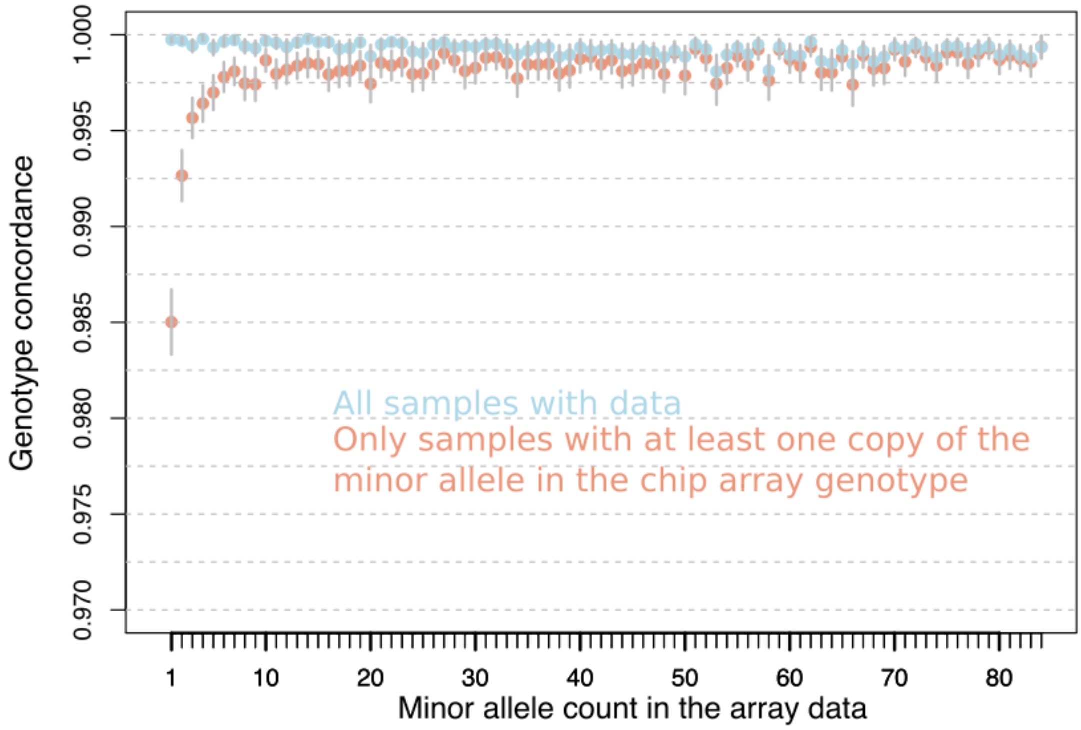


Figure S07.13 Genotype concordance between the array genotype data and the genotype calls from whole-genome of Aboriginal Australian samples.

In addition, we performed a simulation study to assess the severity of potential errors in reconstructing the SFS based on called genotypes. Also, to assess the effect of mixing datasets with different depths of coverage, we simulated NGS data from two populations, assuming that the first population (Pop 1) had a depth of coverage (DP) similar to Denisovan, and that the second population (Pop2) had a DP similar to Yoruba (Figure S07.14). Note that we refer to depth of coverage as being the number of reads per site passing mapping and base quality filters, which will be referred to hereafter as DP. To be conservative, we performed simulations based on the DP of Denisovan (mean ~25x), which is one of the individuals included in the demographic analyses who has the lowest DP. We also considered a case where the depth of coverage would be eight times smaller, with a mean of ~4x (Figure S07.14b). Next generation sequencing data was simulated following Nielsen et al.(2012), assuming that the allele frequencies of the two populations follow the Balding-Nichols model, given ancestral allele frequencies proportional to *1/x* and population-specific F_{ST} values of 0.275 and 0.05 for Pop 1 and Pop 2, respectively. For each individual, the number of reads at each site was drawn from a negative binomial rather than from a Poisson distribution, as the Negative-binomial distribution fitted the observed data very well (Figure S07.14a). The typical *p* and *s* parameters of the negative binomial distribution were estimated from the empirical mean (*m*) and variance (*v*) as $p=m/v$ and $s=m^2/(v-m)$. Errors were introduced uniformly at a rate of 1.0% and genotype likelihoods were computed following Kim et al. (2011), assuming diallelic sites and equal error rates. We considered sample sizes similar to those observed, with 2 diploids from Pop 1, and 4 diploids from Pop 2.

We considered two sampling schemes: (i) low depth of coverage data, without GQ filtering (Figure S07.15a), and (ii) depth of coverage similar to Denisovan and Yoruba, treating as missing data genotypes with a quality less than 30 (GQ<30), and keeping only sites without any missing data across individuals (Figure S07.15b). For each simulation, we generated 100,000 SNPs, randomly selecting 25,000 SNPs that passed the above filters.

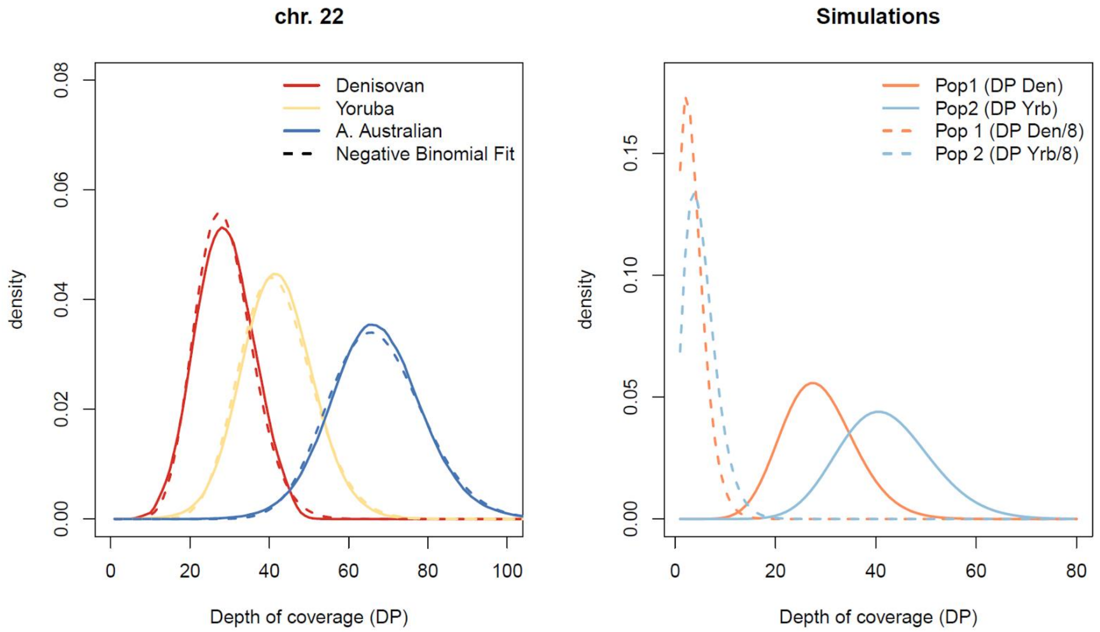
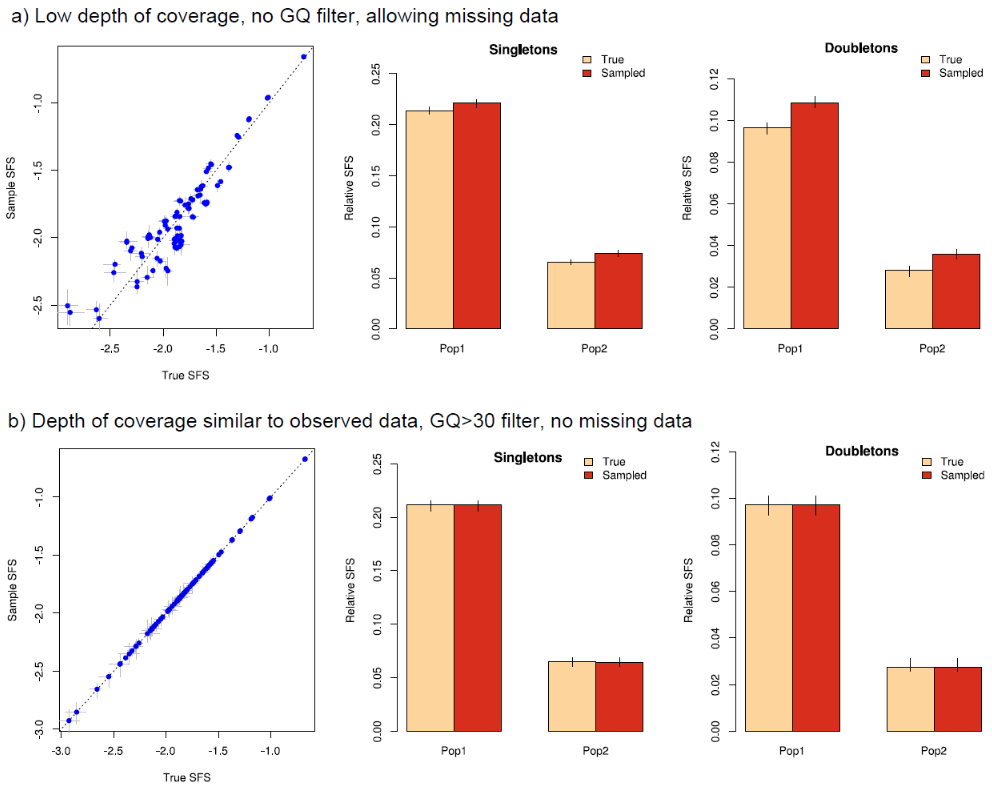


Figure S07.14 a) Distribution of the depth of coverage (DP) for three individuals used in the SFS-based demographic analyses, obtained for sites in chromosome 22 that passed coverage quality filters. As can be seen the negative binomial distribution fits well the observed data. Note that for Aboriginal Australians the DP is based on the monomorphic sites. b) Fitted negative binomial distributions to the DP of Denisovan (DP Den) and Yoruba (DP Yrb), used as reference for simulating data from Pop1 and Pop2, respectively. Dashed lines show the DP distribution used to test the effects of low coverage data (eight times smaller than DP Den and DP Yrb).

As can be seen in the Figure S07.15, our simulations show the expected excess of singletons and doubletons when the coverage is low and no genotype quality filter is applied. However, the 2D SFS is well recovered when the depth of coverage is similar to Denisovan (one of the lowest in our samples), after applying the filters used when processing the real data. Specifically, the estimated proportion of singletons and doubletons were very close to the true values used in the simulations. These results are in agreement with previous studies (Crawford, Lazzaro 2012; Fumagalli 2013; Han, Sinsheimer, Novembre 2014), who showed that when using sites with DP≥ 12x the true SFS is well approximated by the SFS built on called genotypes. Our simulation results suggest this is also the case even when populations have different depths of coverage.

Taken together, our results suggest: (i) a high concordance in genotype calls between the whole genome data and the SNP array data, even for rare variants (singletons and doubletons), and (ii) that differential DP among populations does not affect our ability to recover the SFS for high depth of coverage. We are thus confident that the quality of the genotype calls for singletons and doubletons for the dataset we used for the demographic analyses is sufficiently high and that the joint SFS is unlikely to be affected by potential biases due to differences in coverage among populations.



a) Low depth of coverage, no GQ filter, allowing missing data

b) Depth of coverage similar to observed data, GQ>30 filter, no missing data

Figure S07.15 Comparison of the true 2D-SFS with the estimated 2D-SFS for samples with differential DP across populations. a) Simulations with low DP (mean of ~4x for Pop1, ~5x for Pop2), without filtering steps. b) Simulations with DP similar to observed data (mean ~28x for Pop1, ~42x for Pop2). In the left panels each point indicates the mean, and the error bars indicate the range of values obtained over 10 simulated 2D SFSs. Mid panels: Comparison of the true and inferred proportion of singletons in each population. Right panels: Comparison of the true and inferred proportion of doubletons. The 2D-SFS was simulated for two diploids from Pop1 and four diploids from Pop2, with an error rate of 1.0% (Nielsen et al. 2012) which is larger than the Illumina HiSeq X Ten (Ross et al. 2013) data but also mimics mapping errors. For each simulation 100,000 SNPs were generated, but to keep the same number of SNPs across all simulations, we randomly selected (resampling) 25,000 sites that passed the filters.

Parameter estimates validation - simulation study

Archaic admixture model

To assess whether the SFS-based approach used in this study is able to provide accurate estimates of the parameters under the different models, we simulated 50 SFSs according to a model (see below) and re-estimated model parameters using the simulated SFSs as observed datasets. Focusing more specifically on our ability to estimate the intensity and timing of admixture events between archaic and modern humans, we simulated SFSs under a model involving a sampled archaic population (SA) related to an unsampled population (UNSA) that contributed 2% to the gene pool of the ancestors of modern humans (MH, Figure S07.16) 1,724 generations ago (50 kya, assuming a generation time of 29 years Fenner 2005). Moreover, the MH ancestors went through a bottleneck 1600 generations ago (46.4 kya), thus just before (backwards in time) the admixture event with the UNSA population. The properties of the modelled admixture event were chosen to resemble the admixture event between Neanderthals and modern humans inferred in the previous sections. We assumed that the bottleneck occurred before the admixture pulse from Neanderthals into the ancestors of modern humans, whereas the out of Africa bottleneck inferred in the Out of Africa section was estimated to have occurred after the admixture pulse. Including a bottleneck before the admixture reduces the amount of information available for parameter inference and thus we expect admixture times and proportions to be more difficult to estimate. Note that during the estimation procedures, both with simulated and observed data, admixture can occur before or after the bottleneck event. We also used the estimates provided by Prufer et al. (2014) for the split time between the sampled and unsampled archaic populations (3,820 generations ago), as well as for the age of the sampled archaic population (2,957 generations ago). The split time between archaics and modern humans was set approximately to that estimated in the Out of Africa section, 23,000 generations (~667 kya, Table S07.3). This model is a simplification of that used to study the Out of Africa and colonisation of Australia (see sections above), but it should capture the main features of modern human demographic history and it remains computationally tractable for parameter validation purposes.

To simulate genetic data close to those used to infer the demographic history of the Out of Africa and the colonisation of Australia we simulated 1000 blocks of 1Mb using a mutation rate of 1.25e-8/gen/site (Scally, Durbin 2012). To keep the same proportion between mutation and recombination rate used in MSMC analysis (S08), the recombination rate was set to 1.12x10^{-8}/generation/site. The sample size of the archaic population was set to one diploid individual whereas the sample size of the modern humans was set to four diploid individuals.

Maximum likelihood parameters

We inferred model parameters by considering the simulated SFSs as pseudo-observed data. We estimated the set of parameters that maximize the likelihood for each SFS by using search ranges similar to those used to infer the colonization of Australia and the Out of Africa demographic history. Therefore: (i) for all the population effective sizes, we considered ranges between 100 and 100,000 haploids, but with an open upper bound that is extended if parameters get close to the boundary during the ECM optimization; (ii) for the bottleneck, we fixed to 100 generations with effective sizes varying between 10 and 5,000 haploids. Based on estimates shown in Table S07.3 we set (iii) the time of split between archaic and modern human populations (Tdiv_ARCvsMH) to vary between 15,000 and 25,000 generations; (iv) the bottleneck in the ancestors of modern humans (T_Bot) to vary between the present and the time of split between archaic and modern human populations. The range of the admixture time was set to vary between 1 and 6,900 generations and the range for the admixture proportion was set between 0.0001 and 0.10. The divergence time between the sampled and unsampled archaic populations was, similarly to what was done in the Out of Africa section, fixed to 3,820 generations.

The expected SFS used for the computation of the likelihood of a given set of parameters was obtained by performing 100,000 coalescent simulations. Following what was done for the block bootstrap analysis in the previous sections, we performed 10 optimization runs starting from different initial conditions and selected the run leading to the highest likelihood to get parameter estimates.

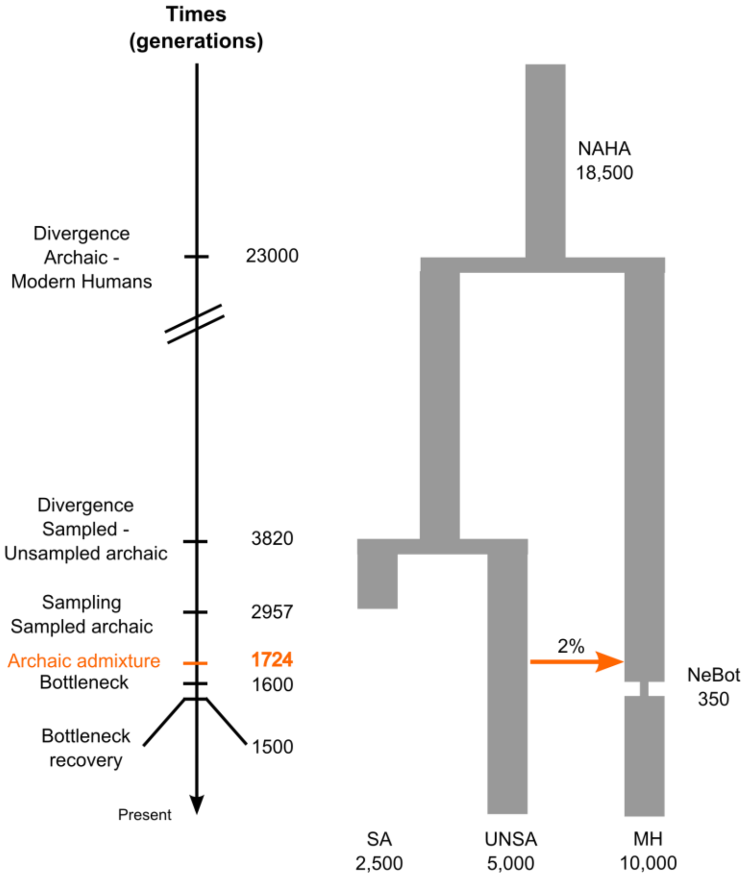


Figure S07.16 Schematic representation of the model used to validate parameter estimations. We simulated 50 datasets under this model from which we re-estimated model parameters. The model is a simplification of that used to infer the demographic history of the Out of Africa and the colonisation of Australia. The parameters related to the admixture event (orange), such as the admixture proportion (admProp) and time (T_adm), were set approximately to those inferred for Neanderthals. Effective population sizes are shown in number of diploids below the population labels, and the times are shown in generations on the left.

Results

The distribution of parameter estimates obtained from 50 SFSs simulated under the model shown in Figure S07.16 suggests that we have a relatively good power to infer model parameters as most of the true parameter values fall within the corresponding parameter distributions. It is worth noticing that we can accurately infer the admixture proportion (admProp) with the median and the mean of the distribution being very close or even equal (2% and 1.8%, respectively) to the true value (Table S07.10). For the time of admixture (T_adm), the mean and the median of the distribution were estimated to be 2,257 and 2,023 generations, respectively, and therefore ~300-500 generations older than the real parameter value. This may suggest that the SFS-based method used here provides an overestimation of the real admixture time. However, the error estimated through this simulation study is still within the non-parametric block bootstrap analysis performed to obtain confidence intervals for the parameter estimates from the real data.

Note that in the models used to infer the demographic history of the colonisation of Australia and the Out of Africa the lowest number of lineages informative about the pulse of Neanderthal admixture in the ancestors of all non-Africans and in the ancestors of Eurasians is equal to four diploid individuals. Indeed, for such a sample size of the modern human population and an admixture proportion of 2%, the probability that a lineage migrates from modern humans to the unsampled archaic population is very low. Thus, the number of coalescent simulations containing information on admixture may be limiting. To compensate this limitation and the larger number of model parameters in the models explored with the real data we used 500,000 simulations to compute the composite-likelihood from the expected SFS. With respect to the Denisovan pulse of admixture a larger Aboriginal Australians (n=7 diploids) provide, thus, more information on the estimation of admixture-related parameters.

In conclusion, our approach seems unbiased regarding the estimates of the admixture proportion and admixture times, but these estimates might be associated to relatively wide confidence intervals, as already visible in our inferences (see Tables S07.3 and S07.9).

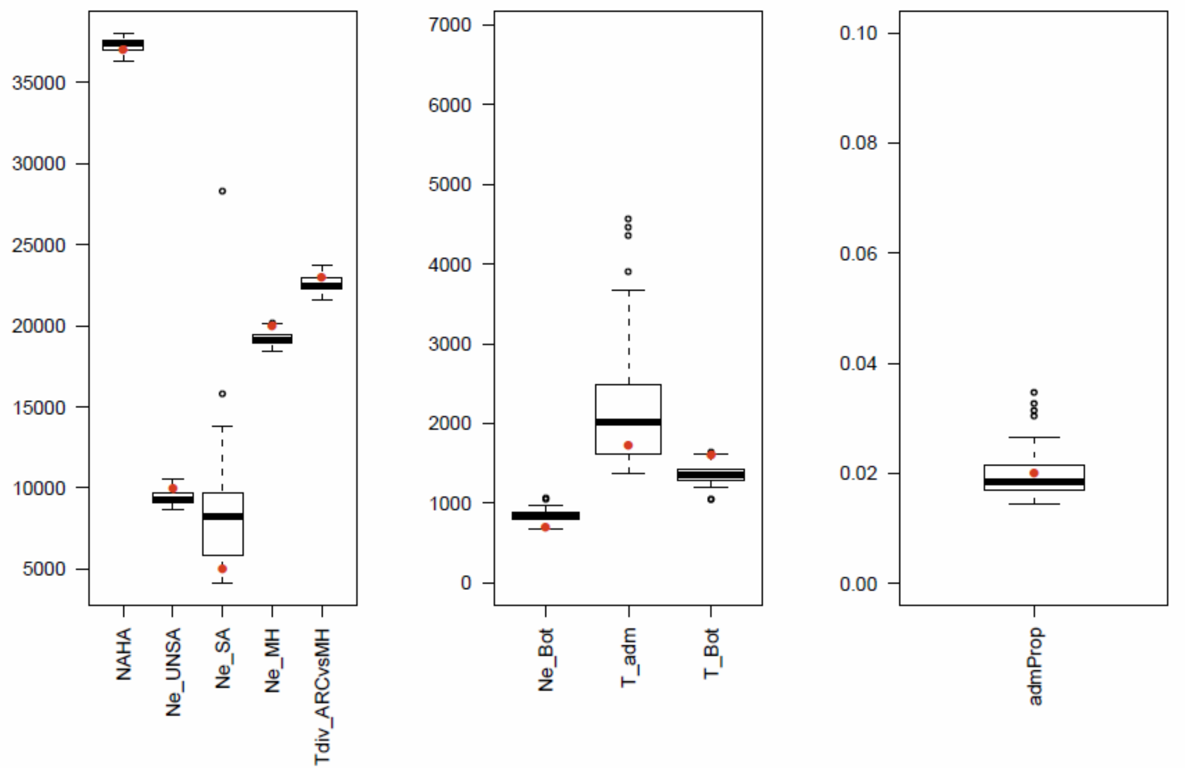


Figure S07.17 Parameter validation. Boxplots showing the distribution parameters estimated from 50 multidimensional SFSs simulated under the archaic admixture model shown in Figure S07.16.True parameter values are shown as red dots. Effective population sizes are shown in number of haploid individuals.

Table S07.10 Inferred parameter distribution properties. Model parameters were estimated from 50 multidimensional SFSs simulated under the archaic admixture model shown in Figure S07.16.

					Quantiles	
Parameter	**True**	**Mean**	**Median**	**25%**	**75%**	**95%**
NAHA	18,500	18,669	18,696	18,524	18,802	18957
Ne_UNSA	5,000	4,719	4,648	4,538	4,855	5173
Ne_SA	2,500	4,225	4,109	2,947	4,847	6695
Ne_MH	20,000	9,596	9,577	9,471	9,715	9872
Ne_Bot	350	426	424	401	443	481
T_adm	1,724	2,257	2,023	1,623	2,494	4153
Tdiv_ARCvsMH	23,000	22,612	22,492	22,249	22,943	23633
T_Bot	1,600	1,358	1,356	1,295	1,424	1606
admProp	0.02	0.020	0.018	0.017	0.021	0.031

S07 References

Abecasis, GR, A Auton, LD Brooks, MA DePristo, RM Durbin, RE Handsaker, HM Kang, GT Marth, GA McVean. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

Adams, AM, RR Hudson. 2004. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. Genetics 168:1699-1712.

Brent, RP. 1973. Algorithms for Minimization without Derivatives. Englewood Cliffs, NJ: Prentice-Hall.

Crawford, JE, BP Lazzaro. 2012. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. Frontiers in genetics 3.

Cunningham, F, MR Amode, D Barrell, et al. 2015. Ensembl 2015. Nucleic Acids Res 43:D662-669.

Davison, AC, DV Hinkley. 1997. Bootstrap methods and their application: Cambridge University Press.

Durand, EY, N Patterson, D Reich, M Slatkin. 2011. Testing for ancient admixture between closely related populations. Molecular Biology and Evolution 28:2239-2252.

Eriksson, A, A Manica. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proceedings of the National Academy of Sciences 109:13956-13960.

Excoffier, L. 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. Molecular Ecology 13:853-864.

Excoffier, L. and H.E. L. Lischer (2010) Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol. Eco. Res. 10: 564-56

Excoffier, L, I Dupanloup, E Huerta-Sanchez, VC Sousa, M Foll. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet 9:e1003905.

Fenner, JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. American Journal of Physical Anthropology 128:415-423.

Flicek, P, MR Amode, D Barrell, et al. 2011. Ensembl 2011. Nucleic Acids Res 39:D800-806.

Fumagalli, M. 2013. Assessing the effect of sequencing depth and sample size in population genetics inferences. PLoS ONE 8:e79667

Green, RE, J Krause, AW Briggs, T Maricic, U Stenzel, M Kircher, N Patterson, H Li, W Zhai, MH-Y Fritz. 2010. A draft sequence of the Neandertal genome. Science 328:710-722.

Gutenkunst, RN, RD Hernandez, SH Williamson, CD Bustamante. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695.

Han, E, JS Sinsheimer, J Novembre. 2014. Characterizing bias in population genetic inferences from low-coverage sequencing data. Molecular biology and evolution 31:723-735.

Hudson, RR. 1998. Island models and the coalescent process. Molecular Ecology 7:413-418.

Kim, SY, KE Lohmueller, A Albrechtsen, Y Li, T Korneliussen, G Tian, N Grarup, T Jiang, G Andersen, D Witte. 2011. Estimation of allele frequency and association mapping using next-generation sequencing data. BMC bioinformatics 12:231.

Meng, XL, DB Rubin. 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80:267-278.

Meyer, M, M Kircher, MT Gansauge, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222-226.

Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154:931-942.

Nielsen, R, T Korneliussen, A Albrechtsen, Y Li, J Wang. 2012. SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. PLoS ONE 7:e37558.

Prufer, K, F Racimo, N Patterson, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43-49.

Racimo, F, S Sankararaman, R Nielsen, E Huerta-Sánchez. 2015. Evidence for archaic adaptive introgression in humans. Nature Reviews Genetics 16:359-371.

Rasmussen, M, X Guo, Y Wang, KE Lohmueller, S Rasmussen, A Albrechtsen, L Skotte, S Lindgreen, M Metspalu, T Jombart. 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. Science 334:94-98.

Rosenbloom, KR, J Armstrong, GP Barber, et al. 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 43:D670-681.

Ross, MG, C Russ, M Costello, A Hollinger, NJ Lennon, R Hegarty, C Nusbaum, DB Jaffe. 2013. Characterizing and measuring bias in sequence data. Genome Biol 14:R51.

Sankararaman, S, N Patterson, H Li, S Pääbo, D Reich. 2012. The date of interbreeding between Neandertals and modern humans.

Scally, A, R Durbin. 2012. Revising the human mutation rate: implications for understanding human evolution. Nature Reviews Genetics 13:745-753.

Slatkin, M, L Voelm. 1991. FST in a hierarchical island model. Genetics 127:627-629.

Varin, C. 2008. On composite marginal likelihoods. AStA Advances in Statistical Analysis 92:1-28.

Varin, C, N Reid, D Firth. 2011. An overview of composite likelihood methods. Statistica Sinica 21:5-42.

Veeramah, KR, D Wegmann, A Woerner, FL Mendez, JC Watkins, G Destro-Bisol, H Soodyall, L Louie, MF Hammer. 2012. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Molecular biology and evolution 29:617-630.

Vernot, B, JM Akey. 2014. Resurrecting surviving Neandertal lineages from modern human genomes. Science 343:1017-1021.

Wakeley, J, N Aliacar. 2001. Gene genealogies in a metapopulation. Genetics 159:893-905.

Wall, JD, MA Yang, F Jay, SK Kim, EY Durand, LS Stevison, C Gignoux, A Woerner, MF Hammer, M Slatkin. 2013. Higher levels of Neanderthal ancestry in East Asians than in Europeans. Genetics 194:199-209.

S08 MSMC analysis

Stephan Schiffels

Brief introduction to MSMC2

For this analysis, we used MSMC2, an updated version of the program published in (Schiffels, Durbin 2014). A detailed publication about this new version is in preparation. MSMC2 differs from MSMC in the following aspect: whereas MSMC estimates the coalescence time to the first common ancestor of any two of multiple haplotypes, MSMC2 runs a separate Hidden Markov Model (HMM) across every possible pair of haplotypes, so it estimates only pairwise times to common ancestors. In this analysis, we used 4 haplotypes from 2 diploid individuals in each run, so the core process consists of 6 HMMs per chromosome. In each step of the expectation-maximization algorithm (see Schiffels, Durbin (2014)), the results from all HMM runs across all chromosomes and haplotype pairs are combined to yield a new parameter estimate in the maximization step.

Similarly to MSMC, MSMC2 yields more recent estimates than PSMC (Li, Durbin 2011), which is implemented via a finer time-discretization in recent times. Note that the time-discretization depends on the number of haplotypes. In contrast to MSMC however, MSMC2 can estimate coalescence rates (i.e. population sizes) also in the more ancient times, since it uses the entire distribution of pairwise coalescence times across all haplotypes, not just the distribution of first coalescence times across multiple haplotypes.

Data processing

We used the whole-genome sequencing data which was (i) phased within the samples in this study while using an existing reference panel, (ii) masked using a global mappability mask (as suggested in Li, Durbin (2011), S03). For all individuals used in the MSMC analysis, we ran the “vcfCaller.py” script (from *www.github.com/stschiff/msmc-tools*) to generate single sample vcfs and masks. We ran MSMC on pairs of samples, and in each case we use the “generate_multihetsep.py” script from the same repository as above, to generate the combined input files for MSMC.

Samples used in this analysis are:

- CAI: CAI01, CAI07
- PIL: PIL06, PIL07
- WCD: WCD01,WCD03
- WON: WON03, WON06
- ECCAC Aboriginal Australians: SS6004477, SS6004478 (Prufer et al. 2014)
- Papuans: HGDP00542, SS6004472
- Yoruba: HGDP00927, SS6004475
- Sardinian: HGDP00665
- Han: HGDP00778
- San: HGDP01029, SS6004473
- Dinka: DNK02, SS6004480

Real time scaling

All results from MSMC were scaled using a mutation rate of 1.25×10^{-8} per basepair per generation (Scally and Durbin 2012), and a generation time of 29 years (Fenner 2005).

Population size inference

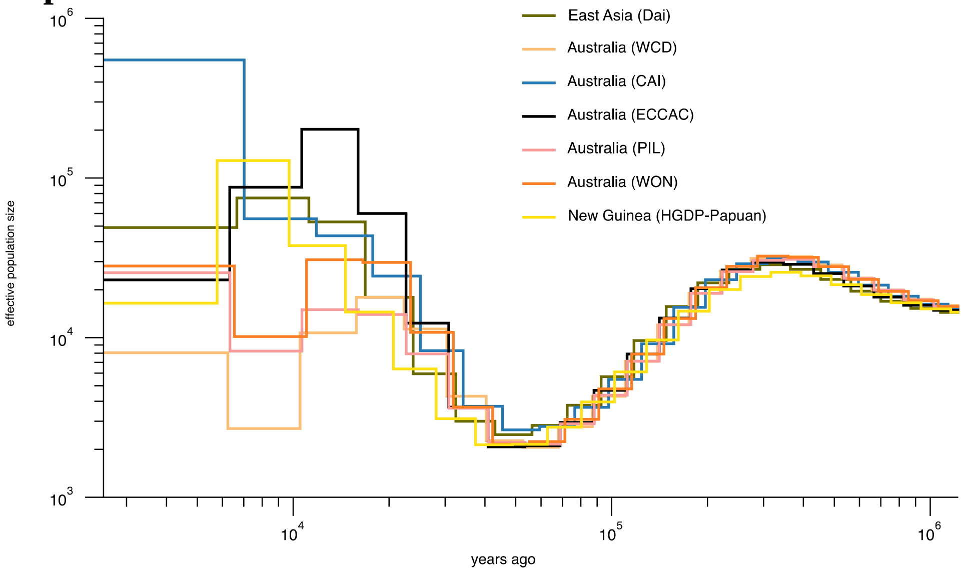


Figure S08.1 Population size estimates from MSMC for pairs of individuals from several populations within and outside of Australia. For each run we use two individuals from each population, i.e., we used four haplotypes in each run.

We find that all populations analyzed here (Aboriginal Australians and East Asians) share a similar population size history before about 20,000 years ago (Figure S08.1 and Extended Data Figure 6a). In particular the population sizes in all non-Africans share a deep and very similar bottleneck around 60,000 years ago. In the more recent past, we observe the following patterns. In the three Aboriginal Australian populations WON, PIL and WCD, we observe a decline of population sizes after 20,000 years ago to about N_e=2,000 (WCD) and N_e=10,000 (WON and PIL). In contrast the population from Cairns, CAI, exhibits a steady increase in population size. These results can be interpreted in a number of ways. For example, one would expect such a result if Australia was peopled from the North East followed by a series of founder effects through the rest of the continent (this hypothesis is supported by the SFS based analysis). We would also expect a similar pattern if there had been recent admixture from Asia and Papua New Guinea – a gene flow that is not only expected given the ethnographic information for those samples (S02) but also by the sNMF results and similar analyses (S05). Indeed, it has been shown previously (Li, Durbin 2011), that admixture can lead to inflated population size estimates with this methodology. The two ECCAC Aboriginal Australian samples published by (Prufer et al. 2014) show a similarly strong increase in population size as the CAI population, but with a recent decline starting at 15,000 BP. This may be caused by consanguinity in these samples, which would lead to artificially low population size estimates in recent times.

