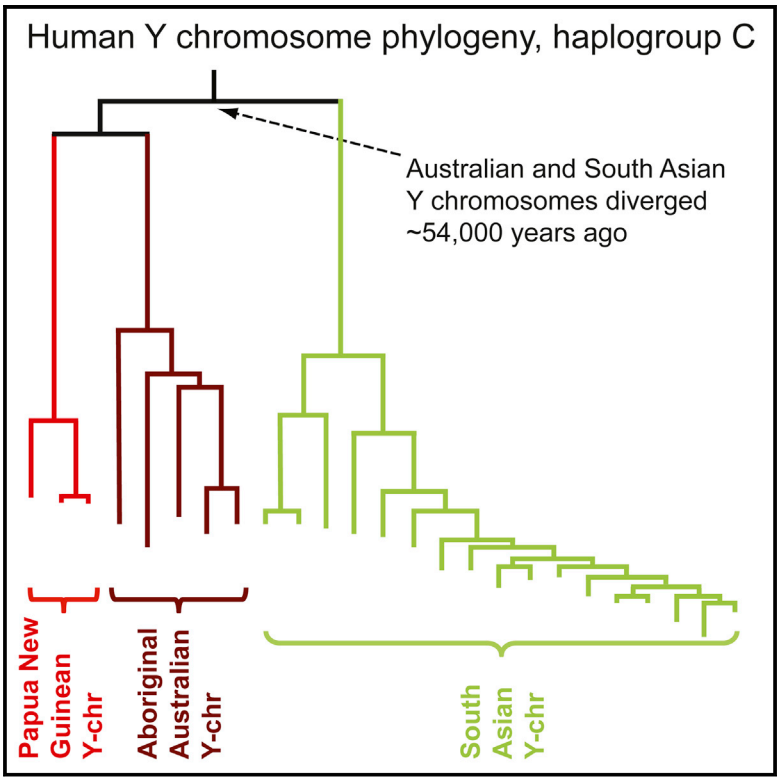
Report

Current Biology

Deep Roots for Aboriginal Australian Y Chromosomes

Graphical Abstract



Highlights

- We have sequenced 13 Aboriginal Australian Y chromosomes
- These diverged from Y chromosomes in other continents around 50,000 years ago
- They diverged from Papua New Guinean Y chromosomes soon after this
- We find no evidence for Holocene male gene flow to Australia from South Asia

Authors

Anders Bergström, Nano Nagle, Yuan Chen, ..., Yali Xue, R. John Mitchell, Chris Tyler-Smith

Correspondence

john.mitchell@latrobe.edu.au (R.J.M.), cts@sanger.ac.uk (C.T.-S.)

In Brief

Bergström et al. show that Aboriginal Australian Y chromosomes diverged from Eurasian, including South Asian, Y chromosomes ~50,000 years ago. This is around the time that Australia was first populated and thus disproves the previous hypothesis of prehistoric Y chromosome gene flow from India ~5,000 years ago.

Bergström et al., 2016, Current Biology *26*, 809–813
March 21, 2016
http://dx.doi.org/10.1016/j.cub.2016.01.028

Current Biology
Report

Deep Roots for Aboriginal Australian Y Chromosomes

Anders Bergström,[1] Nano Nagle,[2] Yuan Chen,[1] Shane McCarthy,[1] Martin O. Pollard,[1,3] Qasim Ayub,[1] Stephen Wilcox,[4,5] Leah Wilcox,[2] Roland A.H. van Oorschot,[6] Peter McAllister,[7] Lesley Williams,[8] Yali Xue,[1] R. John Mitchell,[2,*] and Chris Tyler-Smith[1,*]

[1]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK
[2]Department of Biochemistry and Genetics, La Trobe Institute of Molecular Sciences, La Trobe University, Melbourne, VIC 3086, Australia
[3]Department of Medicine, University of Cambridge, Cambridge CB2 2QQ, UK
[4]Australian Genome Research Facility, Melbourne, Victoria 3052, Australia
[5]Division of Systems Biology and Personalised Medicine, Walter and Eliza Hall Institute of Medical Research, Melbourne, VIC 3052 Australia
[6]Office of the Chief Forensic Scientist, Victorian Police Forensic Services Department, Melbourne, VIC 3085, Australia
[7]Griffith University, Brisbane, QLD 4222, Australia
[8]Community Elder and Cultural Advisor, Brisbane, QLD 4011, Australia
*Correspondence: john.mitchell@latrobe.edu.au (R.J.M.), cts@sanger.ac.uk (C.T.-S.)
http://dx.doi.org/10.1016/j.cub.2016.01.028