Divergence from Africans

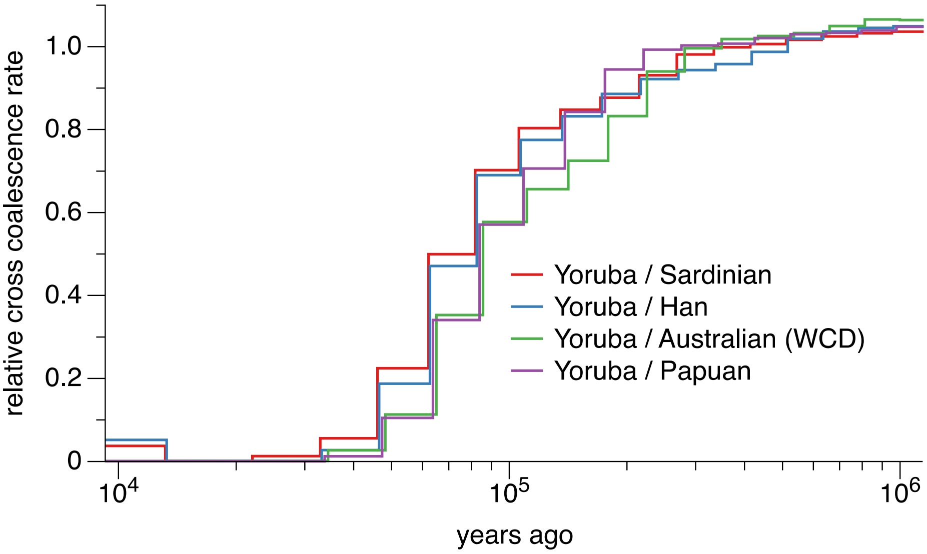


Figure S08.2 Relative cross coalescence rate estimates from MSMC for pairs of individuals of one Yoruban and one other sample, as indicated in the legend.

We generated relative cross coalescence rate (CCR) estimates, which are indicative of the nature and timing of the genetic divergence between populations. We observe a clear difference in the genetic separation time of the European and Asian populations from Yoruba, compared to the genetic separation time of the Papuan and Aboriginal Australian populations from Yoruba (Figure S08.2). The difference between the two separation times is about 10,000 years (obtained by interpolating the CCR estimates through the mid-points of the time intervals). We think this signal is explained in part by a technical artifact caused by insufficient phasing quality in the Australo-Papuan samples, see below. In part, however, we think this signal may reflect post-split gene flow between the ancestors of Yoruba and the ancestors of Eurasians, which would increase the CCR between these two ancestral populations but not between Yoruba and Aboriginal Australians/Papuans. Interestingly, we do not observe this separation when we consider divergence from San instead of Yoruba (Figure S08.3 and Extended Data Figure 4c).

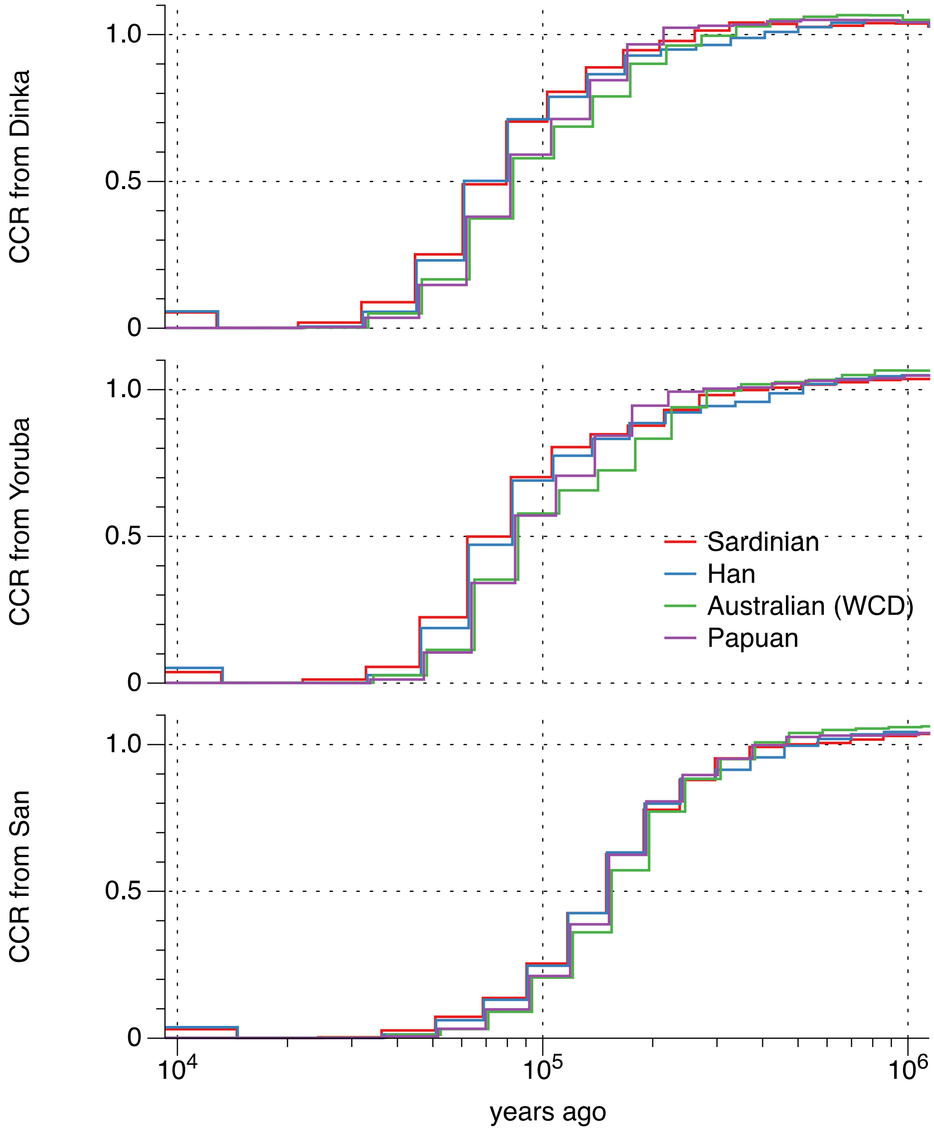


Figure S08.3 Divergence from African populations. Relative cross coalescence rate between three African populations (San, Yoruba and Dinka) to Sardinian, Han, Aboriginal Australian and Papuan samples.

As expected, the three cases result in different divergence times to out-of-Africa populations. The divergence from San is oldest, with a mid-point around 150,000 years BP. The divergence from Dinka is very close to the divergence from Yoruba, with a mid-point around 80,000 years BP (Yoruba) and 70,000 years BP (Dinka). The separation between the Aboriginal Australian divergence and the Eurasian divergence, as seen using Yoruba, is also seen using Dinka. Using San, the four CCR curves look more similar, with the Aboriginal Australian divergence being slightly below (earlier than for) the other three. This suggests that post-split gene flow between Eurasian and African ancestral populations occurred with an African population related to Dinka and Yoruba, but not to San.

We do not think that this separation of the CCR for Australo-Papuan/Yoruba versus Eurasian/Yoruba (or Australo-Papuan/Dinka versus Eurasian/Dinka) provides evidence for a separate split from African ancestors for Australo-Papuan and Eurasian ancestors. Indeed, the population sizes in all non-Africans share a deep and very similar bottleneck around 60,000 years ago, consistent with a single ancestral population splitting from Africa.

Divergence among Papuans and Aboriginal Australians

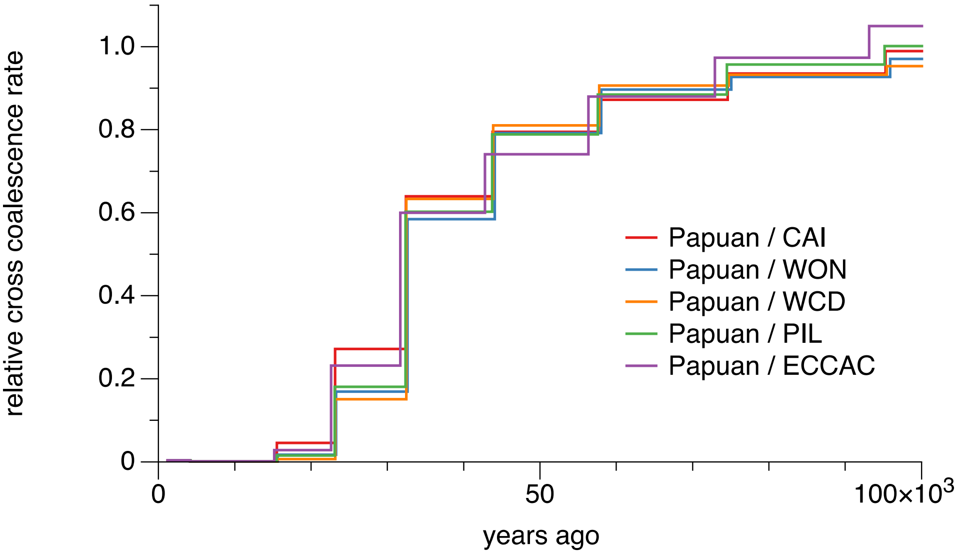


Figure S08.4 Relative cross coalescence rate estimates from MSMC using pairs of individuals of one Papuan and one other individual as indicated. We used CAI01, PIL06, WCD01 and WON03 for this analysis.

Next, we analyzed the divergence between Aboriginal Australian and Papuan samples. Our results suggest that Aboriginal Australian populations split from Papuans (Figure S08.4, and Extended Data Figure 2c for a version smoothed by interpolating the CCR estimates through the mid-points of the time intervals) mostly between 30,000 and 40,000 years ago, with CAI and the two Aboriginal Australian ECCAC individuals diverging slightly more recently than WON, PIL and WCD (e.g. consistent with low levels of post-divergence Papuan gene flow (see S06 and S07)). Finally, we analysed separations inside of Australia, as shown in Figure S08.5.

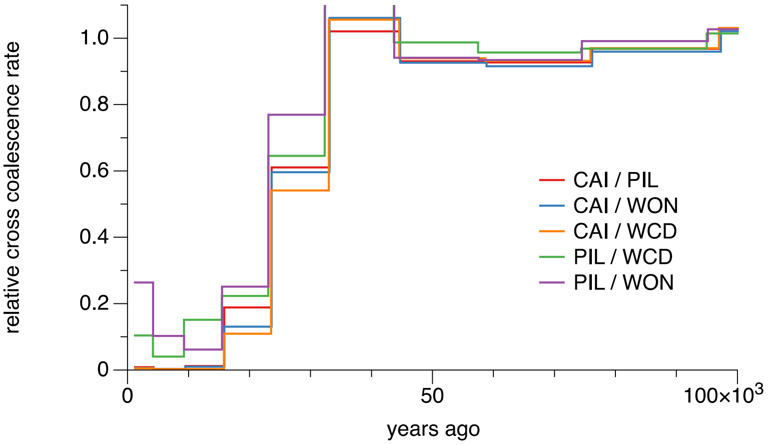


Figure S08.5 Relative cross coalescence rate estimates from MSMC using pairs of individuals of populations within Australia.

CCR estimates within Australia are suspiciously similar in this analysis, suggesting a split between all pairs of populations happening around 25,000 years BP (Figure S08.5). We caution that insufficient phasing quality in the Aboriginal Australian samples may cause an inflation of relatively recent cross divergences. In particular, the reference panel used to assist phasing does not include any Aboriginal Australian samples, so we believe that phasing switch errors may be too frequent to yield the relatively long haplotypes needed for estimates at times more recent than 20,000 years BP. This also may explain the artifact of a lower cross coalescence rate beyond 50,000 years BP seen in Figure S08.5.

Archaic admixture

To study the impact of archaic admixture, we set up a simple simulation using SCRM (Staab et al. 2015) with three populations. Two modern human populations split at 70,000 years ago, and at 35,000 years ago an archaic human population (itself diverged 400,000 years ago from modern humans) admixes into one of the two modern human populations. All population sizes remain fixed. We varied the proportion of admixture and estimated CCR curves between the two modern human populations. The results are shown in Figure S08.6 (and Extended Data Figure 4d).

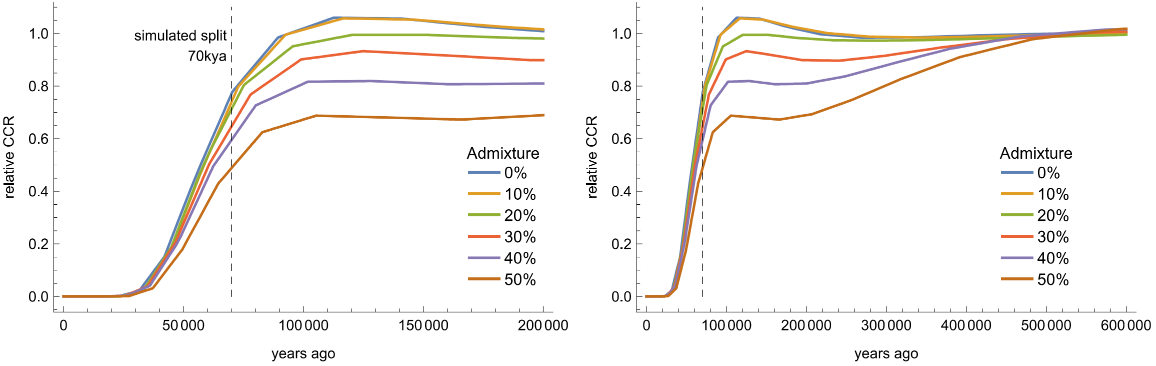


Figure S08.6 Simulations of archaic admixture into modern humans. The curve shows CCR estimates from MSMC2 between two populations in the case where one population receives gene flow from an archaic population. The archaic admixture proportions are given in the legend. The right panel shows the same data on a larger scale. Only strong archaic admixture of at least 20% can explain differences in CCR estimates with MSMC2. To better visualize small differences we show the curves after linear interpolation between the mid-points intervals.

We conclude that archaic admixture cannot explain the difference seen in CCR estimates of Non-African from African ancestors.

The impact of phasing errors

To better understand the separation between the CCR curves seen between African and Non-African ancestors, we tested the effects of differential phasing quality. We expect the Aboriginal Australians and Papuans to be phased more poorly in the real data because of a lack of related reference samples in the 1000G panel used for phasing (see S03). This could in part explain the separation of CCR estimates of non-Africans from Dinka and Yoruba (Figure S08.3). To test this we first simulated two populations that split 70,000 years ago. We then simulated phasing switch errors by walking along each simulated diploid chromosome pair and switching the phase of individuals (simulated as two haplotypes) at exponentially distributed switch points. We then applied MSMC2 to estimate CCR curves between the two simulated populations. Figure S08.7 shows the results, indicating the inverse rate of switch errors in the legend (the smaller the distance in bp between switch errors, the more frequent the errors).

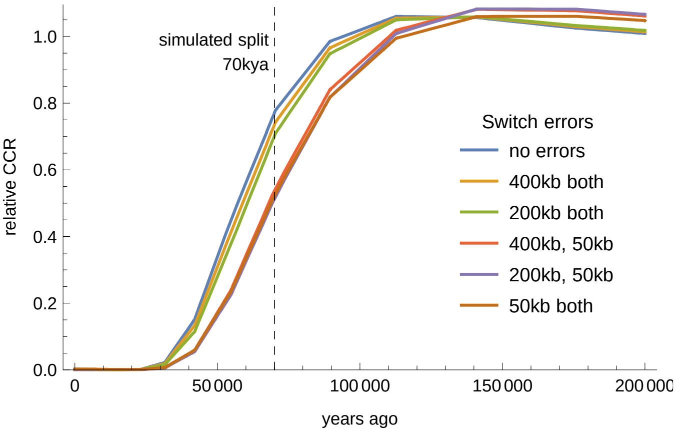


Figure S08.7 Impact of phasing switch errors on MSMC2 CCR rates. We introduced phasing switch errors to simulated data from two populations that split 70,000 years ago. The average distance between switch errors is given in the legend.

We find that phasing switch errors can have a substantial effect on CCR rates, in particular if one of the two involved populations has switch errors as often as 50,000 bp (red, purple and brown curves in Figure S08.7).

To investigate this further, we rephased our data without a reference panel. This should result in poorer phasing in general and similar phasing quality for all samples including the Aboriginal Australians and Papuans. We generated CCR estimates for this data set and the results are shown in Figure S08.8.

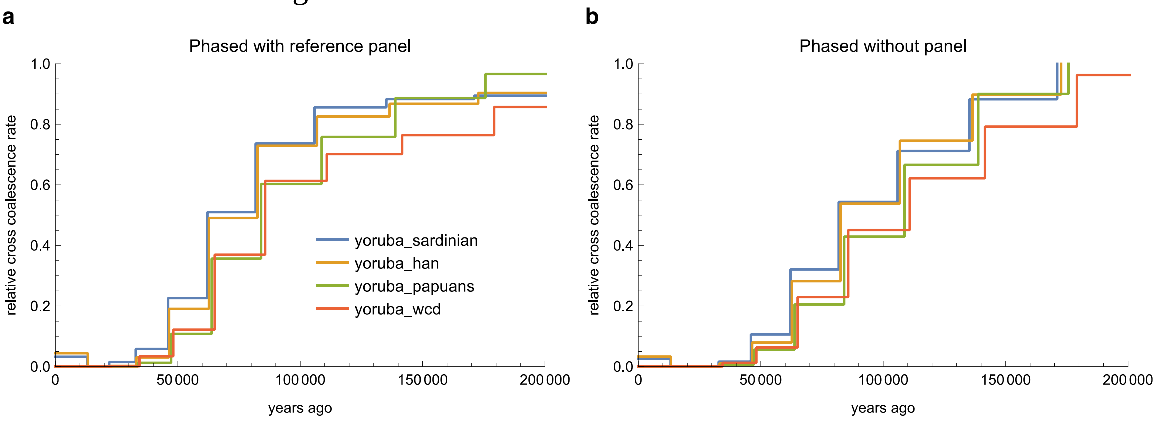


Figure S08.8 Comparing CCR estimates with data generated without a phasing panel. a) Estimates as shown in Figure S08.2. b) CCR Estimates of the same data phased without a reference panel. The separation of curves is smaller, but still evident.

We find that the data phased without a panel (Figure S08.8b) still shows a separation of Australo-Papuan divergence from Eurasian divergence, but the separation between curves is smaller. This suggests that phasing errors can only partially explain the shift in curves.

Uncertainty in CCR estimates, bootstrap analysis and time discretization

To assess the uncertainty of the CCR estimates we generated bootstrap data sets using the block bootstrap method as used in Li, Durbin (2011). We show results for the divergence between Papuans and WCD Aboriginal Australians in Figure S08.9.

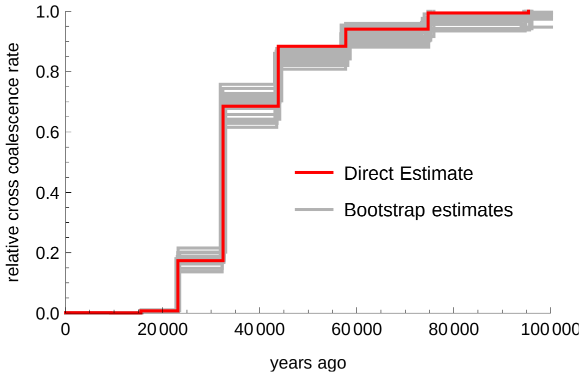


Figure S08.9 Bootstrap estimates for Papuan/Aboriginal Australian divergence. We generated bootstrap samples and estimated CCR curves for the Papuan/WCD divergence (gray). The direct estimate is shown for comparison (red).

The uncertainty of the bootstrap estimates around the estimate based on the whole data is small. We also assessed the impact of time segmentation in MSMC2, as shown in Figure S08.10.

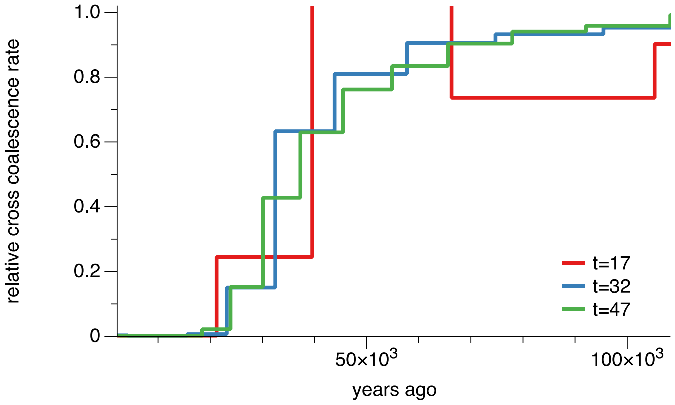


Figure S08.10 Effect of different time patterning. We estimated the Papuan/Aboriginal Australian divergence using varying numbers of time segments. The default (32 patterns) is shown in comparison with estimates using fewer and more segments. The patterns are given as options to MSMC2. The default is “-p 1*2+25*1+1*2+1*3”. Here we also tested “-p 1*2+10*1+1*2+1*3” and “-p 1*2+40*1+1*2+1*3”.

Time segmentation does not affect CCR estimates very much. In particular, the default does not differ much from the case using more time segments. With fewer time segments we see some artifacts such as a non-monotonic CCR curve. We believe the default patterning is appropriate as recommended by MSMC2 developers.

MSMC analysis of simulated data under the model fitted from SFS data

We computed the cross coalescence rates (CCR) curves obtained under our most likely model obtained from the joint site frequency spectra (SFS) analysis, as described in S07 (Figure S07.3). We used the SCRM coalescence simulator (Staab et al. 2015) to simulate African, European and Aboriginal Australian chromosomes with 150Mb each, such that the total amount of sequence equals that of real genomes after masking the regions with poorer mapping quality. The command line for the simulation is:

```
scrm 30 1 -t 84000 -r 75000 1.5e+08 -I 9 0 0 0 0 4 0 4 4
14 0.0 -n 1 0.2556 -n 2 0.6425 -n 3 0.0431 -n 4 0.6425 -n
5 2.49585 -n 6 2.76105 -n 7 0.2537 -n 8 0.61035 -n 9
0.4328 -G 0.0 -eI 0.051975 2 0 0 0 0 0 0 0 0 -eI 0.06595 0
0 2 0 0 0 0 0 0 -m 5 6 8.88 -m 6 5 8.88 -m 6 7 18.28 -m 7
6 18.28 -m 7 8 0.1044 -m 8 7 0.1044 -m 8 9 0.636 -m 9 8
0.636 -eps 0.028175 8 4 0.9962208 -ej 0.03075 8 7 -em
0.03075 7 9 33.84 -em 0.03075 9 7 33.84 -em 0.03075 8 7
0.0 -em 0.03075 7 8 0.0 -em 0.03075 8 9 0.0 -em 0.03075 9
8 0.0 -em 0.03075 7 6 18.28 -em 0.03075 6 7 18.28 -em
0.03075 5 6 8.88 -em 0.03075 6 5 8.88 -en 0.03075 7
4.35125 -eps 0.031525 9 2 0.9606862 -eps 0.03185 9 4
0.99781 -eps 0.035925 7 4 0.9981159 -en 0.043175 7 0.01665
-em 0.043175 5 6 0.0 -em 0.043175 6 5 0.0 -em 0.043175 6 7
0.0 -em 0.043175 7 6 0.0 -em 0.043175 7 9 0.0 -em 0.043175
9 7 0.0 -ej 0.045675 7 6 -en 0.046 9 0.00865 -ej 0.0485 9
6 -eps 0.051025 6 4 0.9652111 -en 0.0542 6 0.02425 -en
0.0567 6 2.76105 -ej 0.12055 6 5 -en 0.12055 5 1.98965 -ej
0.085225 3 4 -ej 0.3029 1 2 -ej 0.380975 2 4 -ej 0.5125 5
4 -en 0.5125 4 1.6121 -p 12
```

This command line simulates 30 haplotypes, which model the genomes of 6 populations: Yoruba (1-4), Sardinian (5-8), Han Chinese (9-12), WCD Aboriginal Australians (13-26), Denisovan (27,28) and Neanderthal (29,30).

We ran three MSMC analyses, each with two haplotypes from Yoruba, and two non-African haplotypes, as shown in Figure S08.11.

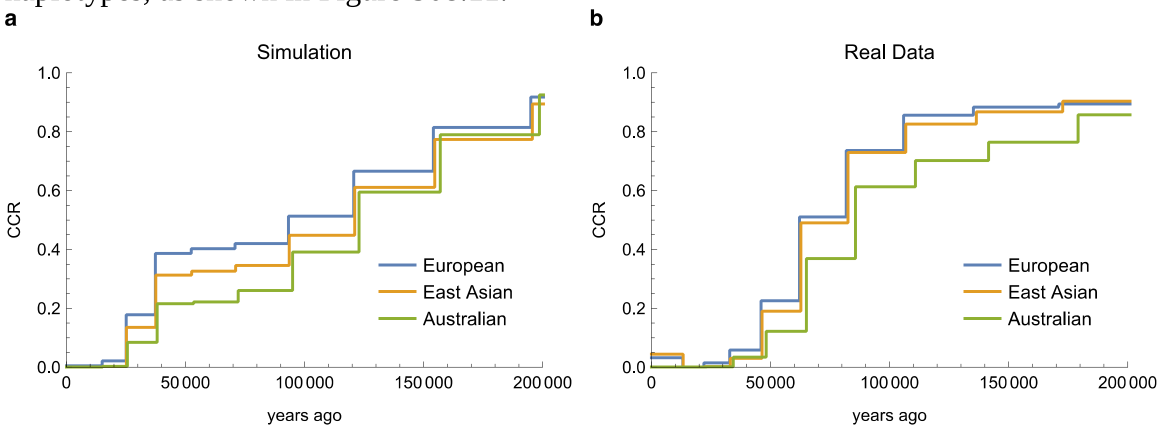


Figure S08.11 MSMC analysis on the African/Non-African divergence on simulated data and real data.

Note that the information used in the data (allele frequencies) to estimate the parameters used for simulations here is different than the information used to compute the CCR curves (haplotypes). We therefore do not expect a perfect fit. In the simulations we observe that up to 130 kya, the CCR curves for Yoruba/Aboriginal Australian is shifted to the right, which is qualitatively similar to the results based on the observed data. However, we note some differences. First, we find that the simulations exhibit a more gradual decline in CCR between Africans and non-Africans than in the real data, where we see a decline with a slower decay phase (before 80kya) and a faster one (between 80kya and present). Second, we also see a difference in the recent European/African CCR (in blue), which is inflated in the simulations compared to the real data in the last 50,000 years. Third, although for the simulations there is a lower CCR from Aboriginal Australians than from Europeans and East Asians the separation of the Aboriginal Australian from the Eurasian CCR curves as seen in the real data is not as clear in the simulated data over the whole time period (until 200’000 years). A potential explanation is that the model contains post-split gene flow between Africa and Europe until present (see e.g. Figure 4), which we think is responsible for the more recent CCR decay from Eurasia than from Australia. The overall similarity between the observed CCR curves and the simulated ones suggest that the SFS based model is sensible. However, the relative lack suggests there is still room for improvement. We could for example consider relaxing some of the simplifying assumptions of the modeled demography, such as constant population sizes along branches and symmetric migration rates, which were introduced to minimize the number of parameters to be estimated from the data. A more sensible future approach will be to combine allele frequencies and haplotype information to estimate demographic parameters.

S08 References

Li, H, R Durbin. 2011. Inference of human population history from individual whole-genome sequences. Nature 475:493-496.

Prufer, K, F Racimo, N Patterson, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43-49.

Schiffels, S, R Durbin. 2014. Inferring human population size and separation history from multiple genome sequences. Nat Genet 46:919-925.

Staab, PR, S Zhu, D Metzler, G Lunter. 2015. scrm: efficiently simulating long sequences using the approximated coalescent with recombination. Bioinformatics 31:1680-1682.

S09 D-statistic based tests using sampled reads from sequencing data

Thomas Mailund, Ida Moltke, Anders Albrechtsen

Computing the D-statistic

Method

To investigate the relationships between different populations we performed D-statistic based tests similar to the one described in Green et al. (2010) and Reich et al. (2010). The statistics were computed based on a single genome from each of the populations of interest. If we let H1, H2 and H3 denote three populations, the D-statistic, D, can be used to test if the data is consistent with the null hypothesis that the tree (((H1, H2), H3), chimpanzee) is correct and there has been no gene flow/population structure between H3 and either H1 or H2. There are several definitions of the D-statistic in the literature. We used the following definition (Durand et al. 2011):

$D = (n_{ABBA} - n_{BABA})/(n_{ABBA} + n_{BABA})$ (eq.S09.1)

where n_{ABBA} is the number of diallelic sites in which the genome from H1 has the same allele as the chimpanzee and the genomes from H2 and H3 have a different allele and n_{BABA} is the number of sites where the genome from H2 has the same allele as the chimpanzee and the genomes from H1 and H3 have a different allele. We computed both standard deviations and significance with a "delete-m Jackknife for unequal m" block-jackknife procedure described in Busing et al. (1999) for 5 Mb blocks of the genome. We considered an absolute Z-value – obtained from the jackknife procedure - higher than 3.0 to indicate a significant deviation from D=0 and thus a rejection of the null hypothesis.

Data

We included a subset of the whole genome sequencing data described in S03. Specifically, we included the following samples:

- Australo-Papuans: Australia: WCD (WCD01, WCD03); HGDP-Papuan: (HGDP00542),
- East Asia: Dai (HGDP01307), Han (HGDP00778),
- Europe: French (HGDP00521), Sardinian (HGDP00665),
- Africa: San (HGDP01029), Yoruba (HGDP00927), Dinka (DNK02),
- Archaic: Altai Neanderthal, Denisovan, Ust Ishim.

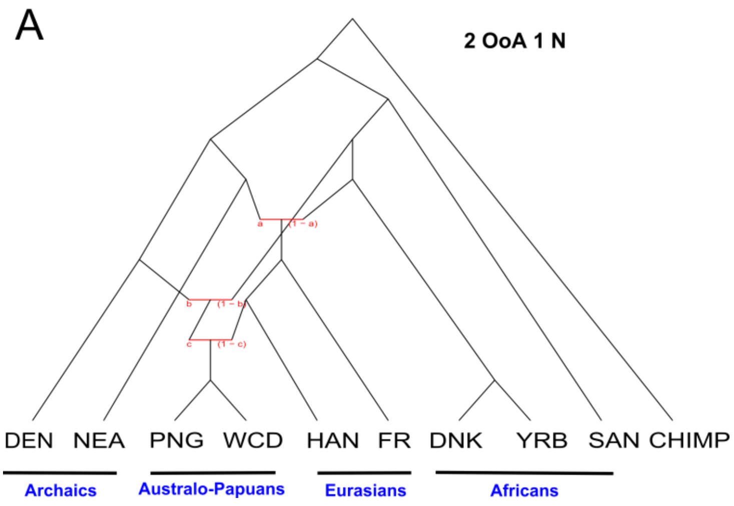
For the chimpanzee outgroup we used the multiway alignment that includes both chimpanzee and human (pantro2 from the hg19 multiz46). Before calculating the D-statistic, the read data was quality filtered as in Orlando et al. (2013). Subsequently, a single base was sampled for each site for each individual. If the sampled bases included a transition, the site was discarded because the archaic genomes that are included in the analysis have postmortem damage (*e.g.*, (Briggs et al. 2007)) resulting in higher errors at these sites. The sampled data from the sites that were not discarded were used to compute the D-statistic.

Admixture graphs: the number of out of Africa events

Background and Method

Admixture graphs provide a framework for fitting a proposed history of populations to the D-statistics computed from the data, which we will here denote "the observed D-statistics". Given a graph of population relationships, including admixture events, the graph predicts expected D-statistics as a function of edge lengths and admixture proportions. Such a graph can be fitted to the observed D-statistics by minimising the distance between the predicted statistics and the observed statistics. We have implemented this approach in an R package available at *https://github.com/mailund/admixture_graph*. The fitting to data is done using a Nelder-Mead numerical optimisation algorithm—implemented in the "neldermead" R package—that minimizes the sum of squared distances between all observed statistics and the expectations from the graph.

We explored a number of admixture graphs involving African, Asian, European, Australian, Papuan and archaic samples. In Figure S09.1, we present only four admixture graphs chosen for further exploration given their higher fit to the data. The admixture graphs differ in two main aspects: 1) the number of out of Africa (OoA) events (one versus two) and 2) the way Neanderthal gene flow is modelled in terms of number of pulses (one or two) and timing of the pulses. For all cases we modelled Asians (Han) as being a sister group to Europeans and included an admixture event between the ancestral population of Australians and Papuans and the ancestral Han population, *i.e.*, two migrations waves into Asia (Rasmussen et al. 2011). Graph 1 ("2 OoA 1 N", Figure S09.1A) involves two OoA and a single gene flow event from Neanderthals. Hence, the Neanderthals contribute to Australo-Papuans "indirectly", by exchanging genes with East Asians. Graph 2 ("2OoA 2N", Figure S09.1B) also involves two OoA but also two independent Neanderthal pulses: one into the ancestral Eurasian population and one into the ancestral Australo-Papuan population. Graph 3 ("1 OoA 1 early N", Figure S09.1C) involves a single out of Africa event and the single Neanderthal pulse takes place before the Eurasian/Australo-Papuan split, into the ancestral population of all non-Africans. Graph4 ("1 OoA 2 late N", Figure S09.1D) involves a single OoA but the gene flow from the Neanderthal goes into the ancestral population of all Eurasians. Note that in all the four graphs the Denisovan contribution was assumed to have occurred only in the ancestral population of Australo-Papuans (Reich et al. 2010; Meyer et al. 2012).



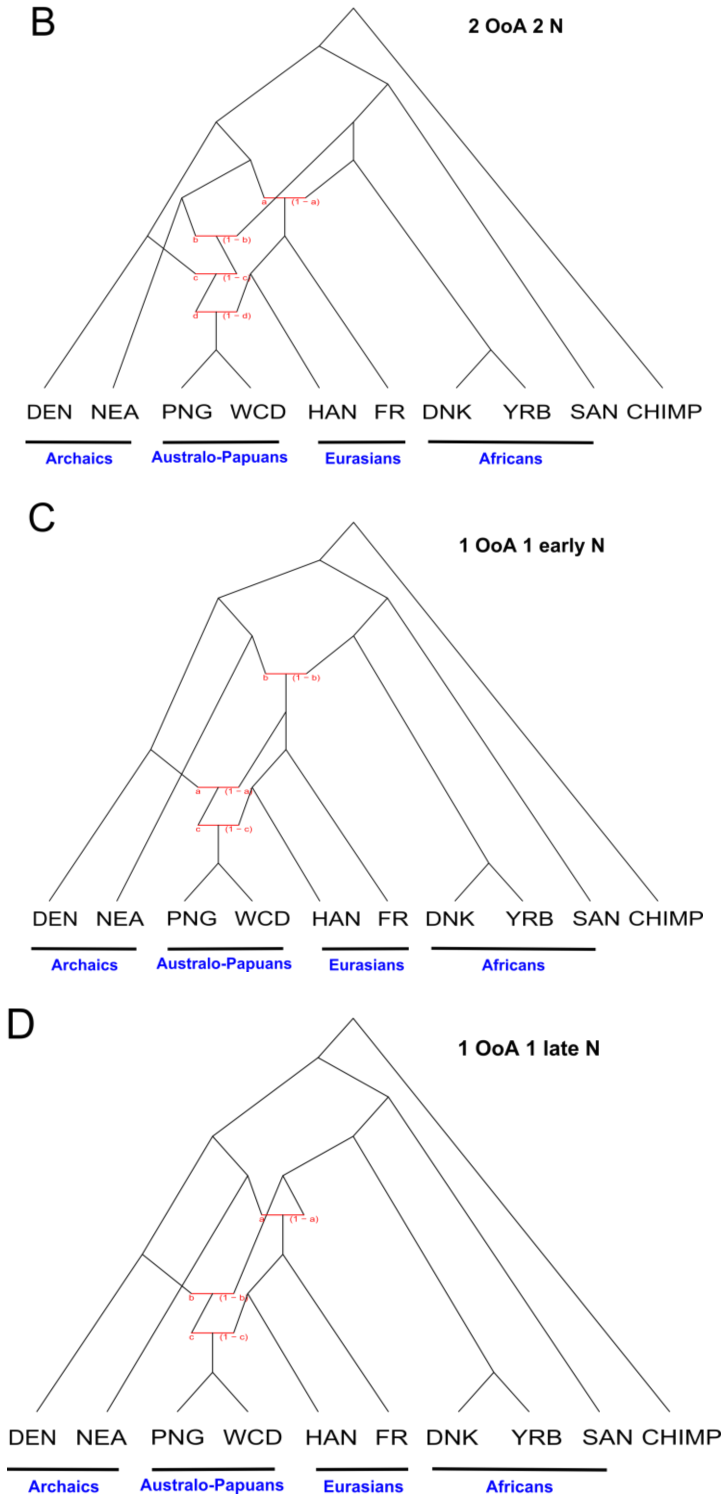


Figure S09.1 Four admixture graphs chosen for further exploration, all involving two archaics (Denisovan, Neanderthal), two Australo-Papuans (an HGDP-Papuan and an Australian-WCD01), two Eurasians (Han and French), three Africans (Dinka, Yoruba, San) and the chimpanzee as an outgroup. All cases involve one Denisovan admixture event into the ancestor of Australo-Papuans and gene flow between the ancestral Han population and the ancestral Australo-Papuan population. **A.** 2 OoA 1 N: two waves out of Africa (OoA) and one Neanderthal admixture event into the Eurasian ancestral population. **B.** 2 OoA 2N: two waves out of Africa (OoA) and two Neanderthal admixture events into the Eurasian ancestral population and the ancestral Oceanians. **C.** 1 OoA 1 early N: one out of Africa event and one Neanderthal gene flow event into the ancestral population of all non-Africans. **D.** 1 OoA 1 late N: one out of Africa event and one late Neanderthal gene flow event into the ancestor of Eurasians**.**

Since the numerical optimisation is sensitive to the choice of the admixture graph parameters, we ran the optimisation ten times with random starting points for each graph and collected the sum of squared errors. In principle, the optimization with the smallest squared errors should be retained but having all of them gives us an idea about how much the fit can vary between optimisations. If two graphs give us roughly the same numbers then we cannot really know if another starting point could change the rank of the graphs, whereas if the fit never overlaps in range we can probably select the graph with the smallest error.

The graph with the smallest errors (best fit) is the **2OoA 2N** but all of the graphs have similar distributions of the sum of squared errors (see Extended Data Figure 4). Note, however, that **2OoA 2N** involves more admixture events (four versus three) and thus more parameters, one would therefore expect a better fit if the topology is correct. We next explored the observed and expected D-statistic for each of the 168 quartets. The results are shown below (Figure S09.2). By plotting the fitted D-statistic for each of the four topologies, one can see that there is no substantial difference between graphs, which explains the difficulty in distinguishing them. Note that the poorly fitted values are mostly the same in these four graphs. Together with the similarity in the error scores in the optimisations, these results suggest that we cannot confidently point towards the graph better fitting the observed data.

Effect of not accounting for the Denisovan admixture in the admixture graphs

We then explored the effect of removing the Denisovan admixture event for the same four graphs as shown in Figure S09.1. Although the ranking among graphs does not change if we remove the Denisovan admixture, the fit gets worse in all cases (see Extended Data Figure 4). Moreover, we now see a much stronger support for two waves out of Africa with two Neanderthal gene flow events, suggesting that accounting for the archaic admixture from Denisovans seems to be key in determining the number of OoA exits. The Denisovan contribution seems to pull the Australians and the Papuans away from the Eurasians, which partly explains why they look like having split from Africans at an earlier time. When accounting for this genetic contribution from Denisovans, however, the support for two waves out of Africa compared to one is much reduced as we saw above.

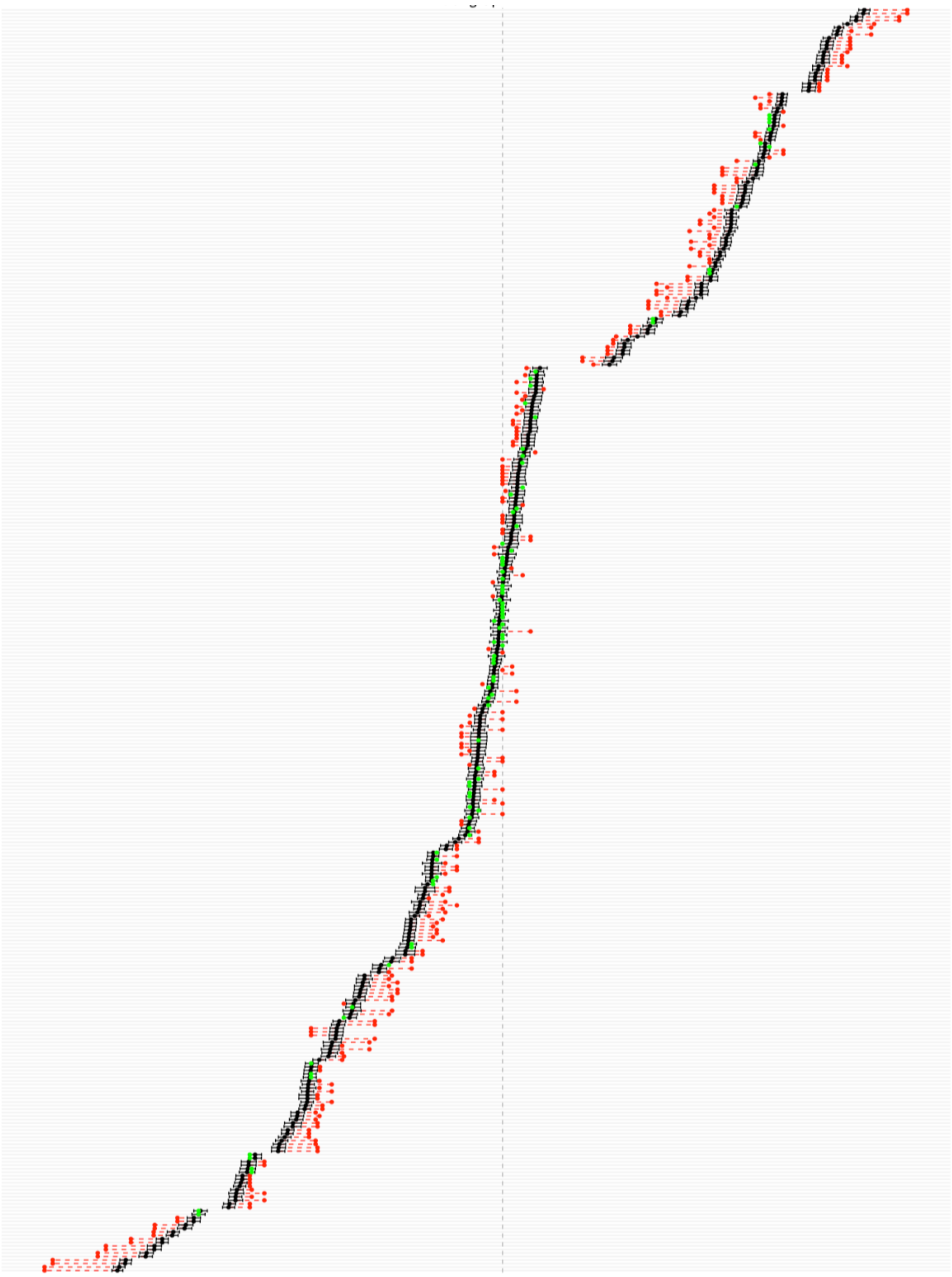


Figure S09.2 An example of the observed D-statistic versus expected D-statistic. The expected D-statistics (with three standard deviations) are shown in black while the observed values are shown in green or in red if the observed D-statistic are within or outside the confidence interval (defined using the standard deviation), respectively. We show the results for the graph **2OoA 1N**. Results are similar for all other graphs. Each row corresponds to the D-statistic computed with a given combination of populations, and all 168 possible combinations of four populations in the graph (quadruplets) are shown.

Correcting the D-statistics for Denisovan admixture

Background

We observed that D-statistics of the form (((H1=Aboriginal Australian,H2=Eurasian), H3=Yoruba),H4=Chimp) are significantly higher than 0. This suggests that Eurasians and Yoruba share more derived alleles than Aboriginal Australians and Yoruba, potentially indicative of two exits out of Africa, with an earlier exit leading to the colonization of Australia.

However, the SFS-based likelihood analysis presented in S07 suggests that the larger divergence observed between Yoruba and Aboriginal Australians than between Yoruba and other non-Africans in part can be explained by the Denisovan ancestry in Australo-Papuans. Furthermore, the above results (Extended Data Figure 4) suggest that accounting for Denisovan admixture impacts our ability to infer the number of exits out of Africa. To assess the effect of such admixture we performed a correction of the D-statistics that tries to account for different levels of Denisovan admixture in the Aboriginal Australians. This approach allowed us to both assess if Denisovan admixture could potentially explain the observed positive D-statistic and to quantify how much Denisovan contribution would be needed in Australian Aboriginals to explain the observed positive D-statistic.

We also used this approach to assess the impact of Denisovan admixture on D-statistics of the form (((H1=Aboriginal Australian,H2=Eurasian), H3=Ust Ishim),H4=Chimp), where Ust-Ishim is the ancient genome from Fu et al. (2014). Several of the uncorrected D-statistics of this form are significantly higher than 0, suggesting that Ust-Ishim are closer to Eurasians than they are to Aboriginal Australians and thus that the Aboriginal Australians are an outgroup to Ust Ishim and Eurasians. We used the approach to assess if the Denisovan admixture could potentially explain the observed positive D-statistics in this case as well.

Method

For each site we let $Y=(Y_{H1},Y_{H2} ,Y_{H3} ,Y_{H4})$ be the pattern of nucleotides for the four individuals H1, H2, H3 and H4. If H1 is the admixed individual of interest, the probability of observing a nucleotide pattern $Y^{admixed}=(Y_{admixedH1},Y_{H2} ,Y_{H3} ,Y_{H4})$ depends on the admixture proportions. We assume that the admixed H1 is the result of a recent admixture event taking place between two source populations called *pop1* and *pop2* (i.e. we ignore drift since the admixture event). In this case, one can expect that a fraction (1-α) of the genome from this population is from *pop1* while the rest, α, is from a *pop2*. The probability of observing the specific nucleotide pattern *y* is then

$$P(Y^{admixed}=y)=(1-\alpha)P(Y^{pop1}=y) + \alpha P(Y^{pop2}=y), \quad \text{(eq.S09.2)}$$

where $Y^{pop1}=(Y_{pop1},Y_{H2} ,Y_{H3} ,Y_{H4})$ and $Y^{pop2}=(Y_{pop2},Y_{H2} ,Y_{H3} ,Y_{H4})$ are the nucleotide patterns with the two source populations as H1 instead of the admixed H1. Based on the nucleotide patterns observed with the two source populations as H1 for any given value of α, we can use this to obtain expectations of all the possible nucleotide patterns after admixture. Likewise we can obtain estimates of the probability of a given nucleotide pattern for a population before the admixture event, e.g., *pop1*, based on the observed patterns with the admixed population and the other source population, *pop2* as H1 by rewriting eq.S09.2:

$$P(Y^{pop1}=y)= (P(Y^{admixed}=y)- \alpha P(Y^{pop2}=y)) / (1-\alpha) \quad \text{(eq. S09.3)}$$

We used this latter equation to calculate a D-statistic corrected for Denisova admixture in Aboriginal Australian samples by setting the admixed population to Aboriginal Australians, and *pop2* to Denisovans. *Pop1* then corresponds to Aboriginal Australian that have not received any gene flow from Denisovans. We first calculated $P(Y^{admixed}=y)$ for the case (((H1=Aboriginal Australian,H2=Eurasian), H3=Yoruba/Ust Ishim),H4=Chimp) and $P(Y^{pop2}=y)$ for the case (((H1=Denisovan,H2=Eurasian), H3=Yoruba/Ust Ishim),H4=Chimp) by counting the number of times the specific pattern *y* is observed and dividing that number by the total number of patterns observed. Next, we calculated $P(Y^{pop1}=y)$ using equation *eq. S09.3* with a given value of α. From these we then calculated the admixture corrected probability of observing an ABBA (BABA) pattern by summing over $P(Y^{pop1}=y)$ for all nucleotide patterns that have an ABBA (BABA) pattern. Finally, we computed admixture corrected D-statistic by plugging the admixture corrected probabilities of observing ABBA and BABA into the formula for the D-statistic (i.e. eq. S09.1).

As noted above, we have assumed that the admixture event took place today, *i.e.*, we ignored drift since the admixture event, which is unrealistic and can bias the results. However, if the observed ABBA or BABA patterns are mainly the result of the admixture event and if drift was negligible since then (e.g. due to large effective sizes), the estimates of corrected D-statistic should still be informative.

Data

We used whole genome sequencing data to obtain counts of the different patterns as described above. We used the ancient Australian Aboriginal sample AusAboriginal (Rasmussen et al. 2011) and the ancient sample Ust-Ishim (Fu et al. 2014). Additionally, we used the following modern samples: an Australian Aboriginal (WCD01), a Yoruba individual (HGDP00927) representing Africa and four Eurasian individuals: a Dai (HGDP01307), a Han (HGDP00778), a Sardinian (HGDP00665) and a French (HGDP00521).

Results

The results for the D-statistics of the form (((H1=Aboriginal Australian,H2=Eurasian), H3=Yoruba),H4=Chimp) are summarized in Figure S09.3 and the results for the D-statistics of the form (((H1=Aboriginal Australian,H2=Eurasian), H3=Ust Ishim),H4=Chimp) are summarized in Figure S09.4.

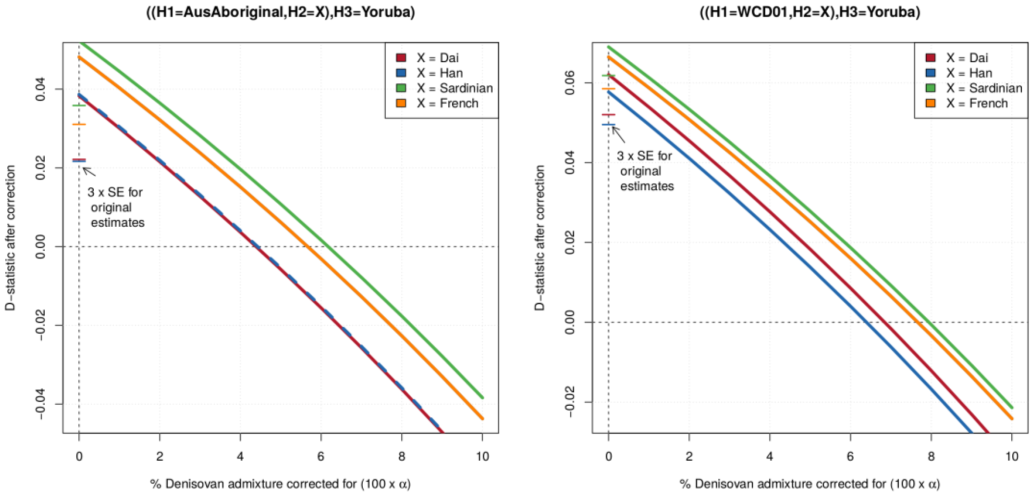


Figure S09.3 D-statistic of the form (((H1=Aboriginal Australian, H2=Eurasian), H3=Yoruba), H4=Chimp) after correction for Denisovan admixture. The x-axis is the admixture proportion (in %) denoted as 100 x α in eq. S09.3. The values at x=0 are the uncorrected D-statistics, which also have 3 times the standard error shown as horizontal lines. The left figure explores the effect on the ancient Australian Aboriginal AusAboriginal (without transitions) and the right figure explores the effect on a modern unadmixed Australian Aboriginal (with transitions) both as H1. Two European and two East Asian populations are used as H2. A Yoruba individual from Africa (Nigeria, HGDP00927) is used as H3.

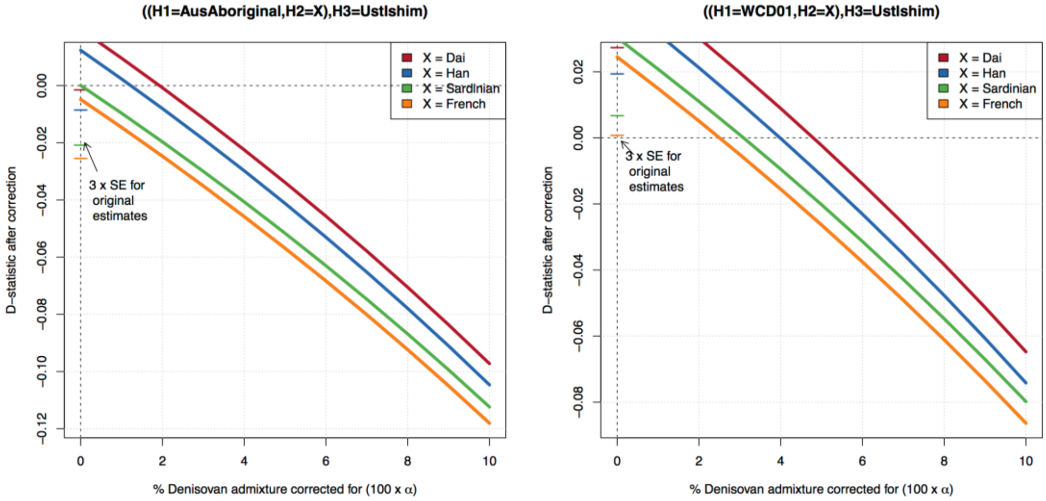


Figure S09.4 D-statistics of the form (((H1=Aboriginal Australian,H2=Eurasian), H3=Ust Ishim),H4=Chimp) after correction for Denisovan admixture. The x-axis is the admixture proportion (in %) denoted as 100 x α in eq. S09.3. The values at x=0 are the uncorrected D-statistics, which also have 3 times the standard error shown as horizontal lines. The left figure explores the effect on the ancient Australian Aboriginal AusAboriginal (without transitions) and the right figure explores the effect on a modern unadmixed Australian Aboriginal (also without transitions) both as H1. Two European and two East Asian populations are used as H2. The ancient sample Ust Ishim was used as H3.

Discussion

As can be seen in Figure S09.3, the rejection of the trees with an African sample as H3 with a positive D-statistics can be explained by a moderate amount of Denisovan admixture: for instance for the tree (((H1=AusAboriginal,H2=Sardinian), H3=Yoruba),H4=Chimp) assuming 4% Denisovan (corresponding to the estimate reported in S07) admixture leads to the corrected D-statistic being markedly smaller, but still significantly different from 0 (Z=3.50 vs. 9.81 for the uncorrected one).

The rejection of the trees with Ust Ishim as H3 with a positive D-statistic can also be explained by a moderate amount of Denisovan admixture (Fig. S09.4). Here, for the tree with the most extreme D-statistic (((H1=WCD01, H2=Dai), H3=Ust Ishim), H4=Chimp) assuming 4% Denisovan admixture leads to the corrected D-statistic being non-significantly different from 0 (Z=1.07 vs 6.41 for the uncorrected D-statistic). We note that the above assumes a simple genetic history of Eurasian populations with, for example, no gene flow from Africa into Eurasia after the out of Africa event. Such events, including the scenarios inferred from the above admixture graphs, will make the D-statistics more positive.

S09 References

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prufer K, Meyer M, Krause J, Ronan MT, Lachmann M, et al. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. Proc. Natl. Acad. Sci. 104:14616–14621.

Busing FMTA, Meijer E, Leeden RVD. 1999. Delete-m Jackknife for Unequal m. Stat. Comput. 9:3–8.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, Johnson PLF, Aximu-Petri A, Prüfer K, de Filippo C, et al. 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature 514:445–449.

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

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, Filippo C de, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338:222–226.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, et al. 2013. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature 499:74–78.

Rasmussen M, Guo X, Wang Y, Lohmueller KE, Rasmussen S, Albrechtsen A, Skotte L, Lindgreen S, Metspalu M, Jombart T, et al. 2011. An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 334:94–98.

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PLF, et al. 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060.

S10 Archaic gene flow

Jeffrey D Wall

Identifying introgressed Neanderthal and Denisovan haplotypes

Method

To identify haplotypes found in contemporary human genomes that were likely to have been inherited due to recent interbreeding with archaic humans, we adopt an approach that combines aspects of enhanced D-statistics (cf. (Meyer et al. 2012)) and linkage-disequilibrium based approaches (e.g., (Wall et al. 2013)). We searched across non-African genomes for clusters of sites in complete linkage disequilibrium ($r^2 = 1$ for all pairwise comparisons) and used these to define potential 'haplotypes'. In practice we actually use diploid genotypes (to account for incomplete phase information), identifying groups of sites with complete correlation between genotypes at each site. We further require individuals to contain all heterozygous or all homozygous genotypes within the identified sites. Equivalently, we identify groups of sites such that there exists a phasing whereby there are only two observed haplotypes. We then filtered those 'diplotypes' to enrich for those that might be the result of Neanderthal (respectively, Denisovan) admixture using the following criteria:

1. the genotypes of 10 sub-Saharan Africans are all homozygous ancestral
2. the genotype of the Altai Neanderthal (respectively, Denisovan) is homozygous derived
3. the genotype of the Denisovan (respectively, Altai Neandertal) is homozygous ancestral
4. the derived allele frequency across all (excluding the Africans) contemporary human samples is > 0 and < 0.5

The final criterion was included to reduce the false positive rate when the ancestral allele is misspecified. The 10 sub-Saharan African samples include 2 San, 2 Mbuti, 2 Yoruba, 2 Mandenka and 2 Dinka (see S04). We implicitly assume that the ancestors of these samples did not interbreed with Neanderthals or Denisovans within the past 200 thousand years.

Then, from this filtered data set, we scanned for extended genotypes containing 4 or more completely correlated SNPs spanning at least 4 Kb (as described above). We further required that there are no gaps of more than 10 Kb in between consecutive SNPs, and call these putative Neanderthal haplotypes (PNHs), respectively, putative Denisovan haplotypes (PDHs). Note that the SNPs in the filtered list taken individually can be used to calculate the equivalent of an enhanced D-statistic. Our rationale for requiring multiple sites in LD rather than individual sites is that this should remove most of the filtered sites that are due to segregation of ancestral haplotypes. Since ancestral haplotypes are generally small, they are unlikely to contain very many diagnostic sites. Instead, recent (e.g., < 100 Kya) admixture will tend to produce longer introgressed haplotypes which we target with our protocol.

For each individual, we tabulated the numbers of PDHs and PNHs, as well as the total length spanned by these (PDHlengths and PNHlengths). Our simulations suggest that each PDH or PNH has a high probability of being true, with a FDR of q ~ 0.05 in Aboriginal Australians without recent European or Asian admixture (results not shown). Note that there are many true Neanderthal and Denisovan tracts that were missed by our method, so the PDHlengths and PNHlengths values are much smaller than the true amount of Denisovan and Neanderthal admixture.

Results

Overall we identified 12,160 PNHs and 7,889 PDHs across all individuals. We also found substantial differences in the numbers and lengths of putative archaic haplotypes across different populations (Table S10.1), consistent with differential levels of Neanderthal and Denisovan admixture found in previous studies (Skoglund and Jakobsson 2011; Meyer et al. 2012; Wall et al. 2013; Prufer et al. 2014). There was also substantial variation in the numbers of PDHs across Aboriginal Australian samples. This latter observation can be explained by differential levels of admixture with Eurasian source populations – across individuals, the number of PDHs is strongly correlated with the estimated amount of Australo-Papuan ancestry (Figure S10.1A, $r^2 = 0.96$, $p = 5.43 * 10^{-49}$).

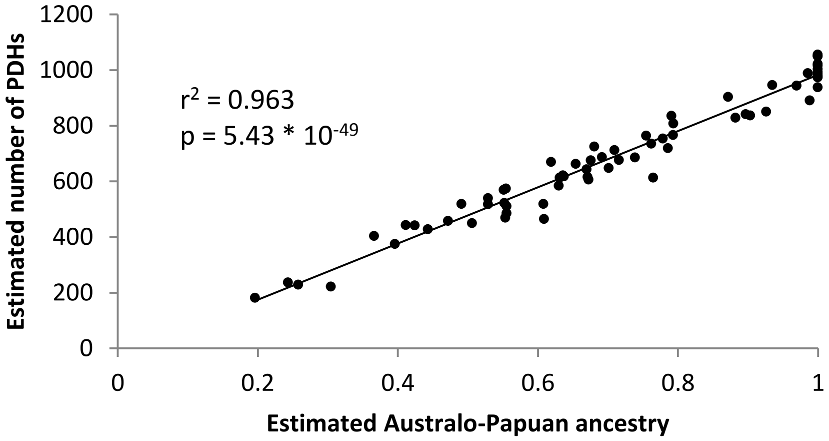
Across Aboriginal Australian samples, there is a positive correlation between the average amount of Neanderthal ancestry and the average amount of Denisovan ancestry, as measured by the number of PNHs and PDHs (Figure S10.1.B, $r^2 = 0.33$, $p = 1.41 * 10^{-8}$). Additionally, across Aboriginal Australians there is a weak negative correlation between the number of PNHs and the average PNH length (Figure S10.1.C, $r^2 = 0.16$, $p = 1.41 * 10^{-4}$). Both of these observations are consistent with a model where the Neanderthal admixture into the ancestors of Aboriginal Australians occurred earlier (in number of generations) than the Neanderthal admixture into the ancestors of Eurasians (as might happen if Aboriginal Australians had a shorter generation time on average).

Finally, we looked at the sharing of PNHs (and PDHs) across samples from different population groups to see if there was evidence for separate Neanderthal (or Denisovan) admixture events in the ancestors of different non-African populations. We tabulated F_{ST} (cf. (Hudson et al. 1992)) for PNHs using F_{ST} for all autosomal SNPs as a control. We find little difference in the two sets of F_{ST} values. For example, the WCD – East Asian and WCD – European F_{ST} values are 0.171 and 0.187 for SNPs and 0.165 and 0.180 for PNHs. This suggests that the primary differentiation between non-African populations likely postdated the primary admixture event with Neanderthals. Similarly, we estimate an F_{ST} of ~0.12 between WCD and Papuans, for autosomal SNPs, PNHs and PDHs. There is also no significant difference between the number of PDHs or the distribution of PDHlengths between unadmixed Australians and Papuans (Mann-Whitney U test, p>0.05). Together, these observations provide strong evidence for a single Denisovan admixture event that predates the population split between Australians and Papuans.

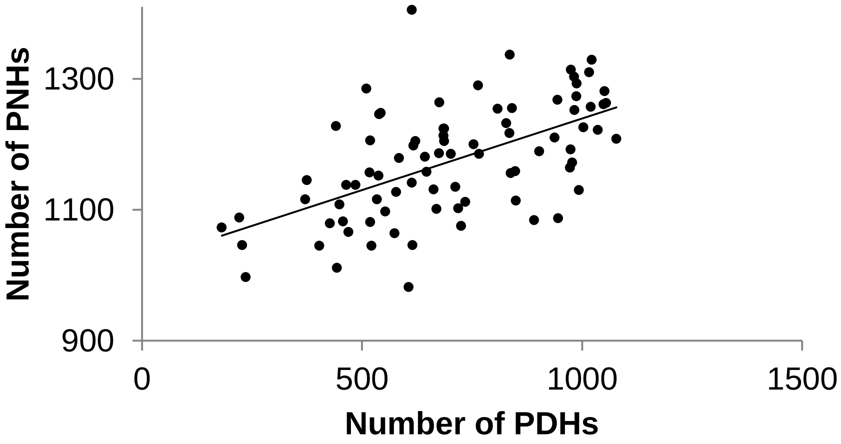
Table S10.1 Average number of PDHs, PNHs, PDHlengths and PNHlengths per individual stratified across populations.

Population	Sample size	PDHs	PDHlengths (Mb)	PNHs	PNHlengths (Mb)
European	2	77	1.57	951	38.5
South Asian	6	124	4.47	966	40.2
East Asian	3	128	4.47	1120	44.3
Amerindian	3	115	3.01	1140	45.6
Polynesian	2	241	8.09	1105	43.6
Papuan	16	974	35.8	1183	41.5
BDV	9	602	22.3	1171	42.2
CAI	10	653	23.6	1188	43.6
ENY	8	533	19.7	1149	42.8
NGA	6	682	25.5	1167	42.9
PIL	12	785	28.5	1225	42.6
RIV	8	508	18.2	1093	41.1
WCD	13	991	38.2	1225	39.9
WON	11	717	26.5	1151	40.8
WPA	6	647	24.2	1148	42.1

A



B



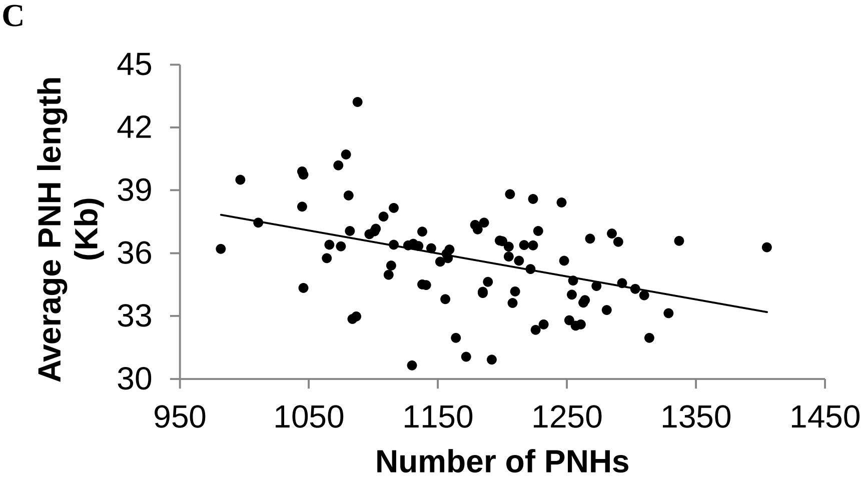


Figure S10.1 Correlations between (A) the estimated amount of Australo-Papuan ancestry (see S05) and the number of identified PDHs, (B) the number of PDHs and the number of PNHs, and (C) the number of PNHs and the average length of each PNH, for all Aboriginal Australian samples.

Detecting admixture from an unknown archaic human source population

Method

We explored the possibility of the Australo-Papuan genomes having experienced admixture from an unknown archaic human source population (sometimes called ‘ghost’ admixture). If such admixture happened, simulations suggest that it would produce short (e.g., 10’s of Kb) introgressed regions that fall outside of the range of modern human variation and do not match either the Neanderthal or Denisovan genome. We searched across 11 HGDP-Papuan and 20 Aboriginal Australian genomes for clusters of sites in complete linkage disequilibrium ($r^2 = 1$ for all pairwise comparisons) and used these to define potential ‘haplotypes’. We included only those individuals that looked to have essentially complete Australo-Papuan ancestry based on chromosome painting and the estimated number of PDHs. These included the following samples: CAI01, ENY02, NGA02, PIL01, PIL06, PIL07, WCD01, WCD02, WCD03, WCD04, WCD05, WCD06, WCD07, WCD08, WCD10, WCD11, WCD12, WCD13, WON03, WON06, 13733_8, 13748_1, 13748_2, 13748_3, 13748_6, 13748_7, 13748_8, 13784_2, 13784_3 and 13784_5 (see S03 and S05). In practice, as above, we actually use diploid genotypes (to account for incomplete phase information), identifying groups of sites consistent with $r^2 = 1$ under a specific phasing. (Equivalently, we are identifying sites where there is a complete correlation of the genotypes across individuals). We then filtered these ‘haplotypes’ to enrich for those that might be due to ancient admixture using the following criteria:

1. Haplotype spanned at least 20 Kb.
2. The density of diagnostic SNPs was at least 1 / Kb.
3. Haplotypes were found in at least one non-African sample.
4. Haplotypes were absent from the panel of 10 sub-Saharan African samples described above.
5. Derived allele frequency for each diagnostic SNP was < 0.5 across all samples.
6. Average frequency of diagnostic alleles was < 0.1 for both the Neanderthal and Denisovan genome sequences.

In all, we identified 167 such regions, which we call ‘putative introgressed haplotypes’ (PIH). Without step 6 we identified 847 regions, the vast majority of which are presumably due to recent Neanderthal or Denisovan admixture.

To assess how many PIH’s are expected by chance, we ran null simulations which incorporated Neanderthal and Denisovan admixture, but no admixture from any other archaic human group (see Figure S10.2). We chose this rather simple model rather than the one estimated from the data using the SFS (S07) because we wanted to isolate the effect of putative archaic admixture on patterns of genetic variation rather than adding the potentially confounding effects of a complicated demographic model. Simulations were run using a modification of a standard coalescent simulator (Hudson 2002) which allowed for the inclusion of archaic human samples. We simulated whole-genome sequence data from 1 Neanderthal, 1 Denisovan, 10 Africans, 11 Papuans and 20 Aboriginal Australians. We took demographic parameters from the literature, such as a mutation rate of 1.25 * 10^{-8} / bp (Scally and Durbin 2012), heterozygosity of 0.001 / bp in the sub-Saharan African samples, and an average generation time of 25 years. We assumed that Neanderthals and Denisovans split from each other 425 Kya, and that their common ancestor split from the ancestors of modern humans 650 Kya (Prufer et al. 2014). The Denisovan and Neanderthal samples were assumed to date from 100 Kya and 130 Kya respectively (Prufer et al. 2014), and the admixture between modern humans and Neanderthals (Denisovans) occurred 60 Kya (55 Kya). One set of simulations assumed that there was no population structure within Denisovans, while the other assumed that the ancestors of the Altai Denisovan sequence and the ancestors of the Denisovans who admixed with Australo-Papuans were completely isolated since 200 Kya. African and non-African populations were assumed to have split 80 Kya without subsequent migration, and Aboriginal Australian and Papuan samples were assumed to have split 20 Kya. We also incorporated a population bottleneck in the non-African samples, from 55 – 60 Kya, with parameters chosen to produce a 40% reduction in genetic diversity (Wall et al. 2008). We assumed the Neanderthal and Denisovan effective population sizes to be 20% of that of modern humans (i.e., N_e = 4,000). Finally, we assumed a 2.5% Neanderthal contribution and a 4% Denisovan contribution to the Australo-Papuan genomes.