SUMMARY

Australia was one of the earliest regions outside Africa to be colonized by fully modern humans, with archaeological evidence for human presence by 47,000 years ago (47 kya) widely accepted [1, 2]. However, the extent of subsequent human entry before the European colonial age is less clear. The dingo reached Australia about 4 kya, indirectly implying human contact, which some have linked to changes in language and stone tool technology to suggest substantial cultural changes at the same time [3]. Genetic data of two kinds have been proposed to support gene flow from the Indian subcontinent to Australia at this time, as well: first, signs of South Asian admixture in Aboriginal Australian genomes have been reported on the basis of genome-wide SNP data [4]; and second, a Y chromosome lineage designated haplogroup C*, present in both India and Australia, was estimated to have a most recent common ancestor around 5 kya and to have entered Australia from India [5]. Here, we sequence 13 Aboriginal Australian Y chromosomes to re-investigate their divergence times from Y chromosomes in other continents, including a comparison of Aboriginal Australian and South Asian haplogroup C chromosomes. We find divergence times dating back to ~50 kya, thus excluding the Y chromosome as providing evidence for recent gene flow from India into Australia.

RESULTS AND DISCUSSION

Genotyping and Sequencing of Aboriginal Australian Y Chromosomes

144 self-identified Aboriginal Australian males who volunteered to participate in the Genographic Project were previously typed with Y SNPs to assign them to major haplogroups [6]. A large fraction (~70%) of Aboriginal Australian males today carry Y chromosomes of Eurasian origin (~59% European) due to admixture in the last ~200 years after the European colonization of Australia [7]. Among the individuals with indigenous Y chromosomes, 44% belong to haplogroup C, with 42% being C-M347 and 2% the basal C-M130*. Paragroup K* constitutes 56% of indigenous Y chromosomes, with 27% being S-P308, 2% being haplogroup M-M186, and 27% being the basal K-M526* [6]. Although we note that other nomenclatures with relevance to these haplogroups exist [8] or could be proposed, these labels suffice for the purposes of our present study, and for simplicity we hereafter refer to C-M347 and C-M130* as Aboriginal Australian C, to S-P308 and K-M526 as K*, and to M-M186 as M. For distinguishing subclades of haplogroup C, we also make use of the haplogroup labels C1, C2, C3, C4, and C5 as they are used in [9]. 31 of the 144 typed individuals carried Y chromosomes belonging to one of the indigenous haplogroups. Among these individuals, five from haplogroup C, six from haplogroup K*, and two from haplogroup M were re-contacted and agreed to further studies, so their genomes were sequenced to high coverage using the Illumina HiSeq platform (Table 1). Consent was provided to study the history of the uniparental chromosomes, and reads mapping to the Y chromosome were identified. These form the basis for the current study. Comparative data on the sequences of Y chromosomes from other continents were obtained from phase 3 of the 1000 Genomes Project [10], comprising 1,244 samples from 26 populations falling into a wide range of haplogroups, as well as from 12 samples from Papua New Guinea [11] which fall into the haplogroups C, M, and K* as expected [12, 13].

Construction of a Y Chromosome Phylogeny

We used the sequence data to infer a maximum-likelihood phylogenetic tree for the 1,269 Y chromosomes (Figure 1A) (see the Experimental Procedures and Supplemental Experimental Procedures). The overall topology of the tree recapitulates the known Y chromosome phylogeny. In agreement with the prior haplogroup assignments, the Aboriginal Australian and Papuan Y chromosomes fall into two distinct monophyletic clades within the C and K*/M haplogroups. Both of these clades

Table 1. Aboriginal Australian Individuals Sampled for This Study

ID	Y Coverage	Haplogroup	Key Variant	Paternal Origin
A45	19.74	C	M130	Uncertain, possibly Normanton, Queensland
A268	13.06	C	M210	Atherton Tablelands, Far North Queensland
A305	18.03	C	M347	The Karryarra group located near Port Hedland, Western Australia
A342	18.06	C	M347	The Karryarra group located near Port Hedland, Western Australia
A343	12.90	C	M347	Northwest coast, near Broome, Western Australia
A136	12.61	K*	M526	Kuranda, Far North Queensland
A179	18.85	K*	M526	Gunganji tribe, Yarrabah, near Cairns, Far North Queensland
A201	12.19	K*	M526	Uncertain, but states father's people from South East Queensland
A266	19.07	K*	M526	Gunganji tribe, Yarrabah, near Cairns, Far North Queensland
A293	12.77	K*	P308	Pilbara, Western Australia
A473	13.73	K*	P308	Mount Isa region, Central Queensland
A238	16.42	M	M186	Mer (Murray Island), Torres Strait, Far North Queensland
A440	15.29	M	M186	Mer (Murray Island), Torres Strait, Far North Queensland

"Y coverage" refers to the average depth of sequencing coverage on the Y chromosome. We note that the geographic information on the origin of the paternal line is sometimes uncertain and, due to the widespread movement of Aboriginal people after European colonization, might not reflect deeper geographic origins.

received high bootstrap support (100% for the haplogroup C samples and 97% for the haplogroup K*/M samples). The shared phylogenetic history of Aboriginal Australian and Papuan Y chromosomes is consistent with the common origin of these populations as previously inferred from genome-wide data [4, 15–17].