Results

For computational tractability, we simulated 540 different 5 Mb blocks rather than whole chromosomes. The number of blocks was chosen to match the total length of the accessible genome obtained from Illumina sequence data. Since the recombination rate is a fundamental determinant of patterns of linkage disequilibrium (and the expected number of PIH’s), and there are no Aboriginal Australian-specific estimates of recombination rates in the literature, we simulated 6 whole-genome data sets, for each of the two scenarios described above, assuming a range of possible scaled recombination rates, corresponding to ρ (= 4Nr) of 0.2, 0.4, 0.6, 0.8, 1 and 1.2 per Kb. We analysed each simulated data set in the same fashion as the real data, and tabulated the genetic distance (in cM) of the largest simulated and actual PIH’s (genetic distances are mostly affected by the time and extent of admixture and are relatively insensitive to the local recombination rate). Local recombination rates for the actual PIH’s were obtained from the UCSC Genome Browser (Recomb Rate track) and are based on the Decode recombination maps.

The actual data had 12 PIH’s $\geq$ 0.1 cM in length and 49 PIH’s $\geq$ 0.05 cM in length, and the simulated data (over all 12 whole-genome simulations) had no PIH’s $\geq$ 0.03 cM in length.

While there is substantial uncertainty regarding the demographic model used and its associated parameters, we still believe this provides some evidence that there was additional (i.e., non-Neanderthal and non-Denisovan) admixture with a diverged hominin group into Australo-Papuans. Our second set of simulations also show that population structure within Denisovans (between the Altai Denisovan and the Denisovans that admixed with ancestral Australo-Papuans) cannot explain the excess of observed PIHs.

While we do not know the potential source population of these PIH's, one possibility is that Asian *Homo erectus* (or its descendants) still occupied Southeast Asia during the time when modern humans first expanded out of Africa. If *H. erectus* were the source population, then we would expect a substantial degree of divergence between them and modern humans, and potentially very little admixture would be necessary to account for the observed numbers of large PIH's. To obtain a ballpark estimate of the admixture proportion, we modified the model in Figure S10.2 to include additional admixture (at 54 Kya) from a source population that had been completely isolated from the ancestors of modern humans, Neanderthals and Denisovans since 1,500 Kya. We found that just a 0.1% contribution of this population to the genomes of Australo-Papuans led to an average of >100 PIH's longer than 0.05 cM in whole-genome simulations over a range of different recombination rates. So, the total contribution of this 'ghost' population to modern human genomes was probably very limited.

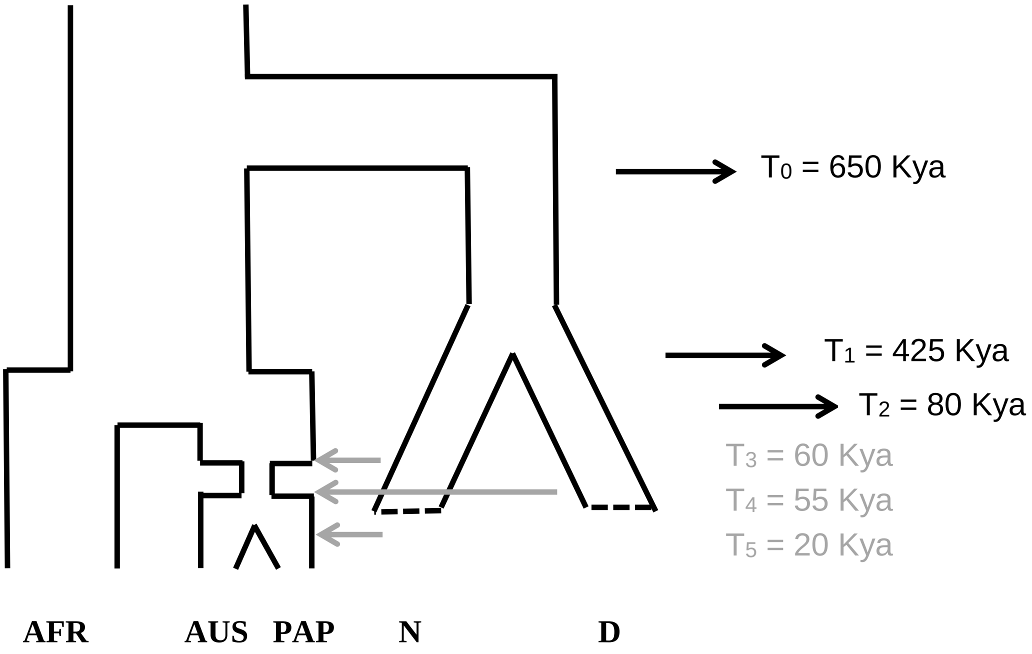


Figure S10.2 Schematic of the demographic model used for the simulations. T_0 is the split time between the ancestors of modern humans and the ancestors of Neanderthals and Denisovans. T_1 is the split time between Neanderthals and Denisovans. T_2 is the split time between African and Australo-Papuan modern human populations. T_3 is the admixture time between Neanderthals and Australo-Papuans. T_4 is the admixture time between Denisovans and Australo-Papuans and T_5 is the admixture time between Aboriginal Australians and Papuans. AFR = African modern humans, AUS = Aboriginal Australian modern humans, PAP = Papuan modern humans, N = Neanderthals, D = Denisovans.

S10 References

Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, Filippo C de, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science 338:222–226.

Prufer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505:43–49.

Scally A, Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. Nat. Rev. Genet. 13:745–753.

Skoglund P, Jakobsson M. 2011. Archaic human ancestry in East Asia. Proc. Natl. Acad. Sci. 108:18301–18306.

Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF. 2008. A novel DNA sequence database for analyzing human demographic history. Genome Res. 18:1354–1361.

Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, Gignoux C, Woerner A, Hammer MF, Slatkin M. 2013. Higher Levels of Neanderthal Ancestry in East Asians than in Europeans. Genetics 194:199–209.

S11 Mutation load analysis

Stephan Peischl, Isabelle Dupanloup, Vitor C Sousa, Laurent Excoffier

Data preparation and processing

In addition to the genomes sequenced in the present study, we processed the genomes of two archaic humans (Denisova, Altai Neanderthal) and 22 Modern humans previously published (Meyer, et al. 2012; Prufer, et al. 2014). The ancestral state of each variant in these genomes were inferred using the ancestral hg19 genome (*http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/*). Only variants for which the ancestral state could be characterized were kept for downstream analysis.

We used the **G**enomic **E**volutionary **R**ate **P**rofiling (**GERP**) scores to quantify the level of evolutionary constraint acting on the SNPs (Davydov, et al. 2010) and ANNOVAR to functionally characterize the SNPs (Wang, et al. 2010). GERP scores computed from the alignment of 35 mammals to the human genome reference sequence hg19 were downloaded from the UCSC platform (Rosenbloom, et al. 2015). GERP scores can be defined as the number of substitutions expected under neutrality minus the number of substitutions observed at that position (Cooper, et al. 2005), and hence are also known as Rejected Substitution (RS) scores. Positive GERP (or RS) scores, larger than 2, represent a substitution deficit, which are expected for sites under selective constraint; while smaller scores, including negative values, indicate that a site is probably evolving neutrally (Davydov, et al. 2010).

Assessment of mutation deleteriousness

GERP RS conservation score as a proxy for mutation effect

We classified all mutations discovered in the dataset into categories based on GERP or Rejected Substitution (RS) scores to categorize mutations by their predicted deleterious effect (Cooper, et al. 2005). Variants were sorted into four groups reflecting the likely severity of mutational effects: "neutral" (-2 < GERP RS < 2), "slightly deleterious" (2 ≤ GERP RS ≤ 4), "deleterious" (4 <GERP RS < 6) and "strongly deleterious" (GERP RS ≥ 6).

Average RS scores

It is an inherently difficult problem to assess mutation load from genomic data (see e.g., Lohmueller 2014 for a discussion of this problem). Instead, we use RS scores as a proxy for selection intensity and calculate, for each individual, the average RS score across all sites at which the focal individual carries a deleterious allele. We focus here on two different measures for the average RS score per site. First, *the average RS score per site* is simply the average of RS scores calculated over all sites at which an individual carries at least one copy of a derived mutation: $\frac{1}{n}\sum RS_i$, where n is the number of segregating sites per individual, and RS_i is the RS score of site i. Note that this measure does not distinguish between heterozygous sites and derived homozygous sites. In contrast, *the average additive RS score per site* is defined as $\frac{1}{n}\sum RS_i\ g_i$, where g_i is the number of derived alleles at site i. For simplicity, we will sometimes call the average additive RS score the "average load per site".

Results

Unless stated otherwise, all results in the following section are based on SNPs located in exonic regions.

Overall distribution of conservation scores

As expected, we find that the number of derived alleles per individual decreases with increasing GERP RS score (Figure S11.1). When focusing on the average additive GERP RS score, we find that the average conservation of sites increases with distance from Africa (Figure S11.2). This is mainly due to an increase in the number of derived homozygous sites with distance from Africa (Figure S11.3). Intriguingly, the number of derived homozygous sites increases across all GERP categories, except for strongly deleterious sites (GERP RS score > 6). This suggests that, during the expansion, slightly and moderately deleterious sites evolved as if they were neutral due to strong genetic drift at the front of the expansion (see also Henn, et al. 2015), whereas the evolution of strongly deleterious sites is mainly driven by selection. We also see a decrease in the number of segregating sites per individual with increasing distance from Africa, in line with strong drift at the expansion front (Figure S11.4).

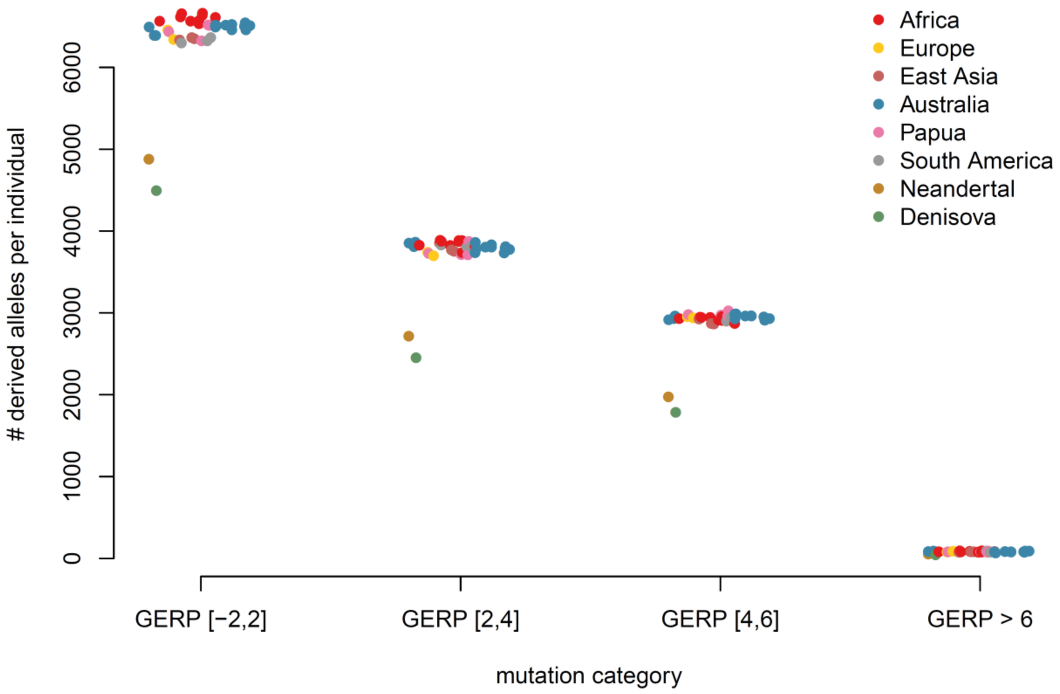


Figure S11.1 Number of derived alleles per individual for different categories of GERP RS scores. Each dot represents an individual and different colours correspond to different populations.

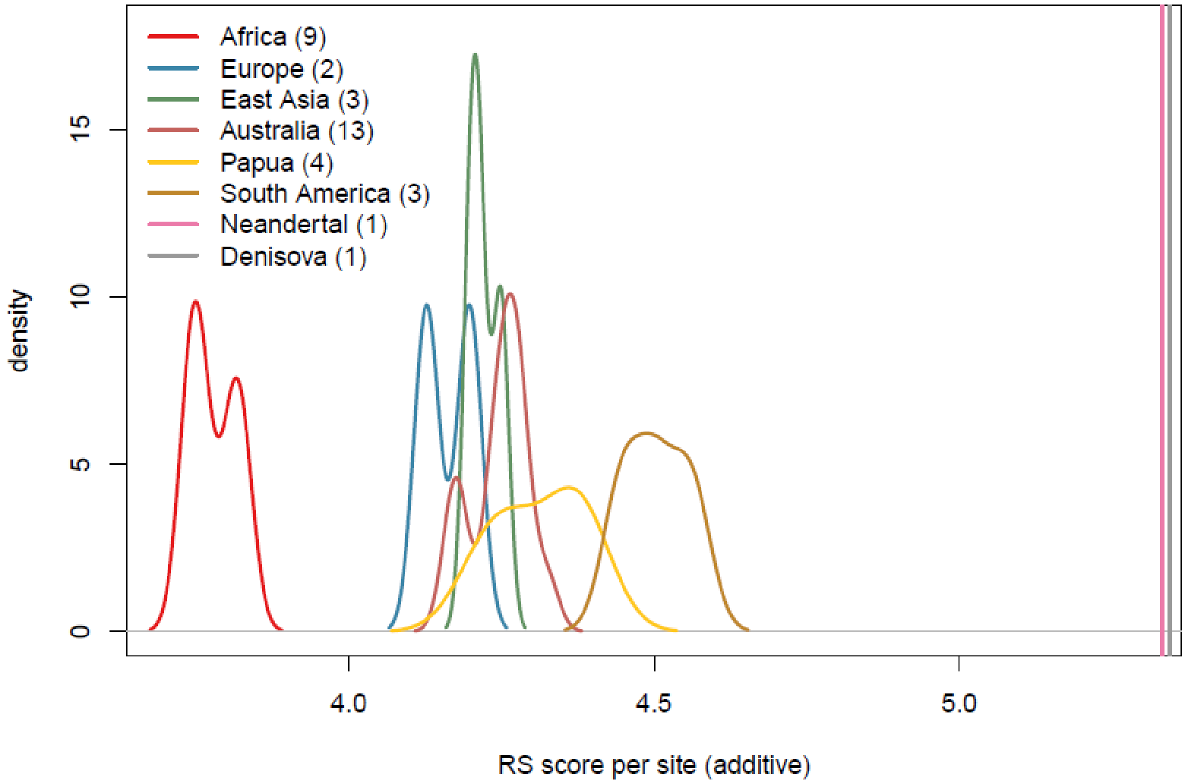


Figure S11.2 Average additive RS score per site. Each curve shows the distribution of average RS scores across individuals belonging to a population. Vertical lines indicate the average RS score for populations with only one individual (Neanderthal and Denisova).

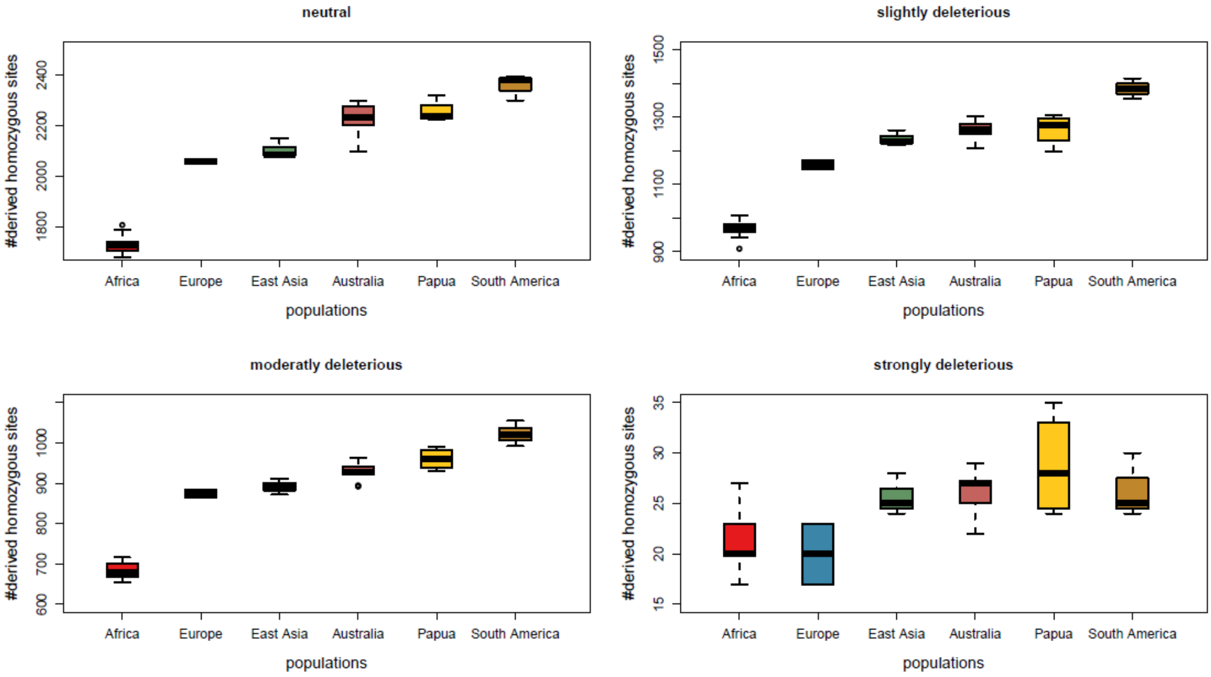


Figure S10.3 Number of homozygous derived sites per individual. Each boxplot shows the distribution of the number of homozygous derived sites across individuals from a population. Each panels shows results for a class of mutations (“neutral” (-2 < GERP RS < 2), “slightly deleterious” (2 ≤ GERP RS ≤ 4), “deleterious” (4 <GERP RS < 6) and “strongly deleterious” (GERP RS ≥ 6)).

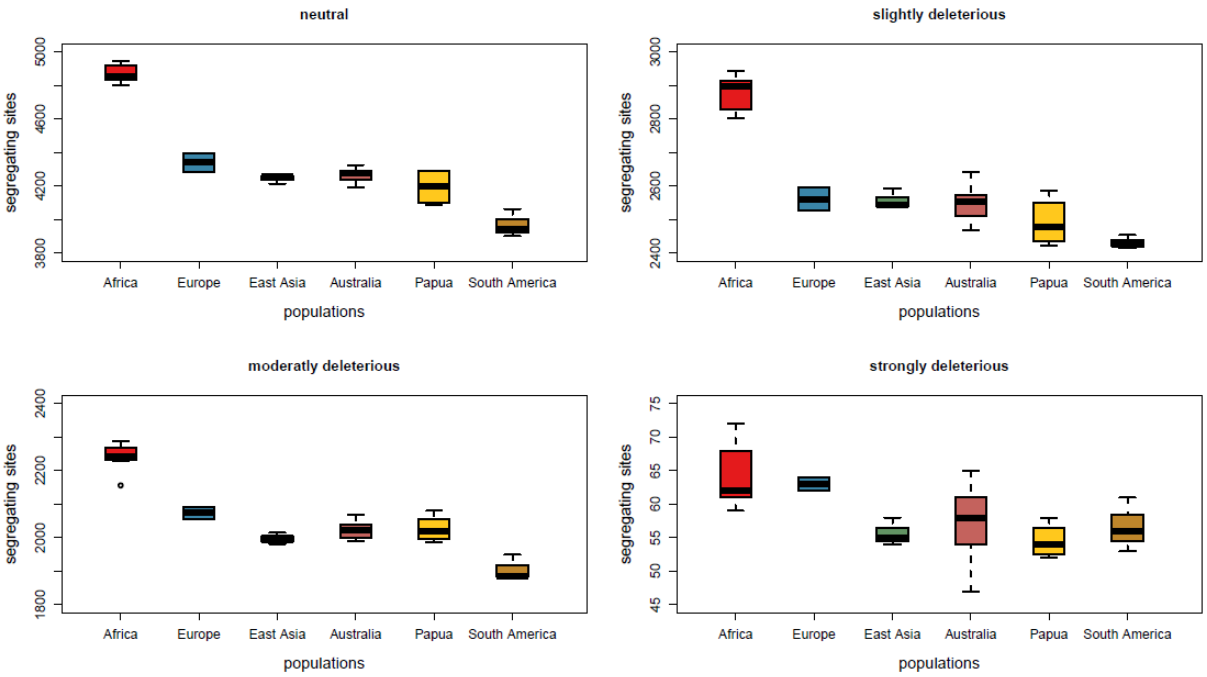


Figure S11.4 Number of segregating sites per individual. Each boxplot shows the distribution of the number of sites at which an individual carries at least one derived allele. Each panels shows results for a class of mutations ("neutral" (-2 < GERP RS < 2), "slightly deleterious" (2 ≤ GERP RS ≤ 4), "deleterious" (4 < GERP RS < 6) and "strongly deleterious" (GERP RS ≥ 6).).

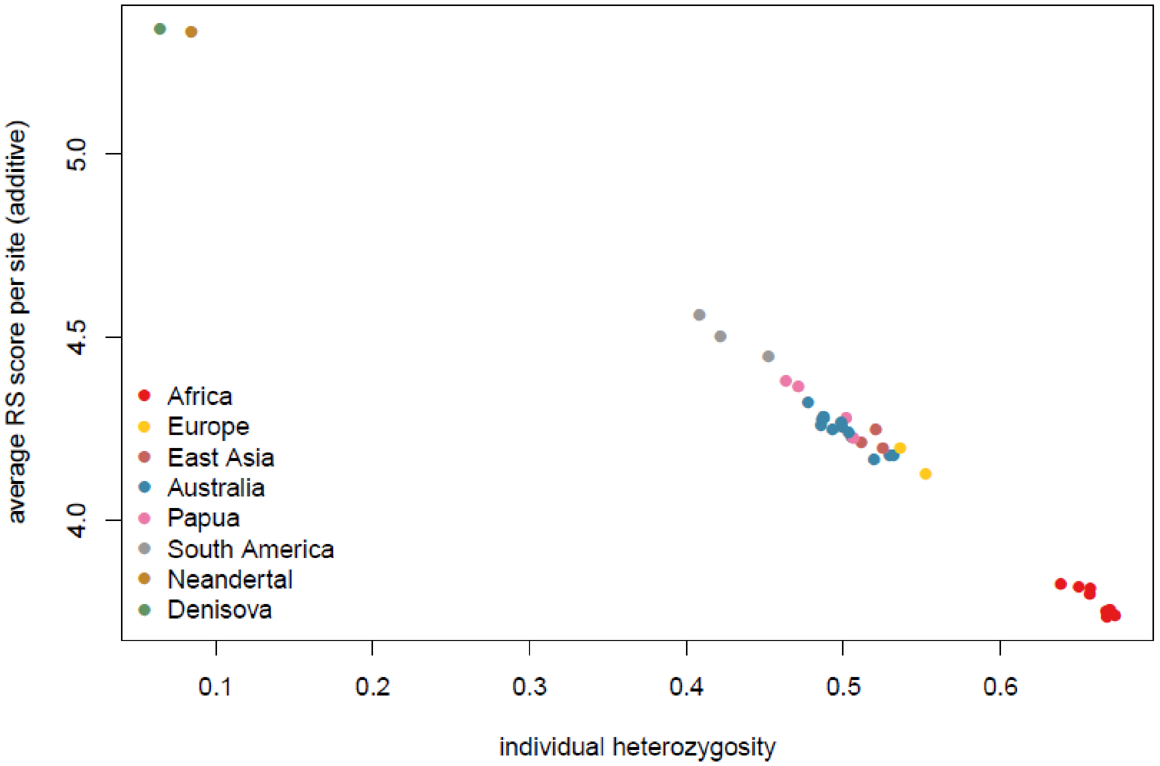


Figure S11.5 Relationship between heterozygosity and additive GERP RS score per individual.

Correlation between additive load and heterozygosity

Further support for the hypothesis that strong drift at the expansion front shaped neutral and functional diversity comes from a close to perfect correlation between per individual heterozygosity (calculated as the fraction of heterozygous sites among all segregating sites) and additive load (Figure S11.5, $R^2 = 0.99$).

Distinction between coding and non-coding regions

The results reported above are consistent with results obtained from non-coding regions (i.e, SNPs used for demographic inference, see Figures S11.6 – S11.8).

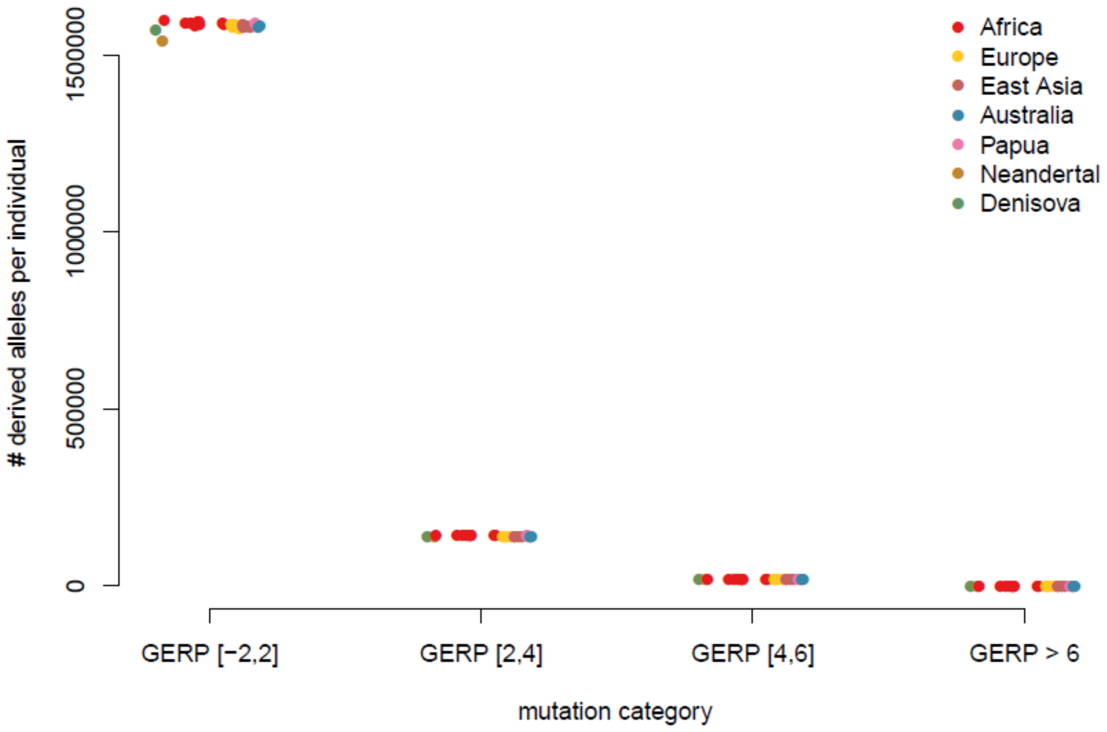


Figure S11.6 Number of alleles per individual for different GERP RS categories for non-coding regions.

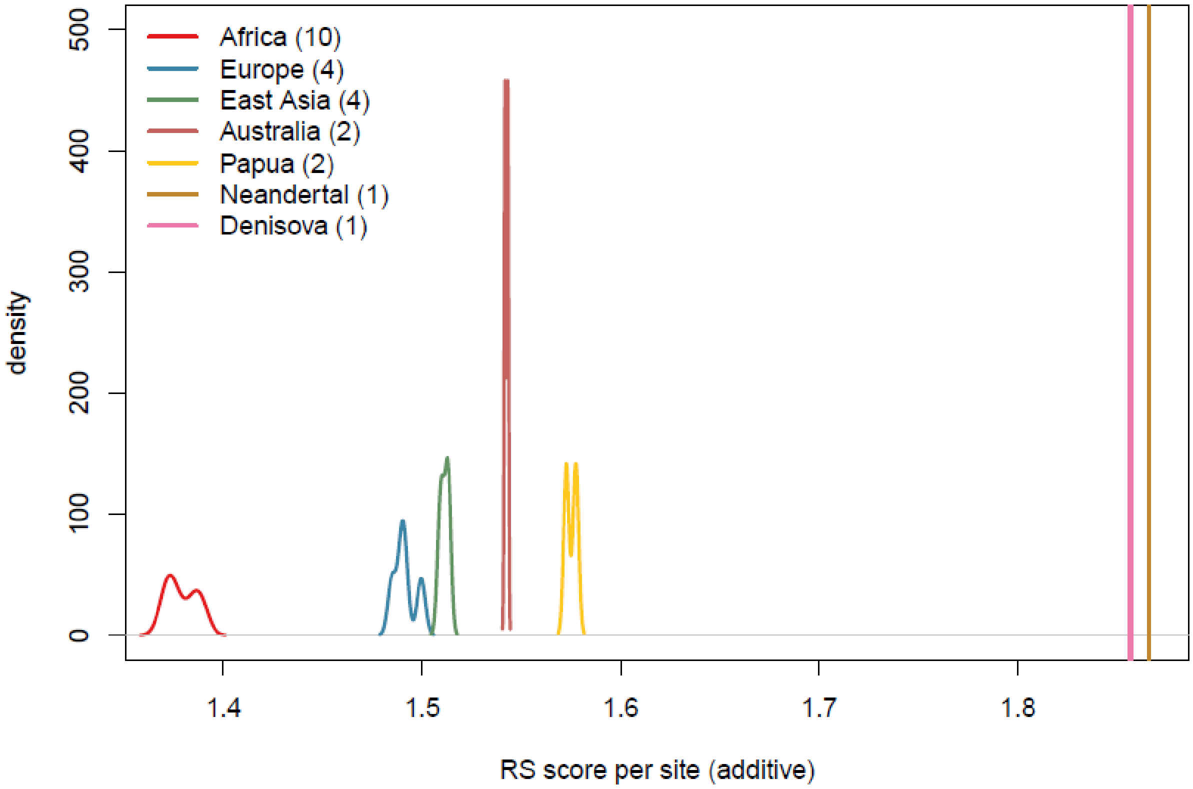


Figure S11.7 Average load per non-coding site for individuals from different populations.

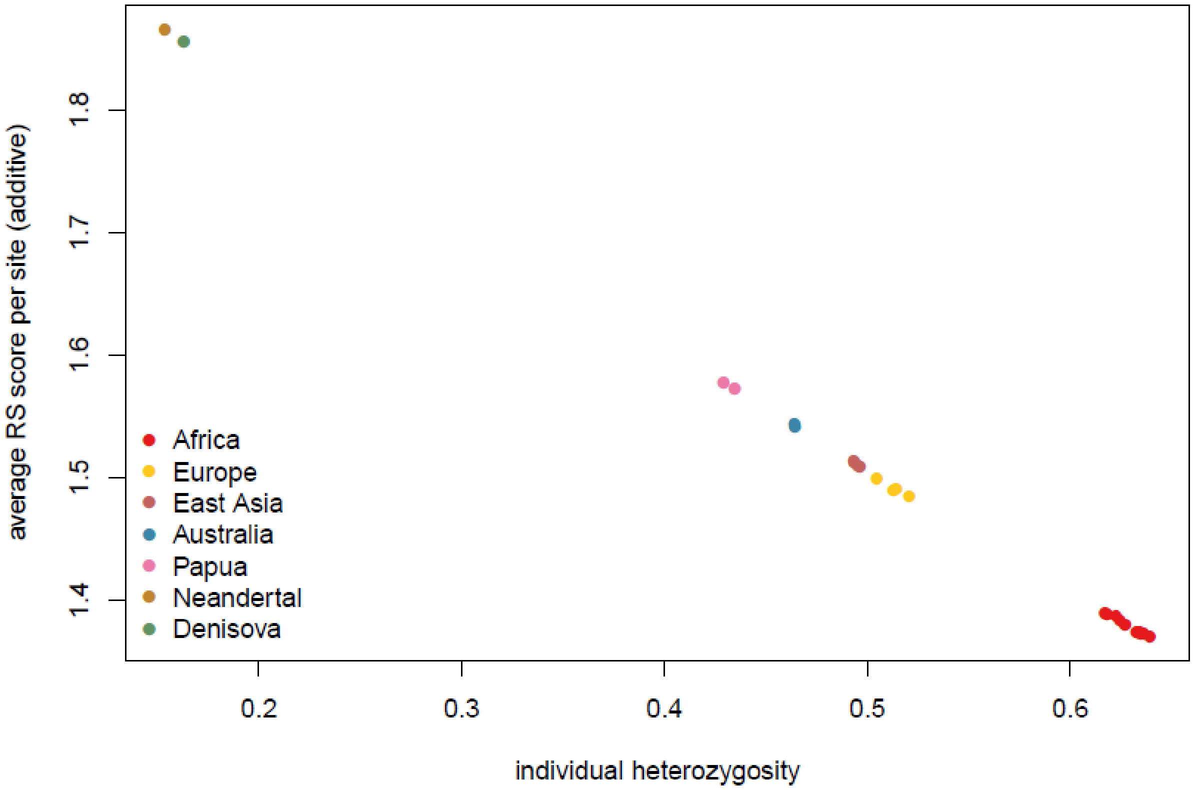


Figure S11.8 Correlation between heterozygosity and additive load per non-coding site.

Mutations shared with archaic populations

We next focus on mutations that are shared between Archaic individuals and non-African individuals, but are absent in African individuals. We tentatively consider these mutations as having introgressed from Archaics into modern humans. To distinguish between the effects of Denisovan and Neanderthal admixture, we also considered the subset of introgressed mutations that are found in either the Neanderthal individual or the Denisova individual, but not in both. We note that some sites that are considered as introgressed here might actually be shared with Africans, but were simply not detected in our nine African samples, and are therefore mis-identified as introgressed.

When focusing on Archaic or Neanderthal introgression (red and green boxplots in Figure S11.9, right column), we see a clear increase in the number of homozygous derived sites with distance from Africa. This increase is compatible with the idea that archaic introgression occurred during the early stage of the out of Africa expansion (e.g., in Europe), and that strong drift at the expansion front has increased homozygosity along the expansion axis. In contrast, the number of introgressed mutations shows a non-monotonic pattern (red and green boxplots in Figure S11.9, left column), with a maximum of introgressed mutations in Australians/Oceanians. This observation seems to be in conflict with the expected effect of the Out of Africa expansion on genetic diversity (i.e., a decrease of diversity along the expansion axis, see Figure S11.4). One explanation for this apparent discrepancy is that introgression would have occurred recurrently over a large geographic distance, stretching from Europe to East Asia, which would produce a gradual increase in the number of introgressed sites from Europe to East Asia. Strong drift and serial founder effects during the colonization of South America would have then led to the observed decrease of introgressed sites in South America. However, a more likely explanation is that the number of putative introgressed Neanderthal mutations is higher in Australo-Papuans because these populations also experienced Denisovan admixture. Given that Neanderthal and Denisovan share a common ancestor, introgressed sites can be misclassified in populations that experienced both Neanderthal and Denisovan admixture (see below). The pattern of Denisova introgression is simpler, and essentially limited to Australian and Oceanian individuals.

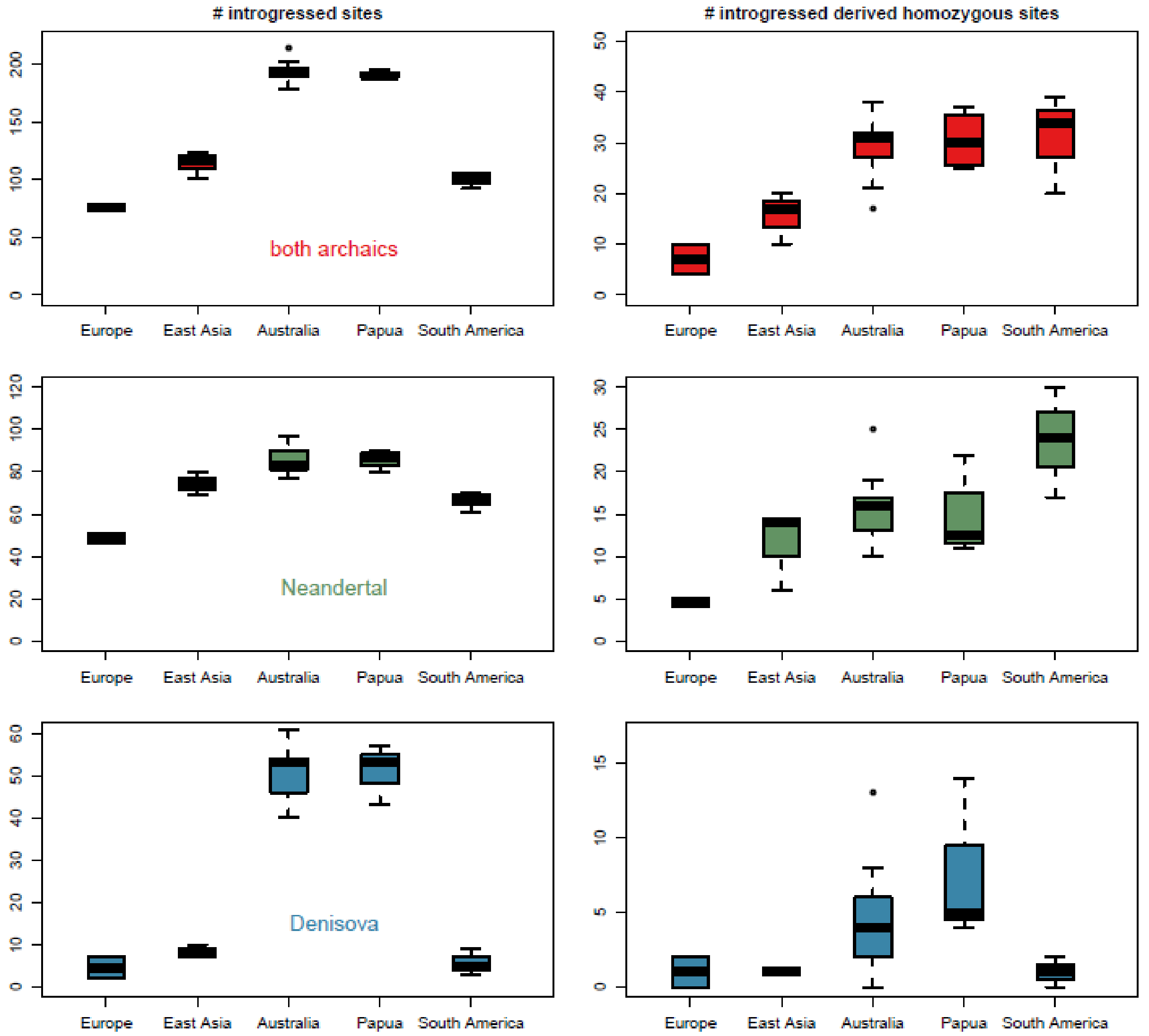


Figure S11.9 Archaic introgression: Left column: number of introgressed sites per individual. Right column: number of derived homozygous sites. Top row shows results for sites that are introgressed from either Neanderthal or Denisovans, middle row shows the subset of introgressed sites that are found only in the Neanderthal individual but not in Denisovans, and bottom row shows results for mutations that are found only in the Denisova individual but not in the Neanderthal individual.

A gradient of shared archaic load from Africa

We next consider the effect of archaic introgression on genetic load in modern humans. For mutations having potentially introgressed from Neanderthals, we find a clear increase of mutation load with distance from Africa for both measures of genetic load (Figure S11.10), whereas no particular spatial pattern is visible for sites introgressed from Denisovans. The latter observation makes sense as Denisovan introgression is limited to Australians and Papuans. The increase of the average load per site at Neanderthal introgressed sites can again be explained by strong drift at the expansion front and subsequent more efficient purging of deleterious mutations in core populations (e.g. Europe).

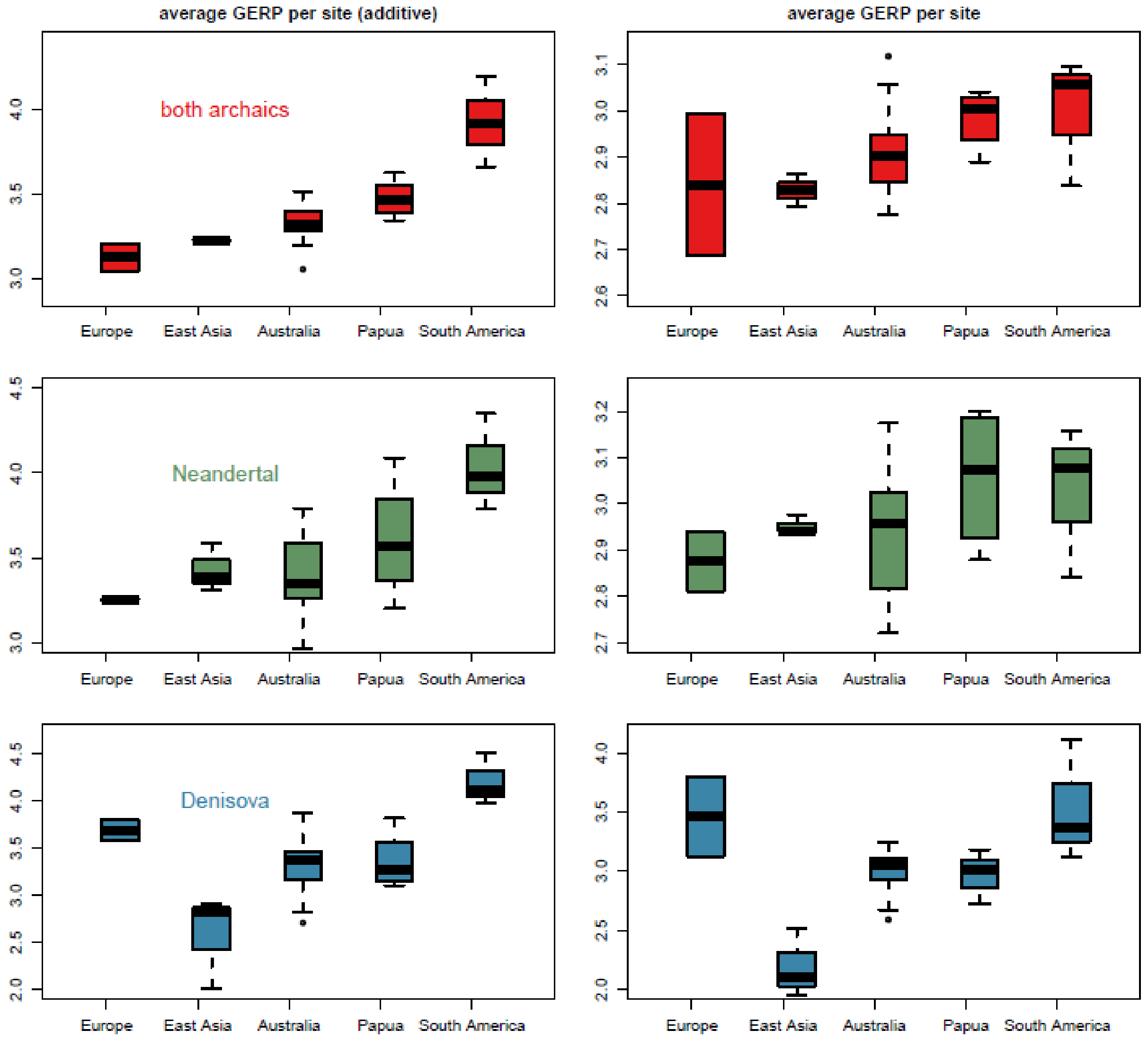


Figure S11.10 Left column: average additive GERP RS score per site. Right column: average GERP RS score per site. Top row shows results for sites that are introgressed from either Neanderthal or Denisovans, middle row shows the subset of introgressed sites that are found only in the Neanderthal individual but not in Denisovans, and bottom row shows results for mutations that are found only in the Denisova individual but not in the Neanderthal individual.

Quite surprisingly, the average RS score per site of mutations introgressed from Neanderthals is larger than the global average RS score in modern humans (red vs. blue boxplots in Figure S11.11), and there is a clear increase in this RS score with distance from Africa (red box plots, Figure S11.11). This striking feature cannot be explained by drift during colonisation since drift should not affect this average score, which is indeed approximately constant over all non-Africans when averaging across all the sites. Interestingly, the average RS score of introgressed sites in South Americans is very similar to the average RS score observed in Neanderthals. In contrast, we find that the average effect of introgressed sites in Europeans matches the global average that is mostly influenced by modern human specific sites. It suggests that the observed gradient could have been obtained by a more efficient purging of introgressed deleterious mutations in core populations, and less efficient with increasing distance from Africa. Difference in the efficiency of this purge among populations could either be due to longer time during which selection would have acted, or differences in population sizes making selection more efficient in large than in small populations.

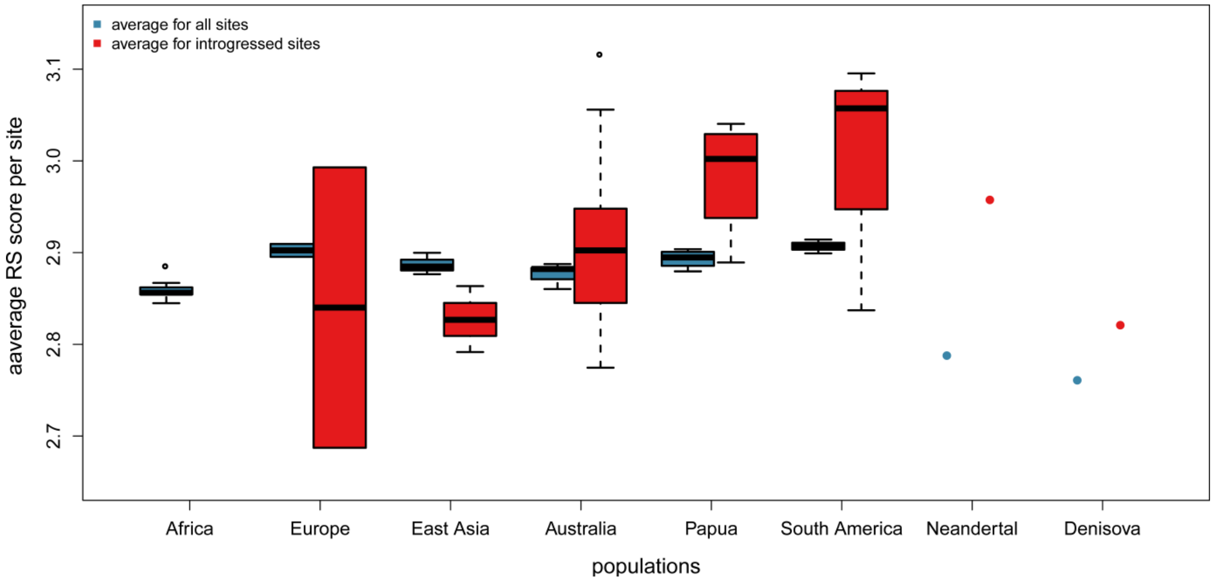


Figure S11.11 Distribution of average RS score per site across individuals from different populations. Red boxplots show the average RS score per site for mutations that are introgressed from Neanderthals and not found in Denisovans. The blue boxplots show results for the average across all sites.

Finally, when considering the effect of introgressed mutations on the average additive RS score per site, we find that even though the average RS score of introgressed sites in Neanderthals is larger than the average for sites located in exons in modern humans (Figure S11.12), Neanderthal introgressed sites have a reduced average additive load per site in modern humans (cf. blue and red boxplots in Figure S11.12). This observation implies that Neanderthal introgressed mutations are present at low frequencies, and are therefore more often found in heterozygotes than in non-introgressed sites with a comparable RS score.

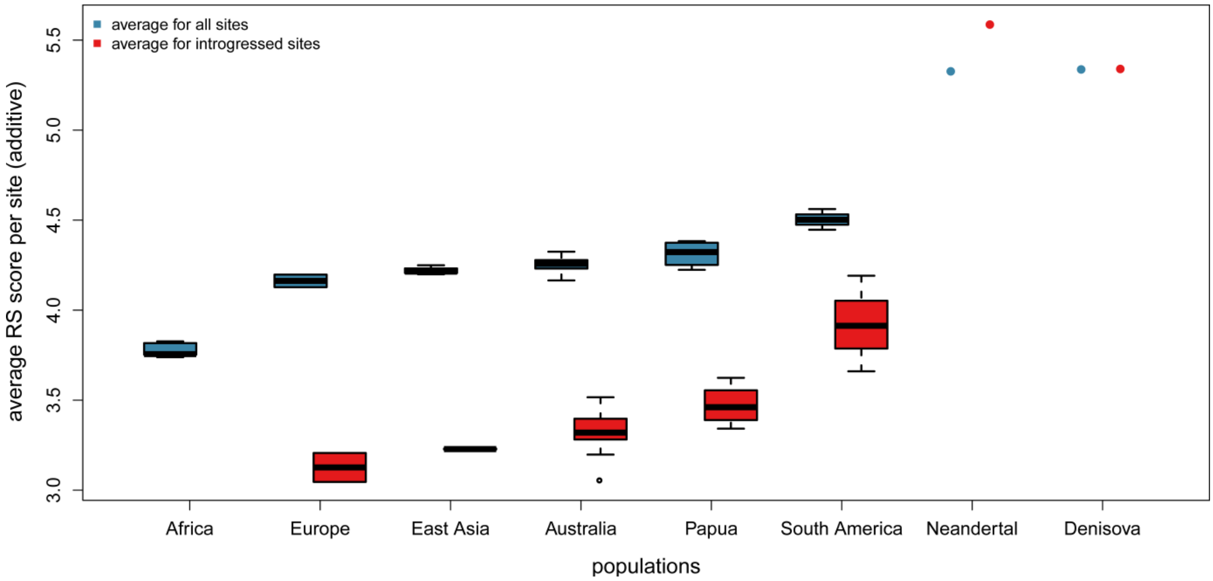


Figure S11.12 Average additive RS score per site. Blue boxplots show results for all mutations observed in modern humans. Red boxplots are for mutations that have introgressed exclusively from Neanderthals and that are not found in Denisovans.

Probability of misspecification of introgressed archaic sites

As seen in Figure S11.9, the observed number of introgressed sites private to Neanderthal vary across populations, with higher values found in Aboriginal Australians and Papuans (New Guinea). As discussed above, this could be interpreted as evidence for an extra pulse of Neanderthal admixture in these populations, even though there is no archaeological evidence that the range of Neanderthals extended that far. However, since these Oceanian populations also exhibit evidence of Denisovan admixture, the fact that Neanderthals and Denisovans share a common ancestor could lead to a misclassification of Neanderthal private sites, and vice-versa. Below we describe how the probabilities of misclassification can be computed using a coalescent argument.

We considered a site as being private introgressed Neanderthal (respectively Denisovan) if the derived allele is found both in a non-African (modern) population and in Altai Neanderthal (respectively Denisovan), and if the ancestral allele is fixed in both Altai Denisovan (respectively Altai Neanderthal) and African populations. Under the simplest case of one lineage sampled from each population (H in non-Africans, D in Altai Denisovan and N in Altai Neanderthal), there are three possibilities to obtain the pattern (D=ancestral allele, N=derived allele, H=derived allele) characteristic of a private Neanderthal site: (a) H and N coalesce in the Neanderthal population (Figure S11.13a); (b) H and N coalesce in the ancestral population of Denisovan and Neanderthal ("ND anc" population), with H introgressing from Neanderthal (Figure S11.13b); or (c) H and N coalesce in the ancestral population of Denisovan and Neanderthal, with H introgressing from Denisovan (Figure S11.13c). The two first cases correspond to correctly identified private introgressed Neanderthal sites in humans as being of Neanderthal origin, but the case (c) corresponds to a misspecification.

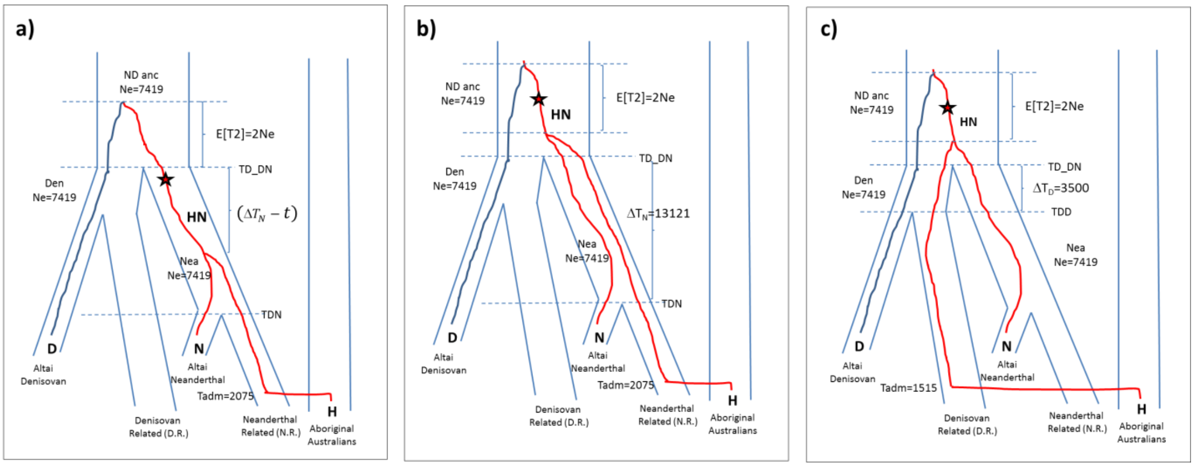


Figure S11.13 Gene trees that could lead to private Neanderthal sites (i.e. sites that are fixed ancestral in Denisovans and Africans, and that are derived in Altai Neanderthal and in a non-African population). For simplicity, in the above figures we only considered three lineages: H (non-African modern human), N (Altai Neanderthal) and D (Altai Denisovan). a) correctly classified Neanderthal private site due to coalescent of N and H in Neanderthal population. b) correctly classified Neanderthal private site due to coalescent of N and H in ancestor of Denisovan and Neanderthal population (i.e. shared ancestral polymorphism). c) gene tree where H introgressed from Denisovan related population (D.R.) but coalesces with N in the ancestor of Denisovan and Neanderthal population, leading to misclassified Neanderthal private sites. Effective population sizes and time intervals reported in the figure are based on the estimates from S07.

For simplicity we ignore the African lineage and assume that Africans are fixed ancestral. Looking backwards in time, a private Neanderthal site would be misclassified if the H lineage moves to the Denisovan related population (D.R.) due to admixture but coalesces with an Altai Neanderthal lineage (N) before coalescing with the Altai Denisovan (D), followed by a mutation in the HN branch (Figure S11.13c). This can only happen if the H lineage reaches the ancestral population of Denisovans and Neanderthals without coalescing with the Altai Denisovan D lineage. Thus, the misspecification is due to incomplete lineage sorting of ancestral polymorphism shared between Neanderthals and Denisovans. The probability that H coalesces with N, even though H is introgressed from a Denisovan related population, Pr(HN coal | H from DR), depends on: (i) the time interval ($\Delta T_{D \text{ in gen}}$) between the divergence of the Denisovan and Neanderthal (TD_DN) and the divergence of the Altai Denisovan from the D.R. (TDD); and (ii) the effective size of the ancestral Denisovan population (Den *Ne*). The larger the time interval (ΔT_D) and the smaller the effective size (Den *Ne*), the lower the probability that H coalesces with N before D. Assuming that all populations have the same effective size *Ne*, and that time is scaled in units of *2Ne*, this is given by

$$\begin{aligned}&\Pr(H,N\ coal\,|\,H\ from\ D.R.)\\&\qquad = \Pr(no\ coal\ in\ Den)\ \Pr(H,N\ coal\ NDanc|\ H,D\ or\ H,N\ coal)\\&\qquad = \frac{1}{2}e^{-\Delta T_D}\end{aligned}$$

where $\Delta T_D = \Delta T_{D \text{ in gen}}/(2Ne)$.

The probability that a mutation occurs in the HN branch, ignoring the split time from archaics and modern humans is given by

$$\Pr(mut\ in\ HN) = \int_0^\infty \theta\, t\, f(t)\, dt = \theta\, E[T_2]$$

Where θ is the scaled mutation rate (θ=2*Ne*), and $E[T_2]$ is the expected time for the coalescence of two lineages, which is $E[T_2]$ =1 in our rescaled time. The probability that a mutation occurs in the branch HN depends on its expected length and on the mutation rate, which depend on the effective size of the ancestral population of Denisovans and Neanderthals. Thus, the probability of a misclassified private Neanderthal site is thus given by

$$\Pr(misclassified) = admD\ \Pr(H,N\ coal\,|\,H\ from\ D.R.)\Pr(mut\ in\ HN)$$

where *admD* is the admixture contribution of Denisovan.

Similarly, we can compute the probability that a human lineage (H) that introgressed from the Neanderthal related population is correctly identified, i.e. H coalesces with N before coalescing with D. In that case, there are two possibilities.

First, HN can coalesce in the ancestor of Denisovan and Neanderthal (Figure S11.13b) due to incomplete lineage sorting. It depends on (i) the time interval ($\Delta T_{N \text{ gen}}$) between the divergence of the Denisovans and Neanderthals (TD_DN) and the divergence of the Altai Neanderthal from the N.R. (TDN); and (ii) the effective size of the ancestral Neanderthal population (Nea Ne, Figure S11.13b). The probability of such event, $\Pr(H,N\ coal\ in\ NDanc|\ H\ from\ N.R.)$, can be computed as above, by replacing ΔT_D by ΔT_N (=$\Delta T_{N \text{ gen}}$/2Ne) and *admD* by the Neanderthal admixture (*admN*).

Second, HN can coalesce in the Neanderthal population, i.e. $\Pr(H,N\ coal\ in\ Nea|\ H\ from\ N.R.)$, which corresponds to a case of complete lineage sorting (Figure S11.13a). The probability of such an event is given by

$$\Pr(H,N\ coal\ in\ Nea\ |\ H\ from\ N.R.) = 1 - e^{-\Delta T_N}$$

In this case, the probability of a mutation in the HN branch depends on the sum of the length in the Neanderthal population and the length in the ancestral NDanc population,

$$\Pr(mut\ in\ HN|\ H,N\ coal\ in\ Nea) = \theta(\ E[T_2] + \int_0^{\Delta T_N} (\Delta T_N -\ t)f(t)\ dt)$$

where $\theta = 2N_e\mu$ is the scaled mutation rate , $E[T_2]$ is the expected time for the coalescence of two lineages in the ancestral Neanderthal and Denisovan population, and $f(t) = e^{-t}$ is the distribution of the coalescent time for a sample of two lineages, assuming time is measured in units of *Ne*. The probability of correctly classifying a site is thus given by

$$\Pr(correct) = admN\ (\Pr(correct|H,N\ coal\ in\ Nea, mut\ in\ HN) \\ + \Pr(correct|H,N\ coal\ in\ NDanc, mut\ in\ HN))$$

where

$$\Pr(correct|H,N\ coal\ in\ Nea, mut\ in\ HN) \\ = (1 - e^{-\Delta T_N})\theta(\ E[T_2] + \int_0^{\Delta T_N} (\Delta T_N -\ t)f(t)\ dt)$$

and

$$\Pr(correct|H,N\ coal\ in\ NDanc, mut\ in\ HN) = \theta\frac{1}{2}e^{-\Delta T_N}$$

The corrected number of private sites ($\#correctedPrivateSites$) was computed based on the number of private sites detected ($\#PrivateSites$) and on the proportion of correctly specified sites, i.e.

$$\#correctedPrivateSites = \frac{\Pr(correct)}{\Pr(correct) + \Pr(incorrect)}\#PrivateSites$$

These probabilities are easily obtained assuming that the divergence times, effective sizes and admixture proportions are known. Substituting the parameters estimated under the SFS based likelihood approach (S07), we infer that the proportion of misspecified private Neanderthal and misspecified private Denisovan sites to be 0.39 and 0.18, respectively. The high value of 0.39 for the misspecification of private Neanderthal sites is mostly due to the old divergence between the Denisovan related population and the Altai Denisovan, which is about 3.5x older than the divergence of the Neanderthal related population (i.e., TDD ~ 3.5xTDN), which is thus decreasing the probability that the H lineage coalesces with the Altai D lineage.

Based on the parameter estimates inferred with the SFS-based analyses, we corrected the number of private Neanderthal (respectively, Denisovan) sites by discarding the proportion of misclassified sites that introgressed from Denisovan (respectively, Neanderthal) but share a more recent common ancestor with Altai Neanderthal (respectively, Altai Denisovan). These results suggest that the apparent excess of introgressed Neanderthal sites in Aboriginal Australians and Papuans can be explained by the Denisovan admixture, rather than by an extra pulse of admixture with Neanderthal on the Oceanian branch.

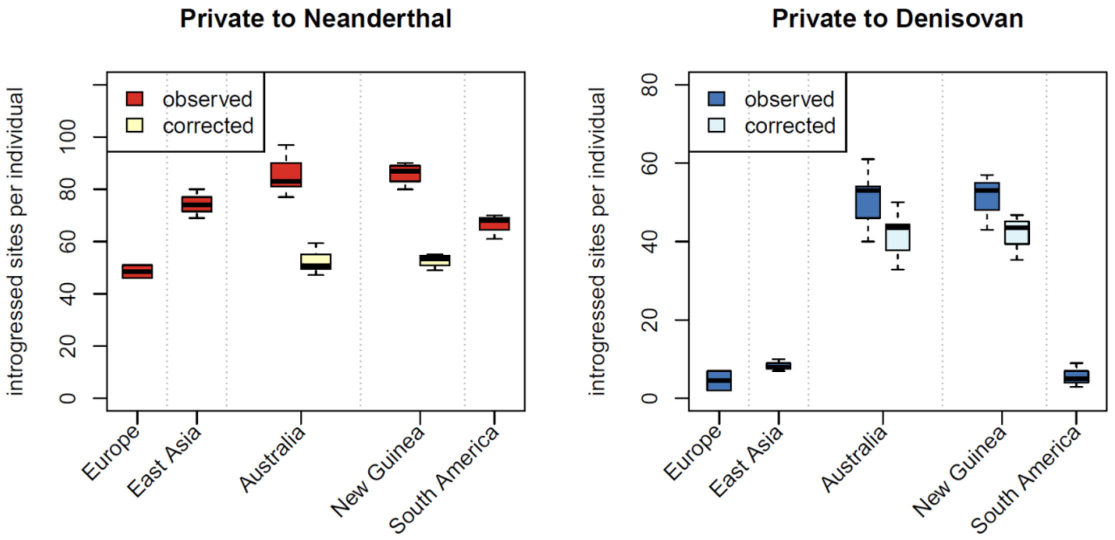


Figure S11.14 Per individual distribution of the number of putative introgressed sites from archaic humans: a) Neanderthal, and b) Denisovan. For populations with evidence of Neanderthal and Denisovan introgression (Australia and New Guinea), the figure shows the corrected proportion of introgressed sites after discarding misspecified sites due to shared ancestry of Denisovan and Neanderthal. Corrections were based on the coalescent argument described above, by replacing the parameter values by the ones estimated with the SFS for the best Out of Africa model (S07).

S11 References

Cooper GM, Stone EA, Asimenos G, Program NCS, Green ED, Batzoglou S, Sidow A 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15: 901-913.

Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6: e1001025.

Henn B, Botigue LR, Peischl S, et al. 2015. Distance from Sub-Saharan Africa Predicts Mutational Load in Diverse Human Genomes. bioRxiv.

Lohmueller KE 2014. The distribution of deleterious genetic variation in human populations. Curr Opin Genet Dev 29: 139-146.

Meyer M, Kircher M, Gansauge MT, et al. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338: 222-226.

Prufer K, Racimo F, Patterson N, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505: 43-49.

Rosenbloom KR, Armstrong J, Barber GP, et al. 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 43: D670-681.

Wang K, Li M, Hakonarson H 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.

S12 Uniparental markers

Enrico Macholdt, Chiara Barbieri, Shengyu Ni, Mark Stoneking

Data

The analysis of maternal and paternal lineages was based on 83 complete mitochondrial genomes (mtDNA) with an average sequencing depth of 3,484 ± 1,515 (mean ± SD) and on 9,601 segregating sites in 13,583,154 callable positions from 44 Y chromosomes (Ychr) with an average sequencing depth of 28.9 ± 4.5 (mean ± SD). Note that the sample size per population is rather low (6-13 for mtDNA and 1-10 for Ychr) which limits our power.

Sequence processing and haplogroup calling

The mtDNA sequence reads were processed by an in-house consensus caller similar to the one developed in Li et al. (2015). The 83 consensus sequences and two mitochondrial references (RSRS and rCRS) were aligned using MUSCLE v3.8.31 (Edgar 2004). The alignment was visually checked for inconsistencies in aligning positions with equal probability. An alignment without the poly C regions was produced by excluding two regions: 16182-16193 and 303-315 (positions refer to RSRS). Subsequently, haplogroup calling was performed with haplogrep (van Oven and Kayser 2009; Kloss-Brandstätter et al. 2011) and haplofind (Vianello et al. 2013). Contradicting calls from these methods were examined manually by checking the overlap of derived positions in the sample and haplogroup defining positions extracted from Phylotree (van Oven and Kayser 2009). Additional information was gathered from a manual screening of a Median Joining Network (Bandelt et al. 1999) .

The data processing and quality filtering of the 44 Aboriginal Australian Y chromosome sequences was performed in multiple steps. First, the total callable region was determined by mapping the sequencing reads to the hg19 Y chromosome using BWA-MEM (Li and Durbin 2009). GATK (McKenna et al. 2010) was then used for realignment and base recalibration with the database in 1000G phase 3 and ISOGG (ISOGG). All emitted sites were selected by the following filter:
(MQ<40.0 || (MQ0>3 && 10*MQ0>DP) || DP<363 || DP>1805). Y chromosome positions with sequencing depth > 23 for female samples were filtered as described in Poznik et al. (2013). One additional filter which retains positions with > 95% of the samples having > 5x coverage was adapted from Karmin et al. (2015). Finally, the callable region containing 13,583,154 bp was submitted to the GATK UnifiedGenotyper (McKenna et al. 2010) with SNP reporting filters: -stand_call_conf 30 -stand_emit_conf 10 and quality filters**:** (QD<2.0 || MQ<40.0 || FS>60.0 || HaplotypeScore > 13.0 || MQRankSum<-12.5 || ReadPosRankSum<-8.0 || (MQ0>3 && 10*MQ0>DP)). Haplogroup assignment was performed with an in-house script that matched our SNPs with the classification provided in ISOGG (ISOGG) version 10.08. Haplogroup assignment was then manually verified comparing our sequence data with the available data for HGDP-CEPH individuals typed for diagnostic SNPs in Lippold et al. (2014).

The haplogroup counts for mtDNA and Y chromosome sequences per population are summarized in Table S12.1. We found predominantly mtDNA haplogroups and subclades which have been previously reported in Australia and Oceania (Ingman and Gyllensten 2003; gounder Palanichamy et al. 2004; Friedlaender et al. 2005; van Holst Pellekaan et al. 2006; Kivisild et al. 2006; Hudjashov et al. 2007), namely M42, M42a, M14, N13, O, O1, O1a, P, P3b, P4b1, P5, R12, S1a, S2 and S5. Furthermore, two samples belonged to characteristic European subclades of H1 and one sample carried Southeast Asian haplogroup E1a2. There are clearly more European lineages in the Y chromosome than in the mtDNA, as 14 out of 44 samples belonged to European haplogroups J2a1b, R1b and I1a2a, whereas only two out of 83 mtDNA sequences belonged to subclades of the European mtDNA haplogroup H1. The non-European Y chromosomes include 25 individuals from Australian haplogroups C1b or K2b, three individuals from East or Southeast Asian haplogroups O1a (BDV05 and WPA06) and O2a (CAI05), and two individuals from haplogroup E1b1b which belong to two distinct sub branches found in Europe and also in the north-east of Africa (ENY07 and WON09) (Cruciani et al. 2004; Semino et al. 2004; Cruciani et al. 2007).

Analysis of genetic variance and population structure

Analyses of mtDNA and Y chromosome variation and structure were carried out using Arlequin (Excoffier and Lischer 2010) version 3.5.1.3, R (R Core Team 2014) and R packages ca**,** ape**,** pegas**,** adegenet**,** ade4**,** MASS**,** lattice**,** vegan**,** fields and gplots (Venables and Ripley 2002; Paradis et al. 2004; Dray and Dufour 2007; Nenadic and Greenacre 2007; Jombart 2008; Sarkar 2008; Paradis 2010; Nychka et al. 2015; Oksanen et al. 2015; Warnes et al. 2015). The Analysis of Molecular Variance (AMOVA, Table S12.2) for mtDNA sequences revealed strong and statistically significant differences among populations that explained 16.82 % (p-value=9.999*10^{-5}) of the total variance. By contrast, only 11.28 % (p-value=0.00889911) of the total variance in the Y chromosome sequences was explained by among population differences. To check if these differences could be explained by the different number of samples for both markers, we performed the AMOVA for 1000 independent random subsamples of 44 mtDNA sequences. A mean among-population variance of 16.45 (95%CI=16.15-16.76), indicates that the different results for mtDNA vs. the Y chromosome are not explained by sample size differences. Note that the among-population variance of mtDNA sequences falls outside of the resampled confidence interval as well. This is probably because not all of the populations are represented in the random subsets. The magnitude of between population variance found in the mtDNA genomes is particularly high for human populations, and it is similar to the values found between 19 Khoisan hunter-gatherers and pastoralist populations of southern Africa, for which a matrilocal or multilocal post marriage residential pattern was suggested (Barbieri et al. 2014). In comparison, 27 patrilocal Bantu speaking populations from different linguistic clusters inhabiting four countries in southern Africa show a between-population variance of 5.5% (Barbieri et al. 2014) and a worldwide panel of 51 populations shows a between-population variance of 25% (Lippold et al. 2014).

We also tested for the presence of geographic structure in the uniparental markers by classifying Australian populations into two sub-groups (Northeast: WPA, CAI and Southwest: WCD, WON, NGA, ENY; the PIL, BDV and RIV groups were omitted from this analysis). MtDNA shows substantial structure corresponding to our specific sampling of these two different geographical groups, because the variance between groups (9.44%, Table S12.2) is higher than the variance among populations assigned to the same group (6.99 %, Table S12.2). The Ychr data does not indicate any structure corresponding to these two ancestries.