Divergence Times between Aboriginal Australian and Other Y Chromosomes

The phylogenetic tree reveals deep divergences between Y chromosomes indigenous to Sahul, the ancient continent that included both Australia and New Guinea, and those from all other populations (Figures 1B and 1C). Complete sequence data allow direct and accurate inference of the timing of these divergences. Applying a point mutation rate of 0.76×10^{-9} per site per year inferred from the number of missing mutations on the Y chromosome of a ~45-ky-old radiocarbon-dated Eurasian sample [18], we infer a divergence time of 54.3 ky (95% confidence interval [CI]: 48.0–61.6 ky) between K*/M chromosomes in Sahul and their closest relatives in the R and Q haplogroups (Figure 1B), and a divergence time of 54.1 KY (95% CI: 47.8–61.4 ky) between Sahul C chromosomes and their closest relatives in the C5 haplogroup (Figure 1C), a distinction noted previously on the basis of a single SNP, M347 [9]. These dates are consistent with the archeological record documenting human occupation in Australia by ~47 kya [2] and with genome-wide analyses that have found an early divergence between the ancestors of Eurasian populations and the ancestors of Aboriginal Australians and Papuans [15]. They thus provide no evidence for any later Y chromosome gene flow into Australia between the early separation and the beginning of recent European colonization. Specifically, these results refute earlier findings based on short tandem repeat (STR) variation that Aboriginal Australian Y chromosomes in the C haplogroup descend from populations in southern India and Sri Lanka 1.3–13.3 kya [5]. Although the closest chromosomes to the Aboriginal Australian Cs in our phylogeny are found in South Asian populations, the deep divergence time and the fact that the Aboriginal Australian Cs share a more recent common ancestor with Papuan Cs show that this is not the result of recent genetic contact. The CIs reported above take into account the uncertainty of the Y chromosome point mutation rate, but not necessarily other possible sources of technical uncertainty (such as read alignment and genotype calling). We tested whether accounting for such additional uncertainty could affect the conclusion of a deep divergence between Aboriginal Australian and South Asian C chromosomes by re-estimating this divergence time from 100 bootstrap samples of sites from the full ~10 million analyzed Y chromosome sites. The 95% CI for these estimates was 50.9–58.1 kya, and very conservative application of the mutation rate uncertainty multiplicatively to the bootstrap estimates gives a combined CI of 44.9–65.9 kya. Technical uncertainty is thus not large enough to affect our overall conclusion. The disparity between our findings and the earlier report can be attributed to improvements in technology, as none of the methods previously used to study the history of the paternal lineage offered the level of phylogenetic or dating precision afforded by complete Y chromosome sequencing. Redd et al. employed ten widely used Y STRs (three simple trinucleotides, three simple tetranucleotides, one of which was bilocal, and four complex tetranucleotides), applying the same fast genealogical mutation rate of 2.08×10^{-3} per STR per 25 years to all of them [5]. It has been shown that Y STRs tend to massively under-estimate ancient divergence times [19], perhaps because of a combination of the fast mutation rate assumed, saturation of STR distances, and in this case the short generation time used.

Although the shared origin of Aboriginal Australians and Papuans is clearly established, and now also supported by the Y chromosome phylogeny presented here, little is known about the history of population separation and gene flow between these groups within Sahul. We observe deep divergences between Aboriginal Australian and Papuan Y chromosomes within the C (50.1 ky; 95% CI: 44.3–56.9 ky) (Figure 1C) and the K* (48.4 ky; 95% CI: 42.8–54.9 ky) (Figure 1B) haplogroups. Although this would be consistent with an early split between the populations,