To further investigate the relationships among populations, we produced a non-metric Multi-dimensional Scaling (MDS) plots based on Φ_{ST} pairwise genetic distances between sequences and Correspondence Analysis (CA) plots using the population by haplogroup contingency table (Figure S12.1 A-D). The mtDNA MDS analysis revealed BDV, RIV and CAI as outliers, while CAI, RIV and WPA were more isolated in the CA plot (caused by haplogroups P, N13 and E1a2 for CAI, P4b1 and European haplogroup H1bs for RIV, and P5 for WPA. Although we present the Ychr MDS and CA plots, the sample sizes are too small to allow for any interpretation ($n \leq 10$).

Low values of mtDNA haplotype diversity along with the highest values of nucleotide diversity (Table S12.1) are found in WCD and PIL, suggesting ancient structure along with a recent decrease in population size in the western part of the Australian continent. Elevated Ychr nucleotide diversity values are mainly but not exclusively found in populations from the same geographical area. There are no significant correlations between mtDNA and Ychr genetic distances (Mantel statistic: r=0.057, p=0.39), nor between genetic and geographic distances (using the average sampling location per population) for both markers (geo vs. mtDNA distance: r=0.223, p=0.14; geo vs. Ychr: r=0.041, p=0.4). The number of sequence differences between individuals (Figure S12.1E-F) are consistent with the AMOVA results in showing mostly large differences between individuals from different populations, although there is some sharing of similar haplotypes between PIL and WCD (both mtDNA and Ychr), and between NGA and WON (mtDNA only).

Bayesian phylogenetic analysis

Bayesian phylogenetic analyses of the uniparental markers were performed with BEAST 1.8.0 (Drummond et al. 2012) using a Markov chain Monte Carlo (MCMC) algorithm. The best fitting substitution model for mtDNA and Ychr sequences was determined by jModelTest v2.7.0 (Darriba et al. 2012) to be Tamura-Nei, 93 model (Tamura and Nei 1993) with Invariant Sites model plus Gamma Rate Distribution (TN93+I+G) and the Generalized time reversible model (Tavaré 1986) (GTR), respectively. The analyses were executed with a *strict* clock model, and *Coalescent: Bayesian Skyline* tree model including *piecewise-linear* skyline model (Drummond et al. 2005). The mtDNA sequences were partitioned into *coding* and *non-coding* regions and the mutation rates were set to $1.708*10^{-8}$ and $9.883*10^{-8}$ substitutions/site/year respectively as previously described (Soares et al. 2009). From the variety of published Y chromosome mutation rates (Xue et al. 2009; Mendez et al. 2013; Poznik et al. 2013; Helgason et al. 2015; Karmin et al. 2015) we chose $0.82*10^{-9}$ mutations/bp/year, which was calibrated by the entry into the Americas and the divergence time estimates of two haplogroup Q lineages (Poznik et al. 2013) and is also quite similar to the estimate from Icelandic pedigrees (Helgason et al. 2015). Because the Y chromosome alignment consists only of variable sites, the BEAST XML input files were corrected with information about the invariant site base composition of the callable region (Barbieri et al. submitted). The first 20% of each BEAST MCMC chain (10 thousand of 50 million steps) were discarded as burn in. BEAST's TreeAnnotator was used to annotate the Maximum Clade Credibilty (MCC) trees (Figure S12.2A-B). Figure S12.2A shows the deep coalescence of all the major mtDNA branches; most of them diverge between ~45-55 kya. Figure S12.2B shows a similar profile, with the coalescence of most of the major Ychr branches ~40-45 kya; only haplogroup R1b, of possible European origin, is characterized by a recent divergence of its lineages ~5kya.

We explored the changes in effective population sizes over time by visualizing them with Bayesian Skyline Plots (BSP, Figure S12.2C-E). For mtDNA, the BSP for all available genomes (n = 83), excluding samples with European mtDNA haplogroup H (n = 81), and the coding region of all samples (n = 83) all show a very similar trend, with an old expansion in population size at 50-40 kya and a tendency for a recent slight expansion followed by a drop in population size in the last 10 kya. By considering groups from northeastern Australia (CAI, WPA) and southwestern Australia (WCD, WON, ENY, NGA), we find indications that the recent trend in population size change is composed of different signals from these two subsets. The BSP based on populations from northeastern Australia shows a constant population size for the past 10ky, whereas populations from southwestern Australia show a slight decrease during this time. The BSP for all Ychr sequences (n = 44) exhibits an old expansion in the same time range as for mtDNA (Figure S12.2E); however, in the last 10,000 years the population size seemed to drop and then expand. This recent signal of population size change disappears when European Ychr sequences are excluded, or when analysing only sequences from autochthonous haplogroups K2b and C1b. This suggests a strong European-related influence on the recent change in population size of Australian Ychrs.

Table S12.1 MtDNA and Y chromosome haplogroup counts per population and diversity statistics, including number of haplotypes, haplotype diversity, and nucleotide diversity and variance; n: number of samples; hts: number of haplotypes

mtDNA	**E1a2**	**H1bs**	**H1e1a3**	**M14**	**M42**	**M42a**	**N13**	**O**	**O1**	**O1a**	**P**	**P3b**	**P4b1**	**P5**	**R**	**R12**	**S1a**	**S2**	**S5**	**nucleotide diversity**	**variance**	**n**	**hts**	**haplotype diversity**
BDV	-	-	-	-	-	-	-	-	-	2	-	-	-	-	-	-	6	1	-	0.00145	$6.41*10^{7}$	9	4	0.8056
CAI	1	-	-	-	1	1	1	-	-	-	6	-	-	-	-	-	-	-	-	0.00208	$1.25*10^{6}$	10	8	0.9333
ENY	-	-	1	-	2	-	-	1	-	-	-	-	1	-	-	1	-	2	-	0.00214	$1.42*10^{6}$	8	8	1.0000
NGA	-	-	-	-	-	-	-	2	1	1	-	-	-	-	-	1	1	-	-	0.00172	$1.04*10^{6}$	6	6	1.0000
PIL	-	-	-	-	5	-	-	1	1	-	-	2	-	-	2	-	-	-	1	0.00239	$1.58*10^{6}$	12	8	0.8939
RIV	-	1	-	-	-	2	-	-	-	-	-	-	5	-	-	-	-	-	-	0.00167	$8.76*10^{7}$	8	7	0.9643
WCD	-	-	-	1	5	-	-	-	1	1	-	1	-	-	1	2	1	-	-	0.00239	$1.56*10^{6}$	13	9	0.8718
WON	-	-	-	-	-	-	-	4	1	3	-	-	-	-	1	2	-	-	-	0.00184	$9.66*10^{7}$	11	8	0.9455
WPA	-	-	-	-	1	1	-	-	-	-	1	-	-	3	-	-	-	-	-	0.00206	$1.48*10^{6}$	6	6	1.0000

Ychr	**C1b**	**E1b1b**	**I1a2a**	**Ja1**	**K2b**	**O1a**	**O2a**	**R1b**	**nucleotide diversity**	**variance**	**n**	**hts**	**haplotype diversity**
BDV	-	-	1	-	-	1	-	2	0.08395	$3.01*10^{3}$	4	4	1
CAI	1	-	-	-	5	-	1	1	0.09898	$2.92*10^{3}$	8	8	1
ENY	1	1	-	-	-	-	-	1	0.14317	$1.14*10^{2}$	3	3	1
NGA	-	-	-	-	-	-	-	1	NA	NA	1	1	NA
PIL	6	-	-	-	1	-	-	1	0.07949	$1.88*10^{3}$	8	8	1
RIV	-	-	1	1	-	-	-	1	0.09564	$5.09*10^{3}$	3	3	1
WCD	5	-	-	1	4	-	-	-	0.10730	$3.22*10^{3}$	10	10	1
WON	-	1	2	-	-	-	-	1	0.11132	$5.28*10^{3}$	4	4	1
WPA	1	-	-	-	1	1	-	-	0.12358	$8.49*10^{3}$	3	3	1

Table S12.2 AMOVA variance estimates for mtDNA and Y chromosome sequences. *significant with alpha=0.05

Source of Variation	mtDNA	Y chromosome
Among nine Population	16.82*	11.28*
Within populations	83.18*	88.72*
Among two geographical groups	9.44*	0.83
Among Population within geographical groups	6.99*	2.62
Within populations	83.57	96.55

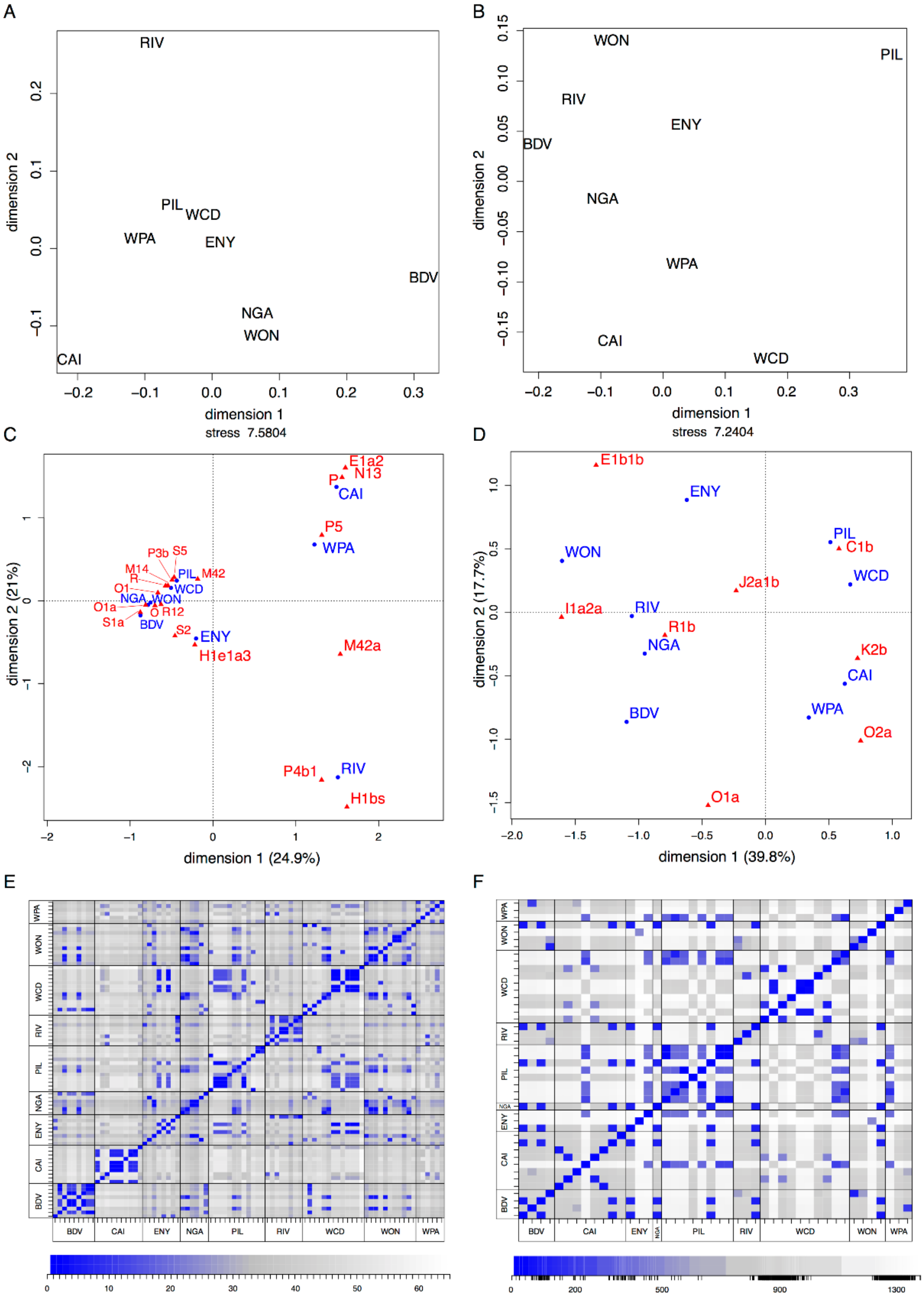
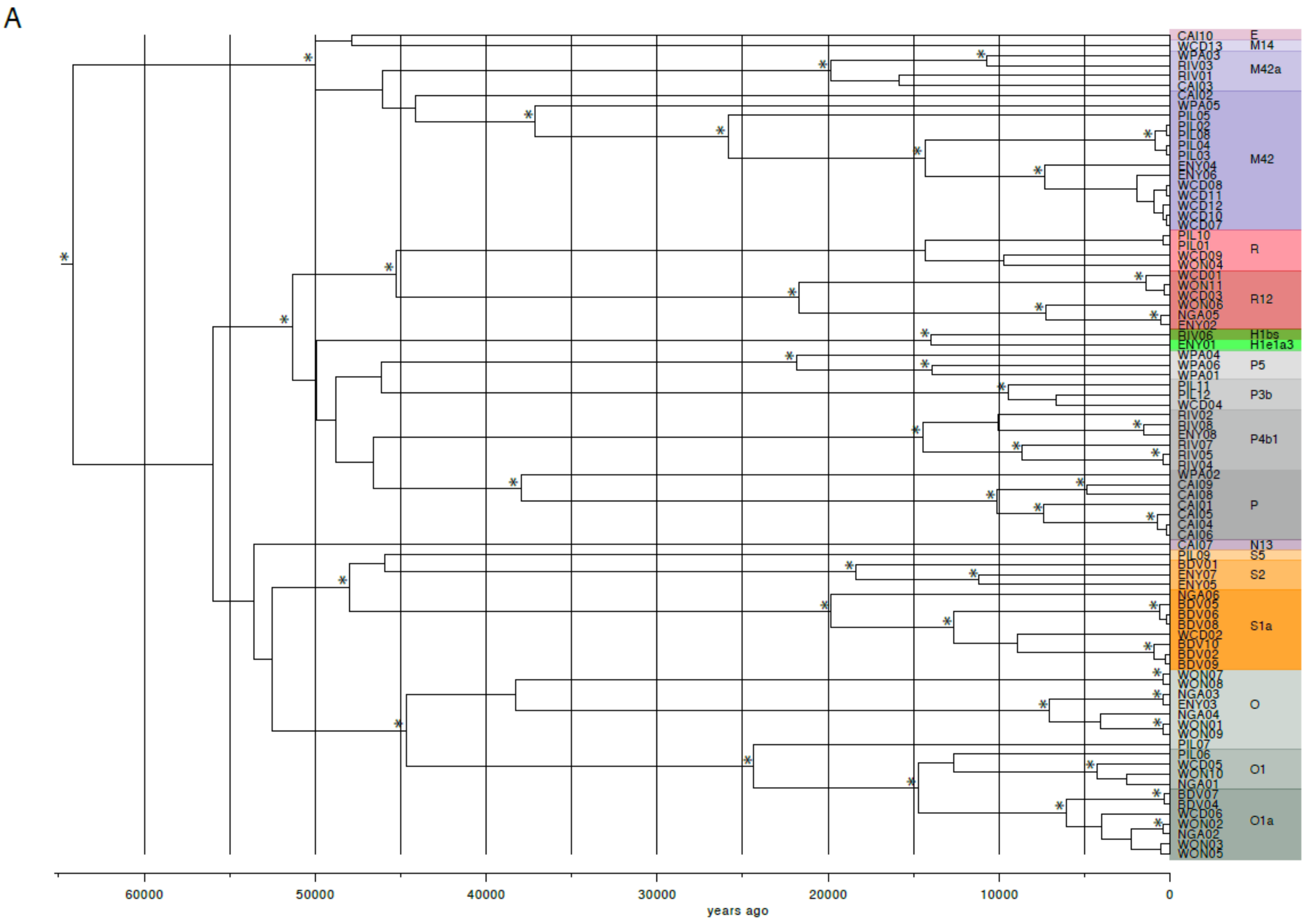


Figure S12.1 MtDNA (left panels) and Y chromosome (right panels) diversity results. A-B Non-metric Multidimensional Scaling Plots based on Φ_{ST} distances; C-D Correspondence Analyses based on haplogroup (red) frequencies within populations (blue); E-F heatplots of pairwise number of differences between individuals.


A
CAI10
WCD13
WPA03
RIV03
RIV01
CAI03
CAI02
WPA05
PIL05
PIL02
PIL08
PIL04
PIL03
ENY04
ENY06
WCD08
WCD11
WCD12
WCD10
WCD07
PIL10
PIL01
WCD09
WON04
WCD01
WON11
WCD03
WON06
NGA05
ENY02
RIV06
ENY01
WPA04
WPA06
WPA01
PIL11
PIL12
WCD04
RIV02
RIV08
ENY08
RIV07
RIV05
RIV04
WPA02
CAI09
CAI08
CAI01
CAI05
CAI04
CAI06
CAI07
PIL09
BDV01
ENY07
ENY05
NGA06
BDV05
BDV06
BDV08
WCD02
BDV10
BDV02
BDV09
WON07
WON08
NGA03
ENY03
NGA04
WON01
WON09
PIL07
PIL06
WCD05
WON10
NGA01
BDV07
BDV04
WCD06
WON02
NGA02
WON03
WON05
E
M14
M42a
M42
R
R12
H1bs
H1e1a3
P5
P3b
P4b1
P
N13
S5
S2
S1a
O
O1
O1a
60000
50000
40000
30000
20000
10000
0
years ago

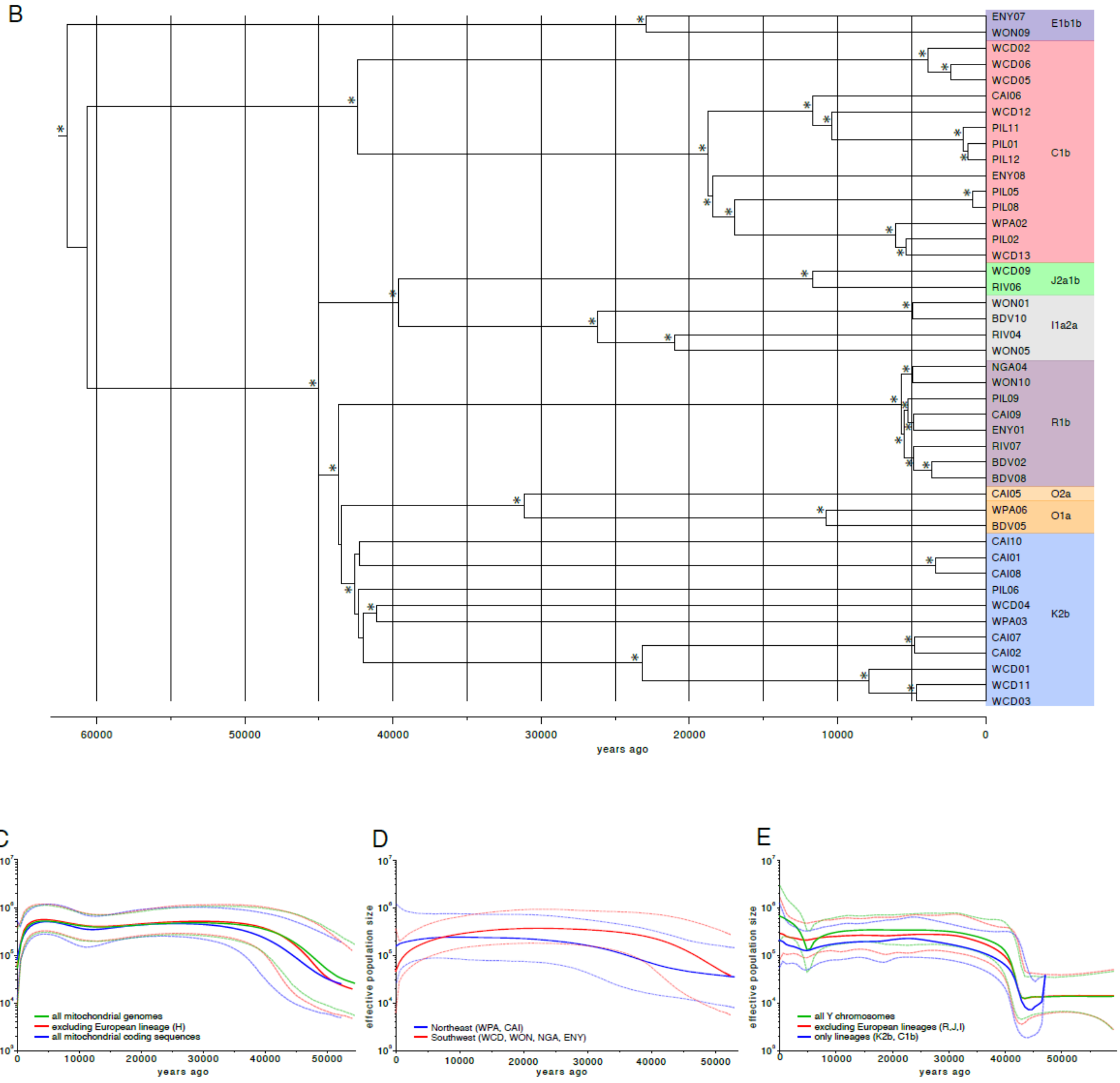


Figure S12.2 Bayesian Maximum Clade Credibility trees of (A) mtDNA sequences and (B) Y chromosome sequences, colour coded by haplogroups and subhaplogroups; asterisk marks nodes with a posterior probability >0.95; Bayesian Skyline plots (C-E) showing effective population size (N_e) estimates over time.

S12 References

Bandelt H-J, Forster P, Röhl A. 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16:37–48.

Barbieri C, Vicente M, Oliveira S, Bostoen K, Rocha J, Stoneking M, Pakendorf B. 2014. Migration and interaction in a contact zone: mtDNA variation among Bantu-speakers in Southern Africa. PloS One 9:e99117.

Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, Moral P, Watson E, Guida V, Colomb EB, Zaharova B, et al. 2004. Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. Am. J. Hum. Genet. 74:1014–1022.

Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, Dugoujon J-M, Crivellaro F, Benincasa T, Pascone R, et al. 2007. Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. Mol. Biol. Evol. 24:1300–1311.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772.

Dray S, Dufour AB. 2007. The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. 22:1–20.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22:1185–1192.

Drummond AJ, Suchard M a, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Excoffier L, Lischer HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10:564–567.

Friedlaender J, Schurr T, Gentz F, Koki G, Friedlaender F, Horvat G, Babb P, Cerchio S, Kaestle F, Schanfield M, et al. 2005. Expanding Southwest Pacific Mitochondrial Haplogroups P and Q. Mol. Biol. Evol. 22:1506–1517.

gounder Palanichamy M, Sun C, Agrawal S, Bandelt H-J, Kong Q-P, Khan F, Wang C-Y, Chaudhuri TK, Palla V, Zhang Y-P. 2004. Phylogeny of Mitochondrial {DNA} Macrohaplogroup N in India, Based on Complete Sequencing: Implications for the Peopling of South Asia. Am. J. Hum. Genet. 75:966–978.

Helgason A, Einarsson AW, Gu\dhmundsdóttir VB, Sigur\dhsson Á, Gunnarsdóttir ED, Jagadeesan A, Ebenesersdóttir SS, Kong A, Stefánsson K. 2015. The Y-chromosome point mutation rate in humans. Nat. Genet. 47:453–457.

van Holst Pellekaan SM, Ingman M, Roberts-Thomson J, Harding RM. 2006. Mitochondrial genomics identifies major haplogroups in Aboriginal Australians. Am. J. Phys. Anthropol. 131:282–294.

Hudjashov G, Kivisild T, Underhill PA, Endicott P, Sanchez JJ, Lin AA, Shen P, Oefner P, Renfrew C, Villems R, et al. 2007. Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. Proc. Natl. Acad. Sci. 104:8726–8730.

Ingman M, Gyllensten U. 2003. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. Genome Res. 13:1600–1606.

ISOGG. Y-DNA Haplogroup Tree 2015. Available from: http://www.isogg.org/tree/

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.

Karmin M, Saag L, Vicente M, Sayres MAW, Järve M, Talas UG, Rootsi S, Ilumäe A-M, Mägi R, Mitt M, et al. 2015. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. Genome Res.

Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, Underhill PA, Scharfe C, Torroni A, et al. 2006. The Role of Selection in the Evolution of Human Mitochondrial Genomes. Genetics 172:373–387.

Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F. 2011. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum. Mutat. 32:25–32.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li M, Schröder R, Ni S, Madea B, Stoneking M. 2015. Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. Proc. Natl. Acad. Sci. 112:2491–2496.

Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schröder R, Stoneking M. 2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investig. Genet. 5:13.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Mendez FL, Krahn T, Schrack B, Krahn A-M, Veeramah KR, Woerner AE, Fomine FLM, Bradman N, Thomas MG, Karafet TM, et al. 2013. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. Am. J. Hum. Genet. 92:454–459.

Nenadic O, Greenacre M. 2007. Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package. J. Stat. Softw. 20:1–13.

Nychka D, Furrer R, Sain S. 2015. fields: Tools for Spatial Data.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H. 2015. vegan: Community Ecology Package.

van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 30:E386–E394.

Paradis E. 2010. pegas: an {R} package for population genetics with an integrated--modular approach. Bioinformatics 26:419–420.

Paradis E, Claude J, Strimmer K. 2004. A{PE}: analyses of phylogenetics and evolution in {R} language. Bioinformatics 20:289–290.

Poznik GD, Henn BM, Yee M-C, Sliwerska E, Euskirchen GM, Lin AA, Snyder M, Quintana-Murci L, Kidd JM, Underhill PA, et al. 2013. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science 341:562–565.

R Core Team. 2014. R: A Language and Environment for Statistical Computing.

Sarkar D. 2008. Lattice: Multivariate Data Visualization with R. New York: Springer

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, et al. 2004. Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am. J. Hum. Genet. 74:1023–1034.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am. J. Hum. Genet. 84:740–759.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.

Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. Fourth. New York: Springer

Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C. 2013. HAPLOFIND: A New Method for High-Throughput mtDNA Haplogroup Assignment. Hum. Mutat. 34:1189–1194.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, et al. 2015. gplots: Various R Programming Tools for Plotting Data.

Xue Y, Wang Q, Long Q, Ng BL, Swerdlow H, Burton J, Skuce C, Taylor R, Abdellah Z, Zhao Y, et al. 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr. Biol. 19:1453–1457.

S13 Spatial analyses

Oscar Lao, Anders Eriksson, Andrea Manica

Correlation between genetics and geography

Background

We investigated to which extent the patterns of genetic diversity present in Aboriginal Australian populations relate to geography. In particular, we studied *i*) the correlation between genetic differentiation and geodesic distance among individuals and *ii*) the main geographic axes of genetic differentiation among Aboriginal Australian groups.

Methods

We studied the genetic relationship among Aboriginal Australian populations by computing a multidimensional scaling (MDS) on an identity-by-state (IBS) matrix between the 69 unrelated Aboriginal Australian individuals (S04) using 433,340 LD pruned SNPs; the geographic resemblance of the first two dimensions of the MDS was estimated by means of a Procrustes analysis (Wang et al. 2010). Anisotropy of genetic diversity within the Australian continent was computed by means of the Bearing correlogram (Rosenberg 2000) as implemented in PASSAGE 2.0 (Rosenberg and Anderson 2011). Briefly speaking, the Bearing correlogram is a procedure for identifying the maximum angle of directional spatial autocorrelation in a variable of interest:

$$G_{ij} = D_{ij} cos^2(\alpha_{ij} - \theta)$$

Where G_{ij} corresponds to the distance for the variable of interest between points *i,j*, D_{ij} corresponds to the spatial distance between points *i,j*, α_{ij} corresponds to the observed angle between spatial points at *i,j* and θ is the angle of anisotropy in the data. Similar formulations have been proposed by Ramachandran and Rosenberg (2011) and Jay et al. (2013) and applied to the identification of the main axes of genetic differentiation in human populations. For G_{ij}, we used the estimated the Weir and Cockerham's F_{ST} (Weir and Cockerham 1984) between pairs of populations using the genomic tracts of Aboriginal Australian ancestry (the "masked" data) as defined in (S06).

Results

Procrustes analysis between the first two dimensions of an MDS for the Australian data only (S06) indicates a strong statistical correlation (R = 0.59, p value < 0.0005; Figure S13.1.A) with geography. Nevertheless, given the presence of recent admixture with allochthonous populations (S05), we repeated the analysis excluding European and East Asian genomic tracts, as identified in S06, from the Aboriginal Australian individuals with at least a 20% of homozygote Australian ancestry in their genome. The correlation became R = 0.77 (p value < 0.0005; Figure S13.1.B) indicating that genetic diversity of aboriginal Australians is directly linked to their geographic sampling location.

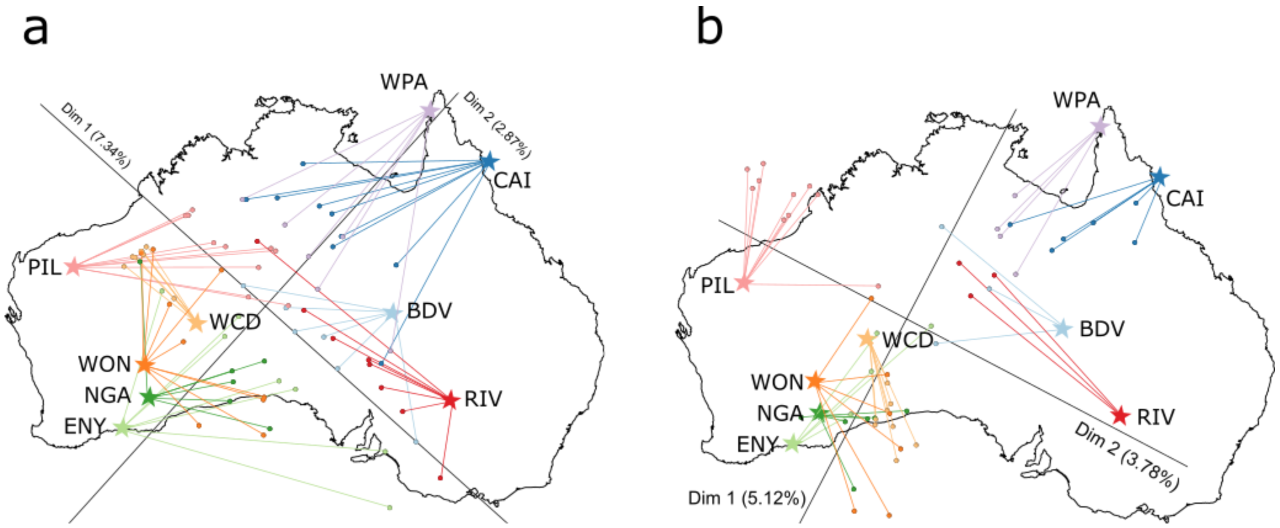


Figure S13.1 First two dimensions of a classical multidimensional scaling (MDS) run on Aboriginal Australians. a) Considering all genetic variants in the genomes of the Aboriginal Australians. b) After masking the tracts assigned to Han Chinese or Brits/Scots (S06). Both MDS plots have been rotated towards the best overlap with geographic sampling locations as defined by Procrustes analysis. In each plot, the arrows indicate the error of the MDS coordinates towards the assigned population sampling geographic coordinates.

The Bearing correlogram on the F_{ST} distance matrix using the masked Aboriginal Australian data identified a maximum gradient angle at 65 degrees compared to the equator, indicating that the strongest genetic differentiation among Aboriginal Australian populations is southwest to northeast.

Inference of geographic position given genetic data

Background

Given the strong correspondence between geography and genetics, we assessed whether the geographic position of the Aboriginal Australian individuals could be inferred from the observed genetic differentiation among individuals.

Methods

In order to determine whether the observed Aboriginal Australian genetic diversity allowed to infer geographic ancestry (Yang et al. 2012), we implemented a simple three layer neural network (Bishop 1996) using the Encog3 JAVA package (Heaton 2011). The input of the neural network were the first two dimensions of the MDS estimated with the pairwise IBS distance matrix between individuals computed on the Aboriginal Australian genomic tracks. The hidden layer consisted of 6 neurons and included a bias neuron. The output layer consisted of the inferred geographic coordinates in Australia. The Sigmoid function was used for neuron activation between the different layers and resilient backward propagation was defined for updating the weights of the neurons. Given the limited number of samples, we adopted a "leave-one-out" strategy. For each individual, we used the remaining individuals to train the neural network with the observed MDS and geographic coordinates. We then used the trained neural network to predict the geographic coordinates of the individual of interest.

As a measure of error between the inferred and the observed geographic coordinates, we computed the geodesic Haversine distance between the two coordinates.

Results

We ran a three layer Neural Network implemented with the Encog3 package, and inferred the geographic coordinates of each individual using the other individuals as the training dataset. The 95% Confidence Intervals (CI) of the Harversine distance between observed and predicted coordinates ranged from 5 km to 1911 km, with the mode at 300 km (Figure S13.2a). The distance between the observed sampled geographic coordinates and those inferred by the neural network depended on which population was considered (Kruskal-Wallis test p-value = 0.04; Figure S13.2b). In particular, it is easier to infer the geographic location of a sample among the WCD, CAI and WON populations compared to BDV or RIV.

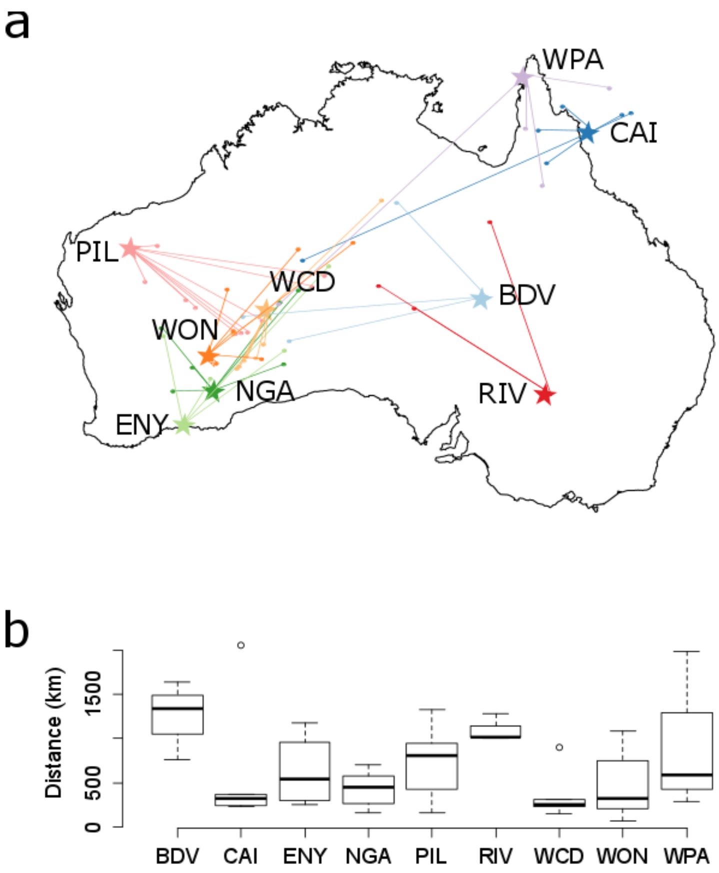


Figure S13.2 a) Inferred geographic coordinates (dots) map and b) boxplot of the distance between inferred and sampled location for each individual by a neural network using all the other individuals as training dataset. The input for the neural network was the first three dimensions of a MDS computed using the IBS distance pairwise on Aboriginal Australian tracks.