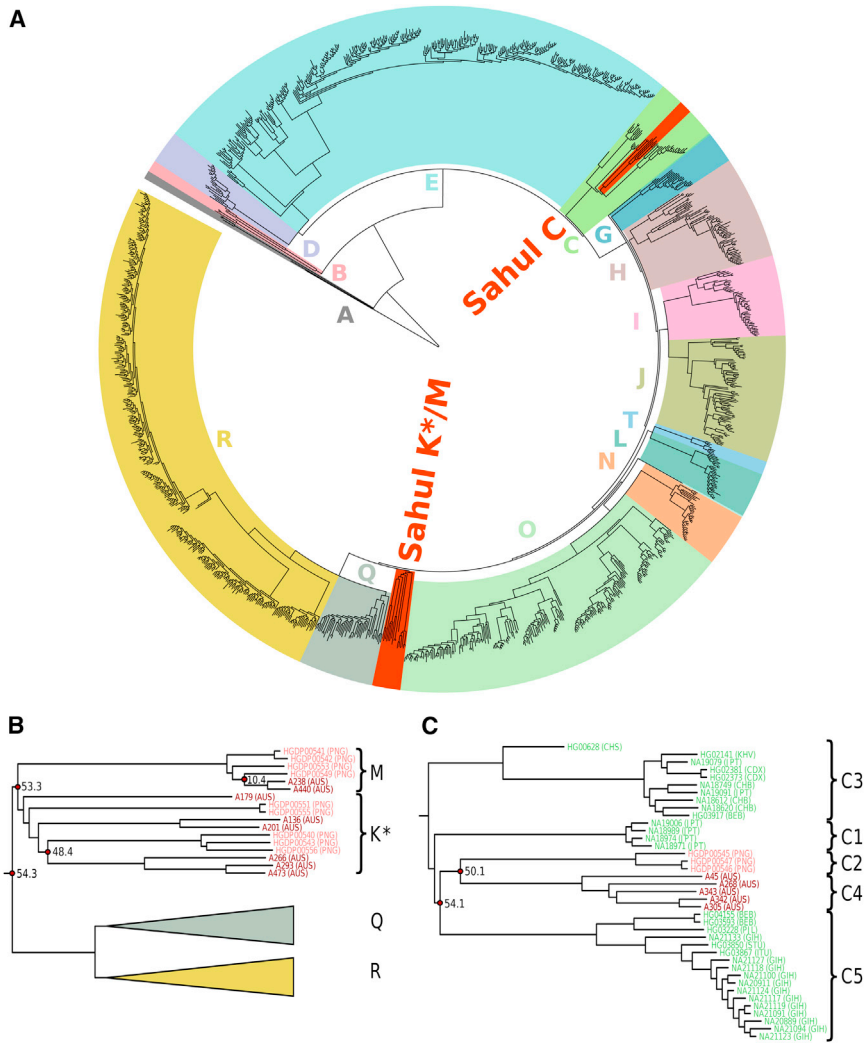


Figure 1. Phylogenetic History of Aboriginal Australian Y Chromosomes

(A) A maximum-likelihood phylogeny was inferred from the Y chromosome data of 1,269 males from worldwide populations, including Aboriginal Australians, using RAxML [14]. High-level haplogroups are colored and labeled along the tree. The two clades that contain the Y chromosomes indigenous to the continent of Sahul (from the Aboriginal Australian and Papuan samples) are indicated in bright red.

(B) The phylogeny of Y chromosomes in haplogroups K* and M. This detailed view of a part of the larger tree displayed in (A) focuses on chromosomes in haplogroups K* and M. Haplogroups Q and R, which are the closest relatives to K* and M in the phylogeny, are represented schematically because they contain very large numbers of samples. Aboriginal Australian and Papuan samples are colored in two different shades of red for easier visual separation. Sample names and population origins are displayed at branch tips (AUS, Aboriginal Australian; PNG, Papua New Guinean). Divergence times in units of thousands of years are indicated on key nodes that correspond to divergences between groups of samples from different populations or haplogroups.

(C) The phylogeny of Y chromosomes in haplogroup C. Sample names and population origins are displayed at branch tips (AUS, Aboriginal Australian; PNG, Papua New Guinean; CHS, Southern Han Chinese in China; KHV, Kinh in Ho Chi Minh City, Vietnam; JPT, Japanese in Tokyo, Japan; CDX, Chinese Dai in Xishuangbanna, China; CHB, Han Chinese in Bejing, China; BEB, Bengali in Bangladesh; PJL, Punjabi in Lahore, Pakistan; GIH, Gujarati Indian in Houston, Texas; STU, Sri Lankan Tamil in the UK; ITU, Indian Telugu in the UK). We note that due to factors associated with missing data arising from the low sequencing coverage of the 1000 Genomes samples, the branch lengths displayed here are not strictly proportional to time.

See also Table S1.

we note that our limited sample size makes it very unlikely that we have observed the most recent divergences, and we therefore cannot rule out more recent split times. Within the M haplogroup, which is found at high frequencies in Papua New Guinea and Melanesia [20] but in less than 1% of Aboriginal Australian males [6], we find a divergence time of 10.4 ky (95% CI: 9.2–11.9 ky) (Figure 1B). Although this coincides approximately with the postglacial geographical separation of Australia and New Guinea after the rise of the sea level ~6–8 kya [21], the fact that the two Aboriginal Australian males who carry the haplogroup M chromosomes trace their paternal ancestry to the Torres Strait Islands (Table 1) makes it more likely that these are very recent introductions into the mainland Australian gene pool. A larger number of geographically diverse Y chromosomes from the different haplogroups indigenous to Sahul would be needed in order to learn more about population relationships within the continent.

Implications for the Peopling of Australia

Y haplogroups from Australia and Papua New Guinea were estimated to diverge from the nearest non-Sahul lineages ~54 kya, and divergences within Sahul-specific lineages date to ~48–53 kya. We note that these times post-date the Mount Toba eruption ~74 kya [22], supporting a model of the initial peopling of this region by modern humans long after this event. The divergence times are close to, but earlier than, the current conservative archaeological date for entry into Sahul, 47 kya [2]. However, the uncertainty in the lineage divergence estimates and the possibility that earlier archaeological sites may be detected make it impossible to determine whether the initial divergence within the Sahul-specific lineages occurred before or after entry into Sahul. The current evidence is consistent with a simple model of a single entry and subsequent rapid lineage divergence.

Around the mid-Holocene (~4–6 kya), small stone tools began to be used extensively in Australia [3], the Pama-Nyungan language family spread over most of the mainland [23], and the first archaeological evidence for the dingo appeared [3]. Genetic patterns proposed as indications of gene flow into Australia from South Asia were dated to approximately the same period. One parsimonious interpretation of these diverse findings could be

that they were all linked, and thus that there was a substantial and influential population influx at this time.

We have taken advantage of improvements of sequencing technology [10] and calibration of the molecular clock [18] to re-examine the claim for male gene flow revealed by Y chromosome relationships [5]. Our sample of 13 Aboriginal Australian Y chromosomes is small, but it includes the relevant haplogroups and conclusively refutes the original basis for this claim. Although this does not demonstrate the absence of any Holocene gene flow or non-genetic influences from South Asia at this time, and the appearance of the dingo remains as strong evidence for external contacts, the evidence overall is consistent with a complete lack of gene flow and indigenous origins for the technological and linguistic changes.

Australia and Papua New Guinea are currently separated only by the 150-km-wide Torres Strait, in which lie many islands. Gene flow across this Strait is both geographically plausible and demonstrated by our data, although we cannot determine when within the last 10 ky it occurred. The analytical techniques now available, applied to larger genetic datasets, including ancient DNA, have the potential to address such questions and provide more detailed insights into the human history of Sahul.

EXPERIMENTAL PROCEDURES

This study received ethical approval from the La Trobe University Human Ethics Committee, Melbourne, Australia (HEC 05/94, April 11, 2006; amended April 18, 2012, June 26, 2012) and The Wellcome Trust Sanger Institute Human Materials and Data Management Committee, Hinxton, UK (12/055). Conclusions from the study have been returned to the participants. We sequenced the whole genomes of 13 Aboriginal Australian males to high coverage on the Illumina HiSeq platform and then analyzed only the reads mapping to the Y chromosome. We used FreeBayes to determine the genotypes of these individuals, along with those of 1,244 males sequenced to low coverage in the 1000 Genomes Project [10] and 12 males from Papua New Guinea sequenced to high coverage [11], at ~10 million Y chromosome sites accessible by short read sequencing. We then used RAxML [14] to infer a maximum-likelihood phylogeny of all the 1,269 Y chromosomes. We estimated the divergence times between clades in the tree by applying the ρ statistic [24], aggregating data across low-coverage samples where relevant, and converted divergence times to units of years by applying a mutation rate of 0.76×10^{-9} per site per year [18]. For more detailed descriptions of the sequence data processing, genotyping and filtering, phylogenetic inference, and dating, see the Supplemental Experimental Procedures. Table S1 provides information on the SNPs called that are phylogenetically informative for the branches of the Y chromosome phylogeny specific to Aboriginal Australians and Papuans (see the Supplemental Experimental Procedures for a description of this table).

ACCESSION NUMBERS

Y chromosome sequence data from the 13 Aboriginal Australians are available for studies of population history under managed access through two separate study accession numbers at the European Genome-phenome Archive (EGA): EGAS00001000315 and EGAS00001000718, both with EGA DAC accession number EGAC00001000205 and EGA policy accession number EGAP00001000210. The correspondence between the sample IDs used in this manuscript and the sample accession numbers is as follows: A45, EGAN00001072788; A136, EGAN00001196788; A179, EGAN00001072787; A201, EGAN00001072789; A238, EGAN00001089015; A266, EGAN00001089016; A268, EGAN00001192719; A293, EGAN00001192720; A305, EGAN00001089017; A342, EGAN00001088616; A343, EGAN00001192721; A440, EGAN00001196789; and A473, EGAN00001196790. Details of SNPs called within the Sahul-specific branches are provided in Table S1. We request that L. Williams (geraniumgroup@gmail.com), R.J.M., and C.T.-S. be consulted before any commercial use is made of novel SNPs within this table.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2016.01.028.