Correlation between genetics and linguistics

Background

In S15 we inferred a Bayesian phylogenetic tree based on the linguistic data for the 83 Pama-Nyungan speakers. It has been argued that it is not possible to recover any continent-wide evidence of genetic relationship from linguistic data because either the ancestor is so remote or that there has been extensive cultural diffusion (see Bowern and Atkinson (2012) and citations therein).We tested this hypothesis by comparing the average gene tree based on the masked Aboriginal Australian data and the linguistic tree (see main text) but also by looking into the correlation of the resulting pairwise matrices.

Methods

We generated three distance matrices between pairs of individuals. The first matrix consisted of the identical by state (IBS) distance between pairs of individuals using the masked Aboriginal Australian data. The second matrix was computed on the Harvesine geodesic distance between the population sampling locations of each pair of individuals. The third matrix contained a measure of linguistic distance computed by measuring the total branch lengths between leaves on the linguistic tree. Since the linguistics tree is based on the languages spoken by the parents, every individual can be represented by more than one leaf. Therefore, we computed the distance for one individual as the mean of the distance for its parents.

The association between linguistics and genetics was estimated by means of a Mantel test using the genetic distance and linguistic distance matrices as implemented in PASSAGE 2 (Rosenberg and Anderson 2011). Since both genetics and linguistics tend to depend on geographic patterns, we repeated the analysis controlling by the putative confounder effect of geography by means of a partial Mantel test.

Results

We observed by means of a Mantel test that genetics and linguistics showed a statistically significant association ($r_{GEN,LAN}$ = 0.4, Mantel test two-tail p-value on 9,999 permutations = 0.0001). Nevertheless, since both genetics and linguistics statistically significantly correlate with geography ($r_{GEO,GEN}$ = 0.36, Mantel test two-tail p-value on 9,999 permutations = 0.0001;$r_{GEO,LAN}$ = 0.83, Mantel test two-tail p-value on 9,999 permutations = 0.0001), we wondered whether the association between linguistics and genetics was spurious due to the confounder effect of geography. However, the association between genetics and linguistics remained statistically significant in a partial Mantel test between genetics and linguistics after controlling by the effect of geography ($r_{GEN,GEO,LAN}$= 0.19, Mantel test p-value = 0.0001). Overall, the Mantel test between genetics and geography is in agreement with the results from the Procrustes analysis above (Figure S13.1); moreover, the identified presence of a correlation between genetics and linguistics while controlling for geography suggests that languages are a good proxy for genetic similarity.

Routes of gene flow across the Australian continent

Background

We investigated the major routes of gene flow across the Australian continent; more specifically, we asked whether movement occurred homogeneously, or it mostly followed coastal routes. To do so, we used an approach implemented in Circuitscape (McRae and Beier 2007; McRae et al. 2008), that represents gene flow as electric currents and uses electrical circuit theory (adapted from electrical engineering) to investigate questions in landscape genetics.

Methods

In Circuitscape v4.0, the landscape is represented as a graph where the nodes are physical locations or populations, and edges between pairs of neighbouring nodes carry resistance (a proxy for the difficulty of moving from one location to the next). Gene flow is then modelled as an electric current flowing between two nodes (populations) of interest, moving through the landscape (circuit) according to the resistance of intermediate edges (the connectivity of the landscape). Current will then flow heterogeneously across the landscape depending on the resistance surface; this can be modified by assigning different resistances to edges connecting nodes with certain characteristics, such as habitat. Edges connecting nodes with different characteristics are given intermediate values. In our case, we considered two habitat classes, coastal and inland.

However, due to sea level changes, we have to account for the fact that nodes might have changed their designation through time (and might have not been available in certain periods, when submerged). To approximate the Australian landscape, we represented Australia and neighbouring islands which were connected to Australia during periods of low sea level, as a network of regular, equal-area hexagonal cells, approximately 100km wide (as used in Eriksson et al. (2012)). Using sea level reconstructions, we then selected all cells that were above sea level at any point over the last 45k years as nodes for our circuit. To determine the resistance to be assigned to the edges connected to any given node, we then assessed in how many 1k years window each node was above sea level, and whether it was coastal or inland. We defined the coastal nodes to be all nodes adjacent to a sea hexagon (thus giving a 100 km band). The results did not change qualitatively when we explored other definitions (nodes within two nodes of a sea hexagon). The resistance of an edge to that node was then computed as $(1/\sum_{i=1}^{n} 1/O_i)$, where O_i is the habitat resistance during time window i. This parameter took a value of infinity for sea, 1 Ω for coastal, and x for inland. We explored values of x going from 1 Ω (same resistance as coastal), to 20 Ω (movement through inland much harder than along coastal), in steps of 1 Ω.

For each pair of populations, we then estimated the current that would flow if we connected a 1 volt battery to the two populations (using either as the positive connection), and the resistance to the flow. The resistance encountered between a pair of populations can be seen as the inverse of movement (and gene flow), and thus a predictor of genetic differentiation between them. We then explored which values of x (the increased difficulty of moving inland) gave the best fit between pairwise F_{ST} and pairwise resistance using a Mantel test, using a Spearman rank correlation to deal with possible outliers and non-linearities in the relationship. We estimated the Weir and Cockerham's F_{ST} (Weir and Cockerham 1984) between pairs of Aboriginal Australian groups after masking for Eurasian tracts (S06).

Results

The best fit between genetic differentiation and landscape resistance (Extended Data Figure 7e; r_S=0.694, p<0.001) was obtained when edges connecting inland nodes had 1.7 times more resistance than coastal edges (Extended Data Figure 7e). We can visualise the major routes of flow among our populations by plotting the cumulative current surface obtained by connecting each pair of populations in turn (Extended Data Figure 7g), clearly highlighting the important role of coastal areas. Note that a number of coastal nodes played only a limited role in terms of gene flow as they were submerged for long periods of time.

S13 References

Bishop CM. 1996. Neural Networks for Pattern Recognition. 1 edition. Oxford : New York: Clarendon Press

Bowern C, Atkinson Q. 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. Language 88:817–845.

Eriksson A, Betti L, Friend AD, Lycett SJ, Singarayer JS, Cramon-Taubadel N von, Valdes PJ, Balloux F, Manica A. 2012. Late Pleistocene climate change and the global expansion of anatomically modern humans. Proc. Natl. Acad. Sci. 109:16089–16094.

Heaton J. 2011. Programming Neural Networks with Encog3 in Java. 2 edition. Heaton Research, Inc.

Jay F, Sjödin P, Jakobsson M, Blum MGB. 2013. Anisotropic isolation by distance: the main orientations of human genetic differentiation. Mol. Biol. Evol. 30:513–525.

McRae BH, Beier P. 2007. Circuit theory predicts gene flow in plant and animal populations. Proc. Natl. Acad. Sci. 104:19885–19890.

McRae BH, Dickson BG, Keitt TH, Shah VB. 2008. Using circuit theory to model connectivity in ecology, evolution, and conservation. Ecology 89:2712–2724.

Ramachandran S, Rosenberg NA. 2011. A test of the influence of continental axes of orientation on patterns of human gene flow. Am. J. Phys. Anthropol. 146:515–529.

Rosenberg MS. 2000. The Bearing Correlogram: A New Method of Analyzing Directional Spatial Autocorrelation. Geogr. Anal. 32:267–278.

Rosenberg MS, Anderson CD. 2011. PASSaGE: Pattern Analysis, Spatial Statistics and Geographic Exegesis. Version 2. Methods Ecol. Evol. 2:229–232.

Wang C, Szpiech ZA, Degnan JH, Jakobsson M, Pemberton TJ, Hardy JA, Singleton AB, Rosenberg NA. 2010. Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat. Appl. Genet. Mol. Biol. 9:Article 13.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358–1370.

Yang W-Y, Novembre J, Eskin E, Halperin E. 2012. A model-based approach for analysis of spatial structure in genetic data. Nat. Genet. 44:725–731.

S14 ABC analysis to characterize recent European, East Asian and Papuan gene flow

Oscar Lao

Background

Given the detected amount in the Aboriginal Australian genome of admixture with Europeans, East Asians and Papuans (S05), we conducted an Approximate Bayesian Computation (ABC) analysis (Beaumont et al. 2002) to: 1) characterize the mode of gene flow (discrete versus continuous) that best explains such admixture and 2) estimate the time and the magnitude of gene flow into the Aboriginal Australian groups (or “populations”).

ABC is a statistical technique to recover the posterior distributions of the parameters of interest in situations where it is difficult to express analytically the likelihood function but it is possible to simulate data and compute summary statistics for the models of interest. The ABC implementation requires: 1) a simulation program that generates datasets given a proposed model, 2) definition of prior distributions for each parameter of the model, 3) a set of summary statistics on the simulated datasets that are informative for the model parameters and 4) an algorithm for ascertaining the closest simulations to the observed data as a function of the summary statistics.

Methods

Simulation of tracts of ancestry proportions and demographic models

We used the demographic simulator implemented in Wollstein and Lao (2015), which can simulate the ancestry of the genomic tracts at the genome of one admixed individual given the known ancestral populations. From a given sample of diploid individuals, the simulator goes backward in time. At each generation, each individual carrying a fragment/chromosome from the sampled individuals takes two parents at random from the pool of parents. In the case of population substructure and migration, a parent is taken from the other population(s) with probability *m* and with probability *1-m* takes a parent from the same population. Once the parents have been ascertained, the fragment is broken with a certain probability given by the recombination rate, and the fragments are assigned to the respective parents. In the case that there was a fragment from another individual already present in the ascertained parent, both fragments coalesce into a single one. Finally, fragments from each ancestral population are identified and the average global genomic ancestry is computed (Figure S14.1).

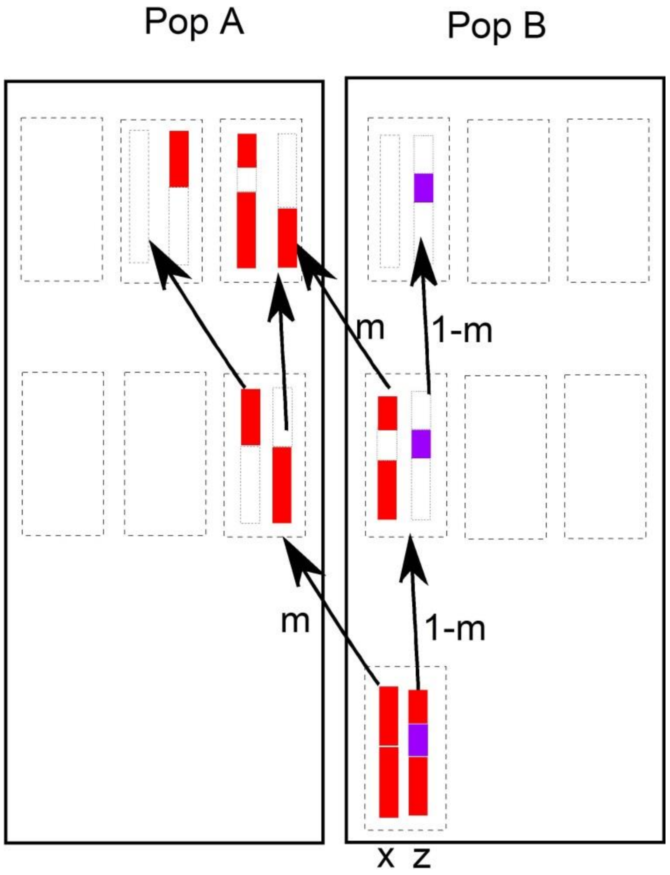


Figure S14.1 Example of how the implemented backward simulator works. Each individual is represented by a dashed box. A diploid individual from population B with two copies from the same chromosome takes two parents from the previous generation. With probability *m*, one of the parents comes from Population A and one of the two chromosomes is assigned at random to the parent individual (in the example, the chromosome X). The chromosome X is the product of a single break by recombination in the parent. Chromosome Z comes from a parent in the same population as the sampled individual and it is broken in three fragments by recombination. Each parent individual takes two parents from the previous generation. In the case of the parent individual of population B, one of the parents comes from population A, whereas another parent comes from population B. Going forward, all the fragments that come from Population A in the sampled individual are painted as red and the remaining fragments coming from Population B as purple.

For each individual, we simulated full genomes with chromosomal size proportional to the real ones as has been done previously (Gravel 2012). Specifically each chromosome was assumed to be of size: 13.65 Mb, 13.15 Mb, 11.20 Mb, 10.65Mb, 10.20 Mb, 9.65 Mb, 9.35 Mb, 8.50 Mb, 8.40 Mb, 8.95 Mb, 7.95 Mb, 8.65 Mb, 6.35 Mb, 5.80 Mb, 6.30 Mb, 6.75 Mb, 6.50 Mb, 5.95 Mb, 5.40 Mb, 5.40 Mb, 3.10 Mb, 3.65 Mb for chromosome 1 to 22 respectively. The recombination rate was set to a constant 1.8 cM/Mb.

Two basic demographic models were considered (Long 1991). In the Hybrid demographic model (HI; Figure S14.2), two ancestral populations (A and B) split long time ago and admix t_{admix} generations ago through a single admixture pulse. The ancestral populations contribute α and 1-α migrants to the ancestral admixed population. In the Continuous Gene Flow model (CGF, Figure S14.2), the ancestral population A starts sending migrants to population B at a rate *m*, and migration continues until the present. For the HI model, we considered that the population C corresponds to Aboriginal Australians, population A corresponds to either Papuans, East Asians or Europeans, and population B corresponds to the non-admixed ancestral Aboriginal Australian. For the CGF model, we considered that population B corresponds to the different Aboriginal Australians populations, and population A corresponds either to Papuans, East Asians or Europeans.

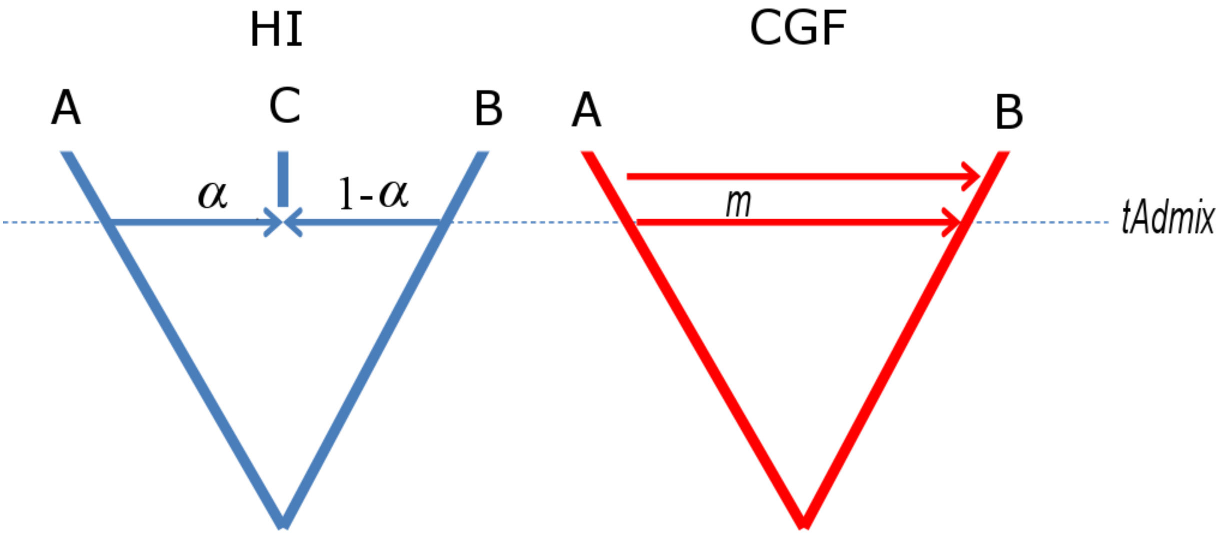


Figure S14.2 Two basic models of population genetic admixture. In the HI model, two populations admix t_{admix} generations ago in a single pulse producing the population C. Population A, respectively B, contributes α, respectively 1-α, migrants to the new admixed C population. In the CGF model, population A starts sending migrants to population B at rate *m* since t_{admix} generations ago.

Prior distributions

For the HI, we considered the following demographic parameters: the effective population size of the hybrid population (Ne_{HI}, twice the number of diploid individuals) that we treated as a nuisance parameter, the time of admixture (t_{adm}) and the proportion α of migrants from population A and 1- α from population B contributing to the ancestral admixed population.
The prior distributions for each parameter were:

Ne_{HI} ~ Uniform(2,000, 10,000)
t_{adm} ~ Uniform(1 generation, 50 generations)
α~ Uniform(0,1)

For the CGF, we considered the time since admixture started (t_{adm}) and the proportion *m* of migrants per generation of the donor population to the acceptor population; the effective population size of the population of acceptor migrants Ne_A was included as a nuisance parameter.
The prior distributions for each parameter were:

Ne_A ~ Uniform(2,000, 10,000)
t_{adm} ~ Uniform(1 generation, 50 generations)
m ~ Uniform(0, $1/t_{adm}$)

Summary statistics and ABC algorithm

Following Verdu and Rosenberg (2011), we used as informative summary statistics the mean (*M*) and variance (*V*) of the ancestry of one of the parental populations in the case of the HI model and the mean and variance of the ancestry of the source population of migrants in the case of the CGF model.

For the model comparison, we combined M and V into a single summary statistic $s = \frac{V}{M}$, which shows a different distribution for each model given the parameter priors we considered (Figure S14.3)

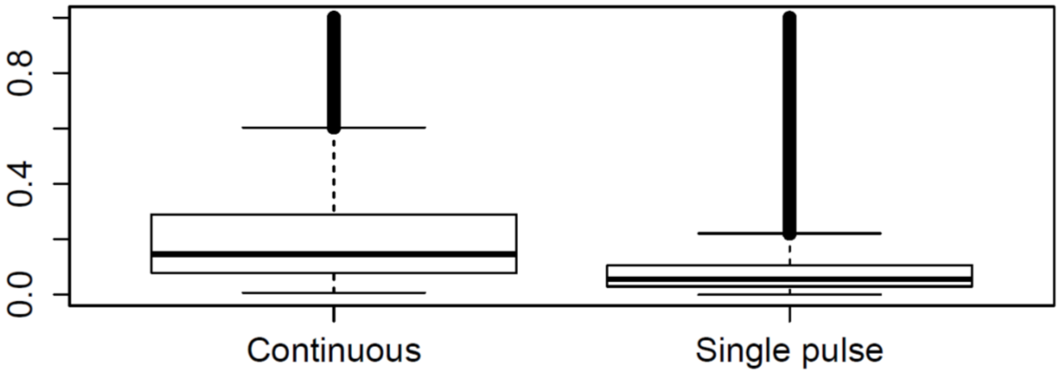


Figure S14.3 Boxplot of V/M ratio computed in 1,000,000 simulations from each model. A tail of V/M > 1 values is observed in both models. CGF simulations (Continuous) tend to show a higher V/M value than HI simulations (Single pulse).

For each model, 1,000,000 simulations were generated, and for each simulation different summary statistics based on observed individual ancestry proportions were computed. In the case of the real data, we used the estimated global ancestry proportions at each Aboriginal Australian individual as estimated by sNMF (S05) to compute the summary statistics. Furthermore, given the observed strong correlation between the percentage of European and Indian ancestry in the Aboriginal Australians (S05), we merged both ancestries.

For parameter estimation, we applied different summary statistics depending on which demographic model was applied and which parameters were considered:

1) In the case of the *HI* model, following Verdu and Rosenberg (2011) we used M for estimating α and $ln\big(M * (1 - M)\big) - \ln(V)$ for estimating t_{adm}.

2) In the case of the *CGF* model, we used V and M for estimating t_{adm}, and M for estimating t_{adm}*m. In order to get the posterior distribution of m, we randomly sampled 1,000 times the posterior distribution of t_{adm} and t_{adm}*m and divided the second value by the first one.

In practice, only the parameters of CGF were computed (see results).

Model comparison was performed for each population using a logistic regression on the closest simulated statistics, treating the two models as categorical dependent variables, as implemented in the *calmod* function (Beaumont 2008). The logistic regression was performed with the 1,000 simulations showing the smallest distance between the simulated and observed summary statistics.

The demographic parameters of the most statistically supported model were estimated using a linear regression on the best 1,000 retained summary statistics and the corresponding parameters, applying a logit transformation in the range of the parameters (Fagundes et al. 2007), as implemented in the *makepd4* function (Beaumont et al. 2002).

In the case of European admixture, we correlated the mean of the posterior distribution for the time of starting admixture with the historical records of first main contact for each Aboriginal population with Europeans (S02). This was not performed for the East Asians and Papuans because of the relative absence of historical records, and due to the fewer Aboriginal Australian groups that show signals of admixture with East Asians and/or Papuans.

Results

We notice that the sensitivity of *V* for estimating t_{admix} decays as t_{admix} increases (Figure S14.4). In practice, we can expect that the currently implemented ABC framework is not going to identify older events than 50 generations in none of the demographic models.

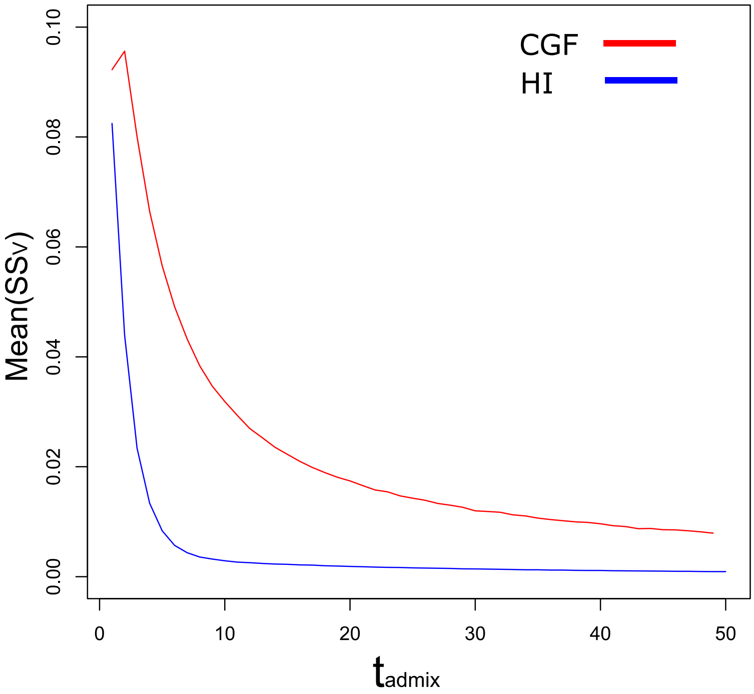


Figure S14.4 Dependence of the time of admixture (t_{admix}) estimated in number of generations and the mean of the *V* summary statistic computed for each simulation and model. As t_{admix} increases the mean value of *V* for each t_{admix} decays rapidly. In practice, V becomes indistinguishable for events that occurred >20 generations in the HI model and >50 generations in the CGF model.

The posterior probability of each model and the posterior distribution of the timing and the intensity of the recent admixture of the Aboriginal Australians with Europeans, East Asians and Papuan populations are reported in Table S14.1. We observed that the most supported model in all the populations and for all three sources of gene flow was the CGF.

Table S14.1 Median and credible interval (between brackets) of the posterior distributions of the estimated time for the beginning of admixture (***t_{adm}***) and migration rate (*m*) with European, East Asian and Papuan populations using the CGF demographic model, which turns out to be the most supported in all the analyses. In the case of East Asian and Papuan admixture, only the populations for which the mean % of observed admixture was > 1 % were considered.

Population	**European ancestry**			**East Asian ancestry**			**Papuan ancestry**		
	$\frac{P(CGF\|D)}{P(HI\|D)}$	t_{adm} *(95 CI)*	*m* *(95CI)*	$\frac{P(CGF\|D)}{P(HI\|D)}$	t_{adm} *(95CI)*	*m* *(95CI)*	$\frac{P(CGF\|D)}{P(HI\|D)}$	t_{adm} *(95CI)*	*m* *(95CI)*
BDV	8.48	15 (3, 44)	0.051 (0.011, 0.148)	-	-	-	-	-	-
CAI	11.48	11 (2, 36)	0.032 (0.004, 0.097)	11.47	20 (5, 46)	0.013 (0.003, 0.04)	7.95	22 (6, 46)	0.006 (0.001, 0.016)
ENY	9.33	25 (5, 47)	0.027 (0.008, 0.092)	-	-	-	-	-	-
NGA	8.49	22 (4, 47)	0.025 (0.006, 0.082)	-	-	-	-	-	-
PIL	9.06	15 (4, 43)	0.012 (0.002, 0.033)	8.05	12 (3, 43)	0.006 (0.001, 0.02)	9.49	29 (11, 48)	0.003 (0.001, 0.008)
RIV	8.68	22 (4, 46)	0.041 (0.013, 0.148)	-	-	-	-	-	-
WCD	6.95	8 (2, 38)	0.005 (0.0003, 0.014)	-	-	-	-	-	-
WON	10.36	11 (3, 37)	0.036 (0.006, 0.105)	-	-	-	-	-	-
WPA	10.45	14 (2, 44)	0.035 (0.005, 0.132)	12.69	10 (2, 44)	0.022 (0.002, 0.064)	12.30	15 (3, 46)	0.02 (0.003, 0.064)

In the case of the European ancestry component, the estimated posterior distributions of the timing of admixture differed among the different Aboriginal Australian groups. Furthermore, the mean of the posterior distribution for the time of admixture estimated at each group statistically significantly correlates with the historical records of the main time of contact with Europeans (Figure S14.5, R^2:0.64, p-value: 0.005).

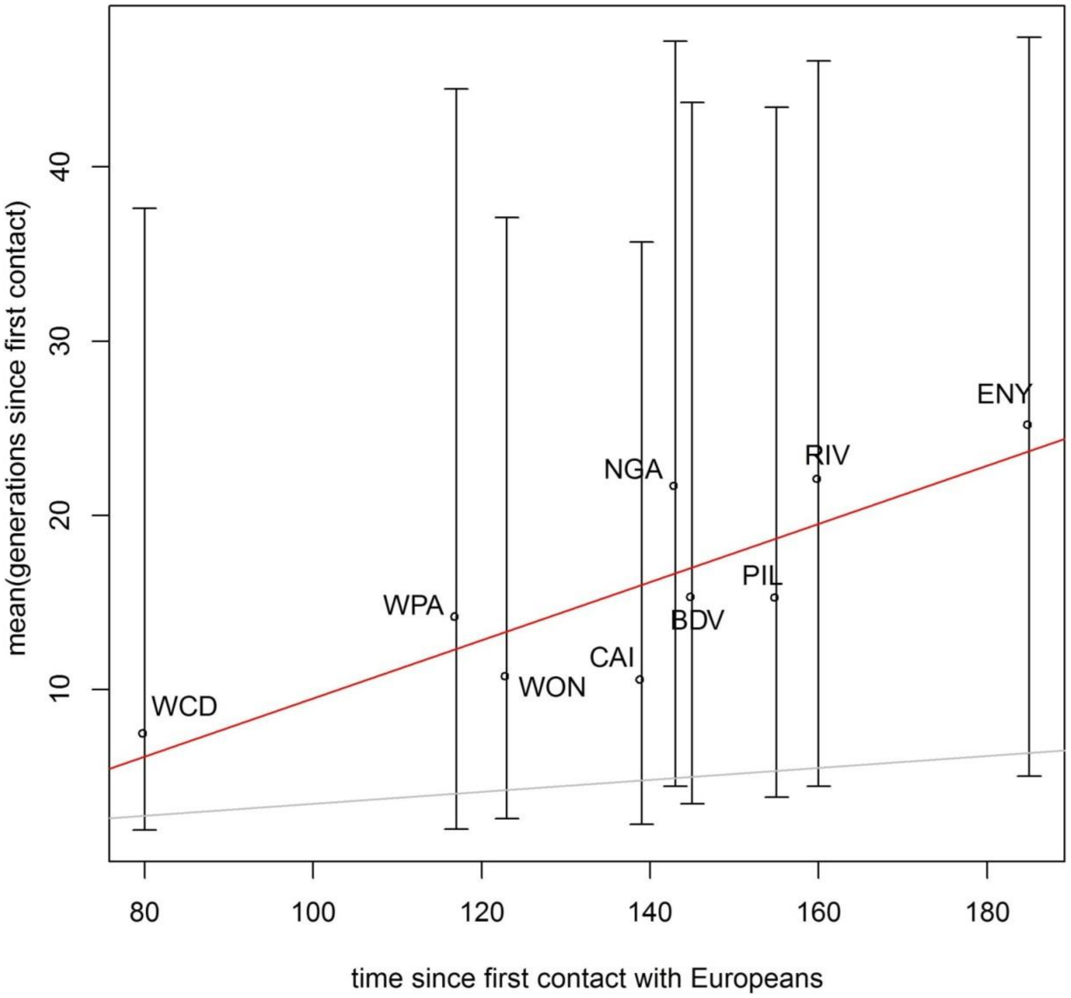


Figure S14.5 Correlation between the time (years) since the beginning of the European contact, as historically recorded, and the mean of the generations of first contact as inferred by our ABC approach. 95% Credible Intervals (CI) for each population are included in the plot. Red line: regression line between the time of first contact with Europeans and the number of generations according to ABC. Gray line: expected regression line with a generation time of 29 years (Fenner 2005).

We observe that to rescale the computed generation time to match the observed historical times requires a much younger generation time than the reference generation time of 29 years (Fenner 2005) (Figure S14.5); nevertheless, the 95% CI of the posterior distributions contains the historical time in all the populations.

The posterior distributions of the amount of European migration rate per generation suggest that some Aboriginal Australian populations were in closer contact with Europeans than others. In particular, WCD and PIL show the smallest migration rates per generation of all the Aboriginal Australian populations, whereas BDV and RIV showed respectively migration rates of 5% and 4% per generation according to the CGF model of continuous admixture.

ABC analysis on the timing of East Asian ancestry in the CAI, PIL and WPA suggested a starting admixture of Aboriginal Australians with East Asians ~10 generations ago in WPA and PIL and ~20 generations ago in CAI. Nevertheless, as observed in the case of European admixture, the credible intervals are quite broad in all the three populations. Similarly, recent Papuan admixture for the same three populations is estimated between 15 and 30 generations ago, with admixture rates per generation ranging from ~0.03% in PIL to 2% in WPA.

S14 References

Beaumont M. 2008. Joint determination of tree topology and population history. In: Simulation, Genetics, and Human Prehistory. McDonald Institute for Archaeological Research: Cambridge. p. 135–154.

Beaumont MA, Zhang W, Balding DJ. 2002. Approximate Bayesian computation in population genetics. Genetics 162:2025–2035.

Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, Bonatto SL, Excoffier L. 2007. Statistical evaluation of alternative models of human evolution. Proc. Natl. Acad. Sci. U. S. A. 104:17614–17619.

Fenner JN. 2005. Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am. J. Phys. Anthropol. 128:415–423.

Gravel S. 2012. Population genetics models of local ancestry. Genetics 191:607–619.

Long JC. 1991. The genetic structure of admixed populations. Genetics 127:417–428.

Verdu P, Rosenberg NA. 2011. A general mechanistic model for admixture histories of hybrid populations. Genetics 189:1413–1426.

Wollstein A, Lao O. 2015. Detecting individual ancestry in the human genome. Investig. Genet. 6:7.

S15 Computational phylogenetics: Pama-Nyungan languages

Claire Bowern

Language Background

Languages change in ways that are broadly similar to genetic mutations. Both linguists and geneticists make use of homologous traits to infer change through time, and to reconstruct ancestral states (cf. Gray et al., 2011). Australian languages at the time of European contact were diverse, with 400 languages belonging to 27 phylic families. In comparison, Europe has only four families of equivalent depth (Indo-European, Finno-Ugric, Turkic, and the Basque isolate). Pama-Nyungan is the largest of these families both in terms of number of languages and in geographical extent, covering 90% of the mainland. All but one of the Aboriginal Australian donors in this project belong to groups whose languages are members of the Pama-Nyungan family.

There are two main linguistic and archaeological scenarios for the spread of Pama-Nyungan. One (e.g., Dixon, 1997) is that Pama-Nyungan is a relic of the initial immigration of Aboriginal Australians to the continent. The second is that Pama-Nyungan is a Holocene expansion from within Australia (Bowern and Koch, 2004; Evans and McConvell, 1997, 1997; McConvell, 1996; Sutton, P., 1990). Hale (1964) places the origin of Proto-Pama-Nyungan in modern Queensland, near the base of the Gulf of Carpentaria. Bowern and Atkinson (2012) concur with this, citing evidence such as the Queensland/Northern territory border region being the joining point of three primary divisions within Pama-Nyungan and a location close to the earliest reconstructed bifurcation of the Western Pama-Nyungan division; this is in line with a principle in linguistic geography of the area of greatest diversity being the most likely point of dispersal. Authors do not explicitly state their hypotheses as to the means by which the Pama-Nyungan language spread across the country. Bellwood (2013) (among others) notes the enigma of large hunter-gatherer families such as Pama-Nyungan; enigmatic given the hypotheses that it was the development of agriculture that facilitated the population increases required to drive the major expansions of several other large language families. Thus Pama-Nyungan could either be a spread into uninhabited territory (as in Dixon's model), a spread where Pama-Nyungan speakers replace earlier inhabitants due to some technological or cultural advantage (as in Evans and Jones (1997)), or some combination of the two.

Data

For each individual, we collected ethnographic data about their genealogies, including (whenever possible) the language or cultural group of parents and grandparents. Where such information was not available, we assigned the language according to the area of country with which the individual identifies (S03). Note that we removed the one individual from the analysis (CAI10) who did not speak a Pama-Nyungan language. The results can be found on Extended Data Table 1.

Basic vocabulary cognate data for most of those languages were previously collected in Bowern and Atkinson (2012), using published and unpublished data from languages across the country. Bowern coded additional languages subsequently.

Method

The method relies on the identification of cognate words (see e.g. Hock and Joseph (1996)). Cognates are the linguistic equivalent of homologous traits: that is, words which are presumed to derive from a common ancestor. We use Bayesian phylogenetic inference as implemented in BEAST (Drummond and Rambaut, 2007) to model language evolution as the gain (0 → 1) and loss (1→0) of cognates along the branches of a language family tree or 'phylogeny'. We assumed a stochastic Dollo model with a relaxed clock (see Nicholls and Gray (2006) and Bowern and Atkinson (2012)) under which cognates can be gained only once but lost multiple times. We assumed a yule prior on branch lengths in the tree. We used BEAST (Drummond and Rambaut, 2007) to run multiple independent MCMC chains for between fifty and one hundred million iterations each, sampled every 10,000 iterations, with the first ten million iterations discarded as burn-in. We used the Tracer component of BEAST to examine the post-burn-in likelihoods and other parameters of interest across the Markov chain. This revealed that runs had reached convergence by this time and effective sample sizes for all parameters were above 2,000.