AUTHOR CONTRIBUTIONS

Project design was carried out by Y.X., R.J.M., and C.T.-S.; community engagement, ethics, and sampling by S.W., L. Wilcox, R.A.H.v.O., P.M., L. Williams, and R.J.M.; data generation, processing, and analysis by A.B., N.N., Y.C., S.M., M.O.P., Q.A., Y.X., and C.T.-S.; data interpretation by A.B., N.N., S.W., R.A.H.v.O., P.M., L. Williams, Y.X., R.J.M., and C.T.-S.; and manuscript writing by A.B., Y.X., R.J.M., and C.T.-S.

ACKNOWLEDGMENTS

We thank the Aboriginal Australian men and their communities for their interest and participation in this study. We thank the Wellcome Trust Sanger Institute Core Pipelines and NPG groups for their special efforts in arranging access to the Y chromosome data for this project. We also thank the 1000 Genomes Project consortium for sharing data and analysis strategies. The GATK3 program was made available through the generosity of Medical and Population Genetics program at the Broad Institute. Our work was supported by Wellcome Trust grant 098051.

Received: December 9, 2015
Revised: January 6, 2016
Accepted: January 12, 2016
Published: February 25, 2016

REFERENCES

1. Roberts, R.G., Jones, R., and Smith, M.A. (1990). Thermoluminescence dating of a 50,000-year-old human occupation site in northern Australia. Nature *345*, 153–156.
2. O'Connell, J.F., and Allen, J. (2015). The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago. J. Arch. Sci. *56*, 73–84.
3. Brown, P. (2013). Palaeoanthropology: of humans, dogs and tiny tools. Nature *494*, 316–317.
4. Pugach, I., Delfin, F., Gunnarsdóttir, E., Kayser, M., and Stoneking, M. (2013). Genome-wide data substantiate Holocene gene flow from India to Australia. Proc. Natl. Acad. Sci. USA *110*, 1803–1808.
5. Redd, A.J., Roberts-Thomson, J., Karafet, T., Bamshad, M., Jorde, L.B., Naidu, J.M., Walsh, B., and Hammer, M.F. (2002). Gene flow from the Indian subcontinent to Australia: evidence from the Y chromosome. Curr. Biol. *12*, 673–677.
6. Nagle, N., Ballantyne, K.N., van Oven, M., Tyler-Smith, C., Xue, Y., Taylor, D., Wilcox, S., Wilcox, L., Turkalov, R., van Oorschot, R.A., et al.; Genographic Consortium (2015). Antiquity and diversity of aboriginal Australian Y-chromosomes. Am. J. Phys. Anthropol. Published online October 30, 2015. http://dx.doi.org/10.1002/ajpa.22886.
7. Taylor, D., Nagle, N., Ballantyne, K.N., van Oorschot, R.A., Wilcox, S., Henry, J., Turakulov, R., and Mitchell, R.J. (2012). An investigation of admixture in an Australian Aboriginal Y-chromosome STR database. Forensic Sci. Int. Genet. *6*, 532–538.
8. Karafet, T.M., Mendez, F.L., Sudoyo, H., Lansing, J.S., and Hammer, M.F. (2015). Improved phylogenetic resolution and rapid diversification of Y-chromosome haplogroup K-M526 in Southeast Asia. Eur. J. Hum. Genet. *23*, 369–373.
9. Hudjashov, G., Kivisild, T., Underhill, P.A., Endicott, P., Sanchez, J.J., Lin, A.A., Shen, P., Oefner, P., Renfrew, C., Villems, R., and Forster, P. (2007).

Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. Proc. Natl. Acad. Sci. USA *104*, 8726–8730.

10. 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
11. Raghavan, M., Steinrücken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Ávila-Arcos, M.C., Malaspinas, A.S., et al. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. Science *349*, aab3884.
12. Shi, W., Ayub, Q., Vermeulen, M., Shao, R.G., Zuniga, S., van der Gaag, K., de Knijff, P., Kayser, M., Xue, Y., and Tyler-Smith, C. (2010). A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. Mol. Biol. Evol. *27*, 385–393.
13. Lippold, S., Xu, H., Ko, A., Li, M., Renaud, G., Butthof, A., Schröder, R., and Stoneking, M. (2014). Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. Investig. Genet. *5*, 13.
14. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.
15. Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K.E., 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.
16. Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M.R., Pugach, I., Ko, A.M., Ko, Y.C., Jinam, T.A., Phipps, M.E., et al. (2011). Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am. J. Hum. Genet. *89*, 516–528.
17. McEvoy, B.P., Lind, J.M., Wang, E.T., Moyzis, R.K., Visscher, P.M., van Holst Pellekaan, S.M., and Wilton, A.N. (2010). Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry. Am. J. Hum. Genet. *87*, 297–305.
18. Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L., 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.
19. Wei, W., Ayub, Q., Xue, Y., and Tyler-Smith, C. (2013). A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping. Forensic Sci. Int. Genet. *7*, 568–572.
20. Kayser, M., Brauer, S., Weiss, G., Schiefenhövel, W., Underhill, P., Shen, P., Oefner, P., Tommaseo-Ponzetta, M., and Stoneking, M. (2003). Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am. J. Hum. Genet. *72*, 281–302.
21. Woodroffe, C.D., Kennedy, D.M., Hopley, D., Rasmussen, C.E., and Smithers, S.G. (2000). Holocene reef growth in Torres Strait. Mar. Geol. *170*, 331–346.
22. Petraglia, M., Korisettar, R., Boivin, N., Clarkson, C., Ditchfield, P., Jones, S., Koshy, J., Lahr, M.M., Oppenheimer, C., Pyle, D., et al. (2007). Middle Paleolithic assemblages from the Indian subcontinent before and after the Toba super-eruption. Science *317*, 114–116.
23. Evans, N., and McConvell, P. (1997). The enigma of Pama-Nyungan expansion in Australia. In Archaeology and Language II: Archaeological Data and Linguistic Hypotheses, R. Blench, and M. Spriggs, eds. (Routledge - Taylor & Francis), pp. 174–192.
24. Forster, P., Harding, R., Torroni, A., and Bandelt, H.J. (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. *59*, 935–945.

Current Biology, Volume *26*

Supplemental Information

Deep Roots for Aboriginal Australian Y Chromosomes

Anders Bergström, Nano Nagle, Yuan Chen, Shane McCarthy, Martin O. Pollard, Qasim Ayub, Stephen Wilcox, Leah Wilcox, Roland A.H. van Oorschot, Peter McAllister, Lesley Williams, Yali Xue, R. John Mitchell, and Chris Tyler-Smith

Table S1. Novel Y-SNPs within haplogroups C and K*/M. Column A: chromosomal position in build GRCh37. B: Reference allele. C: Alternative allele. D: ISOGG information (January 18th, 2014 version (v9.05), http://isogg.org/), where available (. = no information). E-AC: allele called in each of the 25 Aboriginal Australian and Papua New Guinean samples. These are the SNPs that define the branches highlighted in red/dark red in Figure 1A, 1B and 1C. For more information see Supplemental Experimental Procedures.

This table is provided as an Excel file.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Samples

The 13 males included in this study are drawn from a sample collected as part of a larger study of worldwide genomic variation, the Genographic Project, conducted between 2005 and 2013. This project focused on variation within mitochondrial DNA and the Y chromosome. In Australia, the project attempted as wide a geographic coverage of Aboriginal Australians as possible.

We contacted Aboriginal elders in each region to help publicise the Genographic Project and held local meetings to which they were invited. All who identified as Aboriginal were invited to participate in the study, even if they knew they had no direct maternal or paternal line of Aboriginal descent. Aboriginality is a culturally based affiliation and not defined by a person's genetic make-up. DNA was extracted from a saliva sample collected using the Oragene collection kit (http://www.dnagenotek.com/ROW/products/OG500.html).

Following genotyping within the Genographic Project and feedback of results to participants, we attempted to re-contact all haplogroup C, K* and M males to seek their permission for full sequencing of their Y chromosome, and received this permission from 13. Where required, a second saliva sample was collected and DNA extracted as described [S1].

The present study received ethical approval from the La Trobe University Human Ethics Committee, Melbourne, Australia (HEC 05/94, 11th April 2006; amended 18th April 2012, 26th June 2012) and The Wellcome Trust Sanger Institute Human Materials and Data Management Committee, Hinxton, UK (12/055). Conclusions from the study have been returned to the participants.

DNA sequencing, data processing and genotype calling

A single sequencing library was constructed per sample and each library was sequenced across multiple lanes on the Illumina HiSeq platform (read length 100 base-pairs, target fragment length 350 base-pairs).

The sequence reads were mapped to the hs37d5 version of the human reference genome using bwa aln v0.5.9 with the argument "-q 15". Duplicate reads were marked using Picard MarkDuplicates v1.06. The consent obtained from the sample donors only allowed study of the Y chromosome and mitochondrial DNA, and therefore only read pairs mapping properly to these parts of the genome were accessed for further analysis. Mitochondrial DNA results will be reported elsewhere. GATK v.3.3 IndelRealigner was used to perform local realignment around known indels on the Y chromosome (the "Mills-Devine" indel set and the "20101123" 1000 Genomes Phase 2 low coverage indel set, both obtained from the 1000 Genomes Project), with the argument "–LOD 0.4". Base Alignment Qualities (BAQ) were computed using samtools [S2]. These steps were performed so as to largely match the processing applied to the low coverage read alignments in 1000 Genomes Phase 3. Depth of coverage along the chromosome was calculated using samtools [S2].

Y chromosome read alignments for all 1244 male samples in Phase 3 were obtained from the 1000 Genomes Project [S3] (mean Y chromosome coverage = 3.81). Y chromosome read alignments for 12 male samples from Papua New Guinea sequenced to high-coverage were obtained from [S4] (mean Y chromosome coverage = 9.91).

Genotypes were called jointly across all 1269 samples using FreeBayes v0.9.18 [S5] with the arguments "—ploidy 1" to accommodate the haploid state of the Y chromosome and "—report-monomorphic" to obtain genotype calls also at sites without any evidence for polymorphism. Calling was restricted to the regions of the Y chromosome deemed suitable for Illumina short read sequencing [S6], totaling 10,445,993 base-pairs. Furthermore, sites on which the data fulfilled any of the following criteria were excluded:

1) A depth of coverage of reads with mapping quality greater than or equal to 1 in the top or bottom 1% of sites (corresponding to <2401 or >6064 such reads across all samples)

2) A ratio greater than 0.1 of the number of reads with mapping quality 0 to the total number of reads
3) A fraction greater than 0.3 of the number of samples with missing genotype in the unfiltered genotype calls
4) A number of samples greater than 200 having more than 1 read not supporting the called genotype at the site

These filters are similar to those used for the Y chromosome in the 1000 Genomes Project [S3]. The sample genotypes at these sites were then recalled by setting the genotype to the allele with the highest number of supporting reads, or to missing if no allele was supported by at least 2 reads, if more than one allele was supported by more than 1 read or if the fraction of reads supporting the majority allele was lower than 0.75. Multiple alleles at a site were all retained and sites containing possible indel alleles were not removed. Multi-nucleotide variants were decomposed to their constituent SNPs and/or indels using the tool vcfallelicprimitives from the vcflib library (https://github.com/ekg/vcflib). 9,891,532 sites with data remained for analyses. Additional site filters were applied specifically for the purpose of phylogenetic tree inference, but not for dating of split times (see below).

Phylogenetic inference and dating

A maximum likelihood phylogeny of all 1269 Y chromosomes was inferred using RAxML v8.1.15 [S7]. To reduce the computational burden, only SNP sites with a QUAL score greater than or equal to 1 were used as input for this inference, totalling 53,813 sites. RAxML was run with the ASC_GTRGAMMA model of nucleotide substitution and the "stamatakis" correction for ascertainment bias, directly specifying to the program the number of invariable sites. Statistical support for each clade in the tree was assessed from 100 bootstrap replicates. Trees were visualized using the Interactive Tree Of Life [S8] and the FigTree software (http://tree.bio.ed.ac.uk/software/figtree/) and manually rooted at the branch leading to samples in the A0 haplogroup.

The genotypes of the 12 samples in the A and B haplogroups, which are the earliest branching lineages in the Y-chromosomal phylogeny and an outgroup to all the lineages that form the focus of this study, were used to define the ancestral nucleotide state at all sites. Pooling of data across multiple samples in this fashion allows the ancestral state to be accurately inferred at a larger number of sites, as the low sequencing coverage of the 1000 Genomes samples precludes genotype calling at many sites for any given single sample. If there was just one allele present among the called genotypes of these 12 samples at a site, this allele was set as the ancestral nucleotide, while if there were multiple alleles present the ancestral state was not defined. This allowed the ancestral state to be called at all except 4825 sites (0.049%).

The ages of internal nodes in the inferred phylogenetic tree were estimated directly using the ρ statistic [S9] on all called sites. Given a branch between a first sample and an internal node corresponding to the common ancestor with a second sample, the length of the branch was estimated by dividing the number of derived mutations present in the first sample and not the second by the total number of sites with ancestral state in both samples. If either sample had missing data or an indel genotype, the site was ignored. If multiple samples were available in a given clade of the tree, the per-sample divergence estimates were averaged across them. When estimating divergence from a node for which only low coverage 1000 Genomes samples were available for comparison, data were pooled across multiple samples in the way described above for the determination of ancestral states, and the divergence was calculated against the resulting single consensus genotype. Specifically, for dating the divergence of Aboriginal Australian and Papuan haplogroup C chromosomes, data were pooled across the 17 samples in the C5 haplogroup, and for dating the divergence of Aboriginal Australian and Papuan haplogroup K* and M chromosomes data were pooled across 20 randomly selected samples from the R and Q haplogroups.

We converted divergence times in units of mutations per basepair to units of years by applying a point mutation rate of 0.76×10^{-9} mutations per site per year. This rate was inferred from the relative deficiency