Results

As noted in the main text, the genetic and linguistic trees show a surprising degree of congruity (Extended Data Figure 7c,d). However, the genetic splits are dated much earlier than the mid-Holocene (the time period associated with Pama-Nyungan spread by many scholars). While the genetic evidence perhaps most parsimonious fits some variation of Dixon's (1997) scenario of an early dispersal of Pama-Nyungan, Dixon's model is incompatible with all current research into Pama-Nyungan, particularly Bowern and Atkinson (2012), which demonstrates that Dixon's convergence-through-contact model is not supported by the linguistic facts.

There are several ways in which we might reconcile the early dates for genetic split with a more recent dispersal of Pama-Nyungan. The greater within-group diversity of mtDNA, as discussed in the main text (S12), may indicate more male than female migration between groups. If this migration included Pama-Nyungan-language speaking men marrying into non-Pama-Nyungan-language speaking groups, patrilineal linguistic affiliation would cause Pama-Nyungan languages to spread.

Diffusion is another way in which languages may spread. In this case, a group of speakers of one language shifts to another language, as in the case of subjects of the Roman Empire shifting to Latin from Celtic and Germanic languages and Basque (amongst others). Shift is clearly at work in some parts of Pama-Nyungan. An example can be seen from the CAI and WPA genetic populations in our sample (Extended Data Figure 7d). CAI/WPA correspond in the linguistic tree to the Pama-Maric subgroup of Pama-Nyungan. The WPA population includes individuals who speak languages of the Pama-Maric subgroup of Pama-Nyungan, except those Pama-Maric languages of the Cairns rainforest region. CAI is thus monophyletic, but WPA is not. CAI corresponds to the small statured populations of the Cairns rainforest region (Mulvaney and Kamminga, 1999, p. 154). The language tree suggests that the CAI population may have shifted to Pama-Nyungan.

However, we reject cultural diffusion/language shift as the sole mechanism for language spread here, on two counts. First a language shift is culturally costly. Speakers of specific languages do not shift unless they have a strong economic or cultural reason to do so. Many language groups in Australia show stable multigenerational multilingualism (cf. Elwell, V.M.R., 1982) rather than language shift. Secondly, extensive language contact and cultural diffusion tends to leave a signature in linguistic phylogenetic trees, in the form of conflicting evidence for subgrouping and absence of clear binary splits. Conversely, migration can often be "read off" language trees (compare, for example, the migration of Austronesian speakers across insular Southeast Asia and the Pacific in Gray et al., 2009). A similar set of migrations can be seen in the southward branching of Western Pama-Nyungan (in blue in Figure S15.1) and the northward branching northward up the eastern coast of Australia (in green). Note, in comparison, the relative lack of clear branching correlated with geography in the Pama-Maric groups (in purple in figure S15.1), one area of claimed language shift/diffusion.

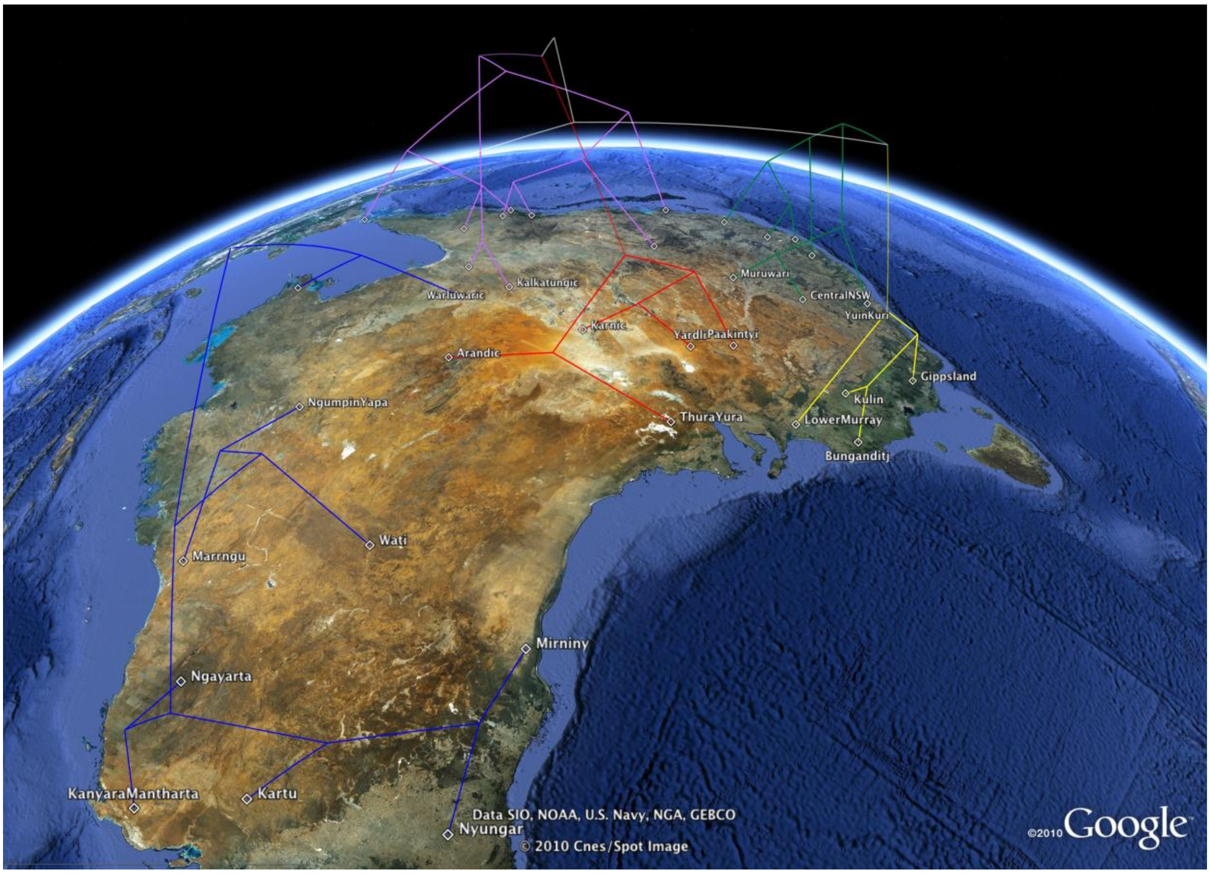


Figure S15.1 Consensus tree from Bowern and Atkinson (2012), with subgroup centroid coordinates projected onto a satellite image of Australia using Mesquite (Maddison, W. P. and D.R. Maddison. 2015. Mesquite: a modular system for evolutionary analysis. Version 3.04 *http://mesquiteproject.org*). Branches are coloured by major subgroup division.

Of course, language histories are complex, and a single mechanistic explanation does not apply to all areas worldwide. Both migration and admixture are at work in the dispersal of Pama-Nyungan. We observe a genetic signal which could be consistent with the dates associated with the Pama-Nyungan dispersal (4~7kya), but we infer in this study that the population ancestral to Pama-Nyungan groups was structured much earlier. This implies a history where at least some language shift has taken place. Further work with samples from non-Pama-Nyungan populations would presumably help clarifying the relative roles of migration and admixture within the history of the Australian continent.

S15 References

Bellwood, P., 2013. First Migrants: Ancient Migration in Global Perspective, 1 edition. ed. Wiley-Blackwell, Chichester, West Sussex, UK ; Malden, MA.

Bowern, C., Atkinson, Q., 2012. Computational phylogenetics and the internal structure of Pama-Nyungan. Language 88, 817–845. doi:10.1353/lan.2012.0081

Bowern, C., Koch, H.J., 2004. Australian Languages: Classification and the Comparative Method. John Benjamins Publishing.

Dixon, R.M.W., 1997. The Rise and Fall of Languages. Cambridge University Press.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214. doi:10.1186/1471-2148-7-214

Elwell, V.M.R., 1982. Some social factors affecting multilingualism among Aboriginal Australians: a case study of Maningrida. Int. J. Soc. Lang. 1982, 83–103.

Evans, N., Jones, R., 1997. The cradle of the Pama-Nyungans: Archaeological and linguistic speculations, in: Archaeology and Linguistics: Aboriginal Australia in Global Perspective. Oxford University Press Australia, Melbourne.

Evans, N., McConvell, P., 1997. The enigma of Pama-Nyungan expansion in Australia. Archaeol. Lang. II 174–191.

Gray, R.D., Atkinson, Q.D., Greenhill, S.J., 2011. Language evolution and human history: what a difference a date makes. Philos. Trans. R. Soc. B Biol. Sci. 366, 1090–1100. doi:10.1098/rstb.2010.0378

Gray, R.D., Drummond, A.J., Greenhill, S.J., 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323, 479–483. doi:10.1126/science.1166858

Hale, K., 1964. Classification of Northen Paman Languages, Cape York Peninsula, Australia: A Research Report. Ocean. Linguist. 3, 248–265. doi:10.2307/3622881

Hock, H.H., Joseph, B.D., 1996. Language History, Language Change, and Language Relationship: An Introduction to Historical and Comparative Linguistics. Mouton, Berlin.

McConvell, P., 1996. Backtracking to Babel: the chronology of Pama-Nyungan expansion in Australia. Archaeol. Ocean. 31, 125–144.

Mulvaney, J., Kamminga, J., 1999. Prehistory of Australia. Allen & Unwin, St Leonards, NSW.

Nicholls, G.K., Gray, R.D., 2006. Quantifying uncertainty in a stochastic model of vocabulary evolution. Phylogenetic Methods Prehistory Lang. Forster P Renfrew C 161–171.

Sutton, P., 1990. The pulsating heart: large-scale cultural and demographic processes in Aboriginal Australia., in: Meehan, B., White, N. (Eds.), Hunter–gatherer Demography. University of Sidney, Sidney, Australia.

S16 Scan for positive selection

Jacob E Crawford, Jade Y Cheng, Fernando Racimo, Rasmus Nielsen

Selection scan in Aboriginal Australians

Background and data

To identify genomic regions putatively involved in genetic adaptation to novel selection pressures in Australia, we compared a panel of genomes from 63 Aboriginal Australian individuals to Europeans and East Asians. We included only the Aboriginal Australian individuals determined to be unrelated and admixed with only European and East Asian ancestry (S05; Excluded individuals: BDV05, BDV06, BDV07, BCF09, CAI05, CAI10, ENY03, NGA01, PIL08, PIL12, WCD06, WCD07, WCD08, WCD10, WCD11, WON01, WON09, WON11, WPA03, WPA05, WPA06). We merged this data with population samples (30 genomes from each population) randomly selected from the British (GBR), Han Chinese (CHB) and Yoruban (YRI) 1000 Genomes project populations (Phase 3, *http://1000Genomes.org*).

Since admixture from other populations could obscure signals of natural selection in the Aboriginal Australians, we compared admixture-corrected allele frequencies in the Aboriginal Australians with allele-frequencies in GBR, CHB and YRI. In other words, we identify loci where allele frequencies in Aboriginal Australians differed from the reference populations. We also searched for genomic regions that differed between the two ancestry components within Australia associated with different ecological regions. To compute admixture corrected allele frequencies, we estimated ancestry proportions based (S05) on the merged dataset assuming either four or five ancestral components (*K*=4 or 5). We only included variable sites with no missing data and a minimum allele frequency of 0.05 across the entire merged panel. In total, we analyzed 5,526,711 variable sites across the genome.

Method

We used Population Branch Statistics (*PBS*) (Yi et al. 2010) to search for genomic regions where allele frequencies were exceptionally differentiated in the Aboriginal Australians relative to frequencies in the Han Chinese and the British. We calculated the *PBS* statistic in blocks of 10 SNPs across the 22 autosomal chromosomes. To filter genomic regions that are highly differentiated between all populations, we defined a new normalized version of the standard *PBS* statistic as:

$$PBS_{n1} = \frac{PBS_1}{1 + PBS_1 + PBS_2 + PBS_3}$$

where PBS_1 indicates *PBS* calculated with Aboriginal Australians as the focal population, PBS_2 indicates *PBS* calculated with the Han Chinese as the focal population, and PBS_3 indicates *PBS* calculated with the British as the focal population. For the comparison among Aboriginal Australian subgroups, we calculated the three *PBS* statistics using the two Australian components and the Han Chinese component. Windowed PBS_{n1} values were sorted into a ranked list and considered putative indications of positive selection if they satisfied two criteria: 1) At least five of the 10 SNPs fell within the top 0.1% tail of SNP-wise PBS_{n1} values and 2) No other window had a higher PBS_{n1} value within one Mb on each side. We present the top hits from this analysis as candidate loci and do not attempt to assign p-values to each locus.

Results

The desert conditions in Australia pose a number of physiological challenges related to harsh arid landscapes such as limited water supplies and extreme oscillation of temperatures, and Aboriginal Australian genomes may possibly harbor adaptation to this environment. The top 10 regions identified in this scan are shown in Extended Data Table 2 and Table S16.1.

While most of the peaks of the PBS statistic covered regions too broad to robustly pinpoint the underlying target gene, we identified one peak that was sufficiently narrow with strong differentiation immediately outside of the exonic regions of the protein coding gene *SLC2A12* (top SNP: chr6:134,391,056; rs4896021; $F_{AUS\text{-}CHB}$ = 0.9603, $F_{AUS\text{-}CEU}$ = 0.9461, $F_{CHB\text{-}CEU}$ = -0.0132; PBS_{n1} = 0.7574). The protein product of this gene, also known as *GLUT-12*, is a member of a family of facilitative diffusion glucose transporters (Rogers et al. 2002). A recent GWAS study in African Americans found a significant association between polymorphisms near *SLC2A12* and serum urate levels (Tin et al. 2011). The pathophysiology of dehydration includes hyperuricemia, or elevated serum urate levels, so these results are suggestive of a locus that may be involved in an adaptive tolerance to dehydration in Australians.

An additional well-known hypothesis for adaptation to extreme desert temperatures posits that Aboriginal Australian populations have several physiological adaptations to tolerate cold nighttime temperatures, including reduced shivering and low body heat conductivity (Scholander et al. 1958; Leppäluoto and Hassi 1991). The thyroid hormone thyroxine regulates metabolism in mammals and is thought to be involved in tolerance to extreme temperatures in Aboriginal Australians (Qi et al. 2014). A genomic scan for genetic associations with variation in thyroid levels had identified a number of significant loci, including a polymorphism near the gene *NETO1*, which codes for Neuropilin And Tolloid-Like Protein 1 (Porcu et al. 2013). Inspection of the top most differentiated windows reveals a peak of high PBS_{n1} values near *NETO1* (top SNP: chr18:70,019,066; rs12455116; $F_{AUS\text{-}CHB}$ = 0.9155, $F_{AUS\text{-}CEU}$ = 0.8669, $F_{CHB\text{-}CEU}$ = 0.0013; PBS_{n1} = 0.6914), consistent with positive selection on a genetic variant at this locus in Aboriginal Australians. Although we do not know the exact causative polymorphism at this locus, it is tempting to speculate that a genetic variant changing expression patterns of *NETO1* and other thyroid-related loci may underlie tolerance of cold nighttime temperatures and high daytime temperatures. Further functional studies would be needed to test this hypothesis.

Various population genetic analyses suggest structure of Aboriginal Australian genetic diversity along a south-west to north-east gradient (Extended Data Figure 2a, Extended Data Figure 7). This gradient correlates with environmental covariates, in that the northeastern groups inhabit generally wet tropical environments, whereas the southwestern groups tend to inhabit dry desert environments. We therefore applied the PBS statistic to two components (*sw* and *ne*) using CHB as an outgroup. The top 10 regions identified for both the *sw* and the *ne* groups are shown in Extended Data Table 2 and Table S16.1. Interestingly, the most extreme signal for the *sw* group lies in close proximity to the gene *KCNJ2* (top SNP:chr17: 68,190,552; rs35167900; $F_{AUSsw\text{-}AUSne}$ = 0.5486, $F_{AUSsw\text{-}CHB}$ = 0.8438, $F_{AUSne\text{-}CHB}$ = 0.1468; PBS_{n1} = 0.5182). *KCNJ2* encodes member 2 of inward-rectifier subfamily J of potassium channel proteins, and polymorphisms near this gene have been associated in a genome-wide association study of thyrotoxic periodic paralysis (Cheung et al. 2012). This disease results from complications related to hyperthyroidism, providing additional support for the thyroid hormone system as a target of desert-related natural selection in Aboriginal Australians.

Table S16.1 Top 10 peaks of differentiation from genome scans of all Aboriginal Australians combined (All) and two Aboriginal Australians subgroups living in different ecological regions in Australia.

Focal Pop	Nearby Gene	rsID	Function of gene product[a]
All	*TMEM86B*	rs734517	Catalyzes the degradation of lysoplasmalogen. Modulates cell membrane proteins.[1]
All	*LRRC52*	rs4147601	Modulates voltage of potassium ion channels. Expressed in testis.[2]
All	*MACROD2*	rs175279	Involved in deacetylase activity.[3,4] Possibly (but not conclusively) causative of Kabuki syndrome.[5]
All	*JRKL*	rs72959058	Homologue to “jerky” gene in mouse.[6]
All	*SPATA20*	rs73338243	Spermatid protein.[7]
All	*NAA60*	rs73503305	Histone acetyltransferase required for nucleosome assembly and chromosome segregation during anaphase.[8] Human-specific imprinted gene[9].
All	*CBLN2*	rs12455116	*CBLN2*: cerebellum-specific protein involved in various signaling pathways.[13,14] Possibly associated with pulmonary arterial hypertension.[15]
	NETO1		*NETO1*: brain-specific transmembrane protein involved in the regulation of neuronal circuitry.[10,11] Associated with thyroid function.[12]
All	*SLC2A12*	rs4896021	Catalyzes sugar absorption[16]. Involved in the pathogenesis of diabetes.[17] Associated with serum urate levels.[18]
All	*LOC101927657*	rs145200081	Unknown (ncRNA).
All	*LOC102724612*	rs113341339	Unknown (ncRNA).
NE	*ZBTB20*	rs9289004	Transcriptional repressor[19] associated with Primrose syndrome.[20]
NE	*ANXA10*	rs2176513	Calcium-dependent phospholipid-binding annexin.[21]
NE	*TRPC3*	rs4502701	Non-selective cation channel[22], associated with spinocerebellar ataxia.[23]
NE	*HS3ST1*	rs7665516	Regulates rate of generation of anticoagulant heparan sulfate proteoglycan.[24]
NE	*MIR548C*	rs2620721	Unknown (microRNA).
NE	*STARD13*	rs7318080	Involved in cell proliferation[25] and fibroblast morphology.[26]
NE	*AKAP11*	rs7319267	Directs protein kinase A activity[27] and is involved in cAMP messenger signaling.[28]
NE	*AGMO*	rs35557899	Catalyzes the cleavage of O-alkyl bonds of ether lipids.[29]
NE	*RUNX1T1*	rs11776341	Involved in transcriptional repression[30]. A translocation involving this gene is associated with acute myeloid leukemia.[31,32]
NE	*FHAD1*	rs2473358	Unknown.
SW	*KCNJ2*	rs35167900	Potassium channel[33], associated with familial atrial fibrillation[34] and periodic paralysis.[35-37]
SW	*TACC2*	rs10159998	Belongs to a family of proteins that interact with the centrosome and microtubules, and that are implicated in cancer.[38]
SW	*LOC101928708*	rs4843556	Unknown (ncRNA).
SW	*C16orf82*	rs72782349	Unknown.
SW	*LOC100507391*	rs56379930	Unknown (ncRNA).

SW	*HAUS4*	rs2008951	A component of a microtubule-binding complex that plays a role in the generation of microtubules in the mitotic spindle.[39,40]
SW	*KNG1*	rs5029990	During the inflammatory response, it is involved in vasodilation, coagulation, enhanced capillary permeability and pain induction.[41-43]
SW	*MYDGF*	rs66891175	Unknown.
SW	*MSMP*	rs1951432	May be involved in the tumorigenesis of prostate cancer.[44]
SW	*VAV2*	rs2519771	Member of an oncogene family.[45] Involved in T-cell receptor signaling.[46]

[a]References for function of gene product:

1 Wu, L. C. *et al.* Purification, identification, and cloning of lysoplasmalogenase, the enzyme that catalyzes hydrolysis of the vinyl ether bond of lysoplasmalogen. *J Biol Chem***286**, 24916-24930, doi:10.1074/jbc.M111.247163 (2011).

2 Yan, J. & Aldrich, R. W. BK potassium channel modulation by leucine-rich repeat-containing proteins. *Proc Natl Acad Sci U S A***109**, 7917-7922, doi:10.1073/pnas.1205435109 (2012).

3 Chen, D. *et al.* Identification of macrodomain proteins as novel O-acetyl-ADP-ribose deacetylases. *J Biol Chem***286**, 13261-13271, doi:10.1074/jbc.M110.206771 (2011).

4 Jankevicius, G. *et al.* A family of macrodomain proteins reverses cellular mono-ADP-ribosylation. *Nat Struct Mol Biol***20**, 508-514, doi:10.1038/nsmb.2523 (2013).

5 Maas, N. M. *et al.* The C20orf133 gene is disrupted in a patient with Kabuki syndrome. *J Med Genet***44**, 562-569, doi:10.1136/jmg.2007.049510 (2007).

6 Zeng, Z. *et al.* Cloning, mapping, and tissue distribution of a human homologue of the mouse jerky gene product. *Biochem Biophys Res Commun***236**, 389-395, doi:10.1006/bbrc.1997.6935 (1997).

7 Shi, H. J. *et al.* Cloning and characterization of rat spermatid protein SSP411: a thioredoxin-like protein. *J Androl***25**, 479-493 (2004).

8 Van Damme, P. *et al.* NatF contributes to an evolutionary shift in protein N-terminal acetylation and is important for normal chromosome segregation. *PLoS Genet***7**, e1002169, doi:10.1371/journal.pgen.1002169 (2011).

9 Nakabayashi, K. *et al.* Methylation screening of reciprocal genome-wide UPDs identifies novel human-specific imprinted genes. *Hum Mol Genet***20**, 3188-3197, doi:10.1093/hmg/ddr224 (2011).

10 Stöhr, H., Berger, C., Fröhlich, S. & Weber, B. H. A novel gene encoding a putative transmembrane protein with two extracellular CUB domains and a low-density lipoprotein class A module: isolation of alternatively spliced isoforms in retina and brain. *Gene***286**, 223-231 (2002).

11 Michishita, M. *et al.* A novel gene, Btcl1, encoding CUB and LDLa domains is expressed in restricted areas of mouse brain. *Biochem Biophys Res Commun***306**, 680-686 (2003).

12 Porcu, E. *et al.* A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function. *PLoS Genet***9**, e1003266, doi:10.1371/journal.pgen.1003266 (2013).

13 Kavety, B., Jenkins, N. A., Fletcher, C. F., Copeland, N. G. & Morgan, J. I. Genomic structure and mapping of precerebellin and a precerebellin-related gene. *Brain Res Mol Brain Res***27**, 152-156 (1994).

14 Wei, P. *et al.* The Cbln family of proteins interact with multiple signaling pathways. *J Neurochem***121**, 717-729, doi:10.1111/j.1471-4159.2012.07648.x (2012).

15 Germain, M. *et al.* Genome-wide association analysis identifies a susceptibility locus for pulmonary arterial hypertension. *Nat Genet***45**, 518-521, doi:10.1038/ng.2581 (2013).

16 Rogers, S. *et al.* Identification of a novel glucose transporter-like protein-GLUT-12. *Am J Physiol Endocrinol Metab***282**, E733-738 (2002).

17 Linden, K. C. *et al.* Renal expression and localization of the facilitative glucose transporters GLUT1 and GLUT12 in animal models of hypertension and diabetic nephropathy. *Am J Physiol Renal Physiol***290**, F205-213, doi:10.1152/ajprenal.00237.2004 (2006).

18 Tin, A. *et al.* Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele. *Hum Mol Genet***20**, 4056-4068, doi:10.1093/hmg/ddr307 (2011).

19 Sutherland, A. P. *et al.* Zinc finger protein Zbtb20 is essential for postnatal survival and glucose homeostasis. *Mol Cell Biol***29**, 2804-2815, doi:10.1128/MCB.01667-08 (2009).
20 Cordeddu, V. *et al.* Mutations in ZBTB20 cause Primrose syndrome. *Nat Genet***46**, 815-817, doi:10.1038/ng.3035 (2014).
21 Morgan, R. O. *et al.* Novel human and mouse annexin A10 are linked to the genome duplications during early chordate evolution. *Genomics***60**, 40-49, doi:10.1006/geno.1999.5895 (1999).
22 Becker, E. B. *et al.* Candidate screening of the TRPC3 gene in cerebellar ataxia. *Cerebellum***10**, 296-299, doi:10.1007/s12311-011-0253-6 (2011).
23 Fogel, B. L., Hanson, S. M. & Becker, E. B. Do mutations in the murine ataxia gene TRPC3 cause cerebellar ataxia in humans? *Mov Disord***30**, 284-286, doi:10.1002/mds.26096 (2015).
24 Shworak, N. W. *et al.* Molecular cloning and expression of mouse and human cDNAs encoding heparan sulfate D-glucosaminyl 3-O-sulfotransferase. *J Biol Chem***272**, 28008-28019 (1997).
25 Ching, Y. P. *et al.* Deleted in liver cancer (DLC) 2 encodes a RhoGAP protein with growth suppressor function and is underexpressed in hepatocellular carcinoma. *J Biol Chem***278**, 10824-10830, doi:10.1074/jbc.M208310200 (2003).
26 Leung, T. H. *et al.* Deleted in liver cancer 2 (DLC2) suppresses cell transformation by means of inhibition of RhoA activity. *Proc Natl Acad Sci U S A***102**, 15207-15212, doi:10.1073/pnas.0504501102 (2005).
27 Lester, L. B., Coghlan, V. M., Nauert, B. & Scott, J. D. Cloning and characterization of a novel A-kinase anchoring protein. AKAP 220, association with testicular peroxisomes. *J Biol Chem***271**, 9460-9465 (1996).
28 Logue, J. S. *et al.* AKAP220 protein organizes signaling elements that impact cell migration. *J Biol Chem***286**, 39269-39281, doi:10.1074/jbc.M111.277756 (2011).
29 Watschinger, K. *et al.* Identification of the gene encoding alkylglycerol monooxygenase defines a third class of tetrahydrobiopterin-dependent enzymes. *Proc Natl Acad Sci U S A***107**, 13672-13677, doi:10.1073/pnas.1002404107 (2010).
30 Kumar, R. *et al.* CBFA2T3-ZNF652 corepressor complex regulates transcription of the E-box gene HEB. *J Biol Chem***283**, 19026-19038, doi:10.1074/jbc.M709136200 (2008).
31 Erickson, P. *et al.* Identification of breakpoints in t(8;21) acute myelogenous leukemia and isolation of a fusion transcript, AML1/ETO, with similarity to Drosophila segmentation gene, runt. *Blood***80**, 1825-1831 (1992).
32 Miyoshi, H. *et al.* The t(8;21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript. *EMBO J***12**, 2715-2721 (1993).
33 Raab-Graham, K. F., Radeke, C. M. & Vandenberg, C. A. Molecular cloning and expression of a human heart inward rectifier potassium channel. *Neuroreport***5**, 2501-2505 (1994).
34 Xia, M. *et al.* A Kir2.1 gain-of-function mutation underlies familial atrial fibrillation. *Biochem Biophys Res Commun***332**, 1012-1019, doi:10.1016/j.bbrc.2005.05.054 (2005).
35 Tristani-Firouzi, M. *et al.* Functional and clinical characterization of KCNJ2 mutations associated with LQT7 (Andersen syndrome). *J Clin Invest***110**, 381-388, doi:10.1172/JCI15183 (2002).
36 Plaster, N. M. *et al.* Mutations in Kir2.1 cause the developmental and episodic electrical phenotypes of Andersen's syndrome. *Cell***105**, 511-519 (2001).
37 Cheung, C. L. *et al.* Genome-wide association study identifies a susceptibility locus for thyrotoxic periodic paralysis at 17q24.3. *Nat Genet***44**, 1026-1029, doi:10.1038/ng.2367 (2012).
38 Gangisetty, O., Lauffart, B., Sondarva, G. V., Chelsea, D. M. & Still, I. H. The transforming acidic coiled coil proteins interact with nuclear histone acetyltransferases. *Oncogene***23**, 2559-2563, doi:10.1038/sj.onc.1207424 (2004).
39 Goshima, G., Mayer, M., Zhang, N., Stuurman, N. & Vale, R. D. Augmin: a protein complex required for centrosome-independent microtubule generation within the spindle. *J Cell Biol***181**, 421-429, doi:10.1083/jcb.200711053 (2008).
40 Uehara, R. *et al.* The augmin complex plays a critical role in spindle microtubule generation for mitotic progression and cytokinesis in human cells. *Proc Natl Acad Sci U S A***106**, 6998-7003, doi:10.1073/pnas.0901587106 (2009).
41 Cheung, P. P., Kunapuli, S. P., Scott, C. F., Wachtfogel, Y. T. & Colman, R. W. Genetic basis of total kininogen deficiency in Williams' trait. *J Biol Chem***268**, 23361-23365 (1993).

42 Merkulov, S. *et al.* Deletion of murine kininogen gene 1 (mKng1) causes loss of plasma kininogen and delays thrombosis. *Blood***111**, 1274-1281, doi:10.1182/blood-2007-06-092338 (2008).
43 Takagaki, Y., Kitamura, N. & Nakanishi, S. Cloning and sequence analysis of cDNAs for human high molecular weight and low molecular weight prekininogens. Primary structures of two human prekininogens. *J Biol Chem***260**, 8601-8609 (1985).
44 Valtonen-André, C. *et al.* A highly conserved protein secreted by the prostate cancer cell line PC-3 is expressed in benign and malignant prostate tissue. *Biol Chem***388**, 289-295, doi:10.1515/BC.2007.032 (2007).
45 Henske, E. P. *et al.* Identification of VAV2 on 9q34 and its exclusion as the tuberous sclerosis gene TSC1. *Ann Hum Genet***59**, 25-37 (1995).
46 Moores, S. L. *et al.* Vav family proteins couple to diverse cell surface receptors. *Mol Cell Biol***20**, 6364-6373 (2000).

Enrichment analysis

Method

We performed an enrichment analysis with *Gowinda* (Kofler and Schlötterer 2012), a method relying on raw SNP scores (in our case the likelihood ratios) as input and which corrects for over-counting genes that contain many SNPs that are near each other, using a permutation test.

Results

As our test SNPs, we used all SNPs with a likelihood ratio score higher than the 99.5% quantile of the score distribution. We used all SNPs that had a score as the background set. The results of the enrichment analysis for a false discovery rate less than or equal to 0.1 are shown in Table S16.2. The results include interesting functions such as epidermis development, hyaluronoglucuronidase activity and vitamin K metabolism. Future work to detect selection in Aboriginal Australians should allow to further assess whether these genes have indeed been under selection for this population.

Table S16.2 Results of the enrichment analysis based on the results of the selection scan described above.

GO_term	#genes	p_value uncorrected	FDR	Num_genes	Description	Gene_IDs
GO:0033906	2	0.00004	0.080	2	hyaluronoglucuronidase activity	ensg00000186792,ensg00000068001
GO:0008544	19	0.00008	0.080	19	epidermis development	ensg00000169594,ensg00000169814,ensg00000127863,ensg00000053747,ensg00000080166,ensg00000109101,ensg00000135960,ensg00000185652,ensg00000167769,ensg00000167754,ensg00000135925,ensg00000146648,ensg00000167880,ensg00000092969,ensg00000137709,ensg00000169744,ensg00000075275,ensg00000169692,ensg00000182568
GO:0045240	5	0.00009	0.080	5	dihydrolipoyl dehydrogenase complex	ensg00000105953,ensg00000119689,ensg00000083123,ensg00000103507,ensg00000248098
GO:0042373	4	0.0001	0.080	4	vitamin K metabolic process	ensg00000167397,ensg00000120942,ensg00000186204,ensg00000196715
GO:0042503	4	0.00011	0.080	4	tyrosine phosphorylation of Stat3 protein	ensg00000099985,ensg00000151422,ensg00000101213,ensg00000128342
GO:0071493	3	0.00012	0.080	3	cellular response to UV-B	ensg00000186792,ensg00000114378,ensg00000068001
GO:0045239	6	0.00015	0.080	6	tricarboxylic acid cycle enzyme complex	ensg00000091483,ensg00000105953,ensg00000119689,ensg00000083123,ensg00000103507,ensg00000248098

S16 References

Berg JJ, Coop G. 2014. A population genetic signal of polygenic adaptation. PLoS Genet. 10:e1004412.

Bryk J, Hardouin E, Pugach I, Hughes D, Strotmann R, Stoneking M, Myles S. 2008. Positive selection in East Asians for an EDAR allele that enhances NF-kappaB activation. PLoS One 3(5):e2209.

Cabral A, Fischer DF, Vermeij WP, Backendorf C. 2003. Distinct functional interactions of human Skn-1 isoforms with Ese-1 during keratinocyte terminal differentiation. J Biol Chem. 278(20):17792-9.

Cheung C-L, Lau K-S, Ho AYY, Lee K-K, Tiu S-C, Lau EYF, Leung J, Tsang M-W, Chan K-W, Yeung C-Y, et al. 2012. Genome-wide association study identifies a susceptibility locus for thyrotoxic periodic paralysis at 17q24.3. Nat. Genet. 44:1026–1029.

Kofler R, Schlötterer C. 2012. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. Bioinformatics 28:2084–2085.

Leppäluoto J, Hassi J. 1991. Human Physiological Adaptations to the Arctic Climate. ARCTIC 44:139–145.

Porcu E, Medici M, Pistis G, Volpato CB, Wilson SG, Cappola AR, Bos SD, Deelen J, den Heijer M, Freathy RM, et al. 2013. A Meta-Analysis of Thyroid-Related Traits Reveals Novel Loci and Gender-Specific Differences in the Regulation of Thyroid Function. PLoS Genet 9:e1003266.

Qi X, Chan WL, Read RJ, Zhou A, Carrell RW. 2014. Temperature-responsive release of thyroxine and its environmental adaptation in Australians. Proc. R. Soc. Lond. B Biol. Sci. 281:20132747.

Rogers S, Macheda ML, Docherty SE, Carty MD, Henderson MA, Soeller WC, Gibbs EM, James DE, Best JD. 2002. Identification of a novel glucose transporter-like protein—GLUT-12. Am. J. Physiol. - Endocrinol. Metab. 282:E733–E738.

Scholander PF, Hammel HT, Hart JS, LeMessurier DH, Steen J. 1958. Cold Adaptation in Australian Aborigines. J. Appl. Physiol. 13:211–218.

Tin A, Woodward OM, Kao WHL, Liu C-T, Lu X, Nalls MA, Shriner D, Semmo M, Akylbekova EL, Wyatt SB, et al. 2011. Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele. Hum. Mol. Genet. 20:4056–4068.

Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS, et al. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science 329:75–78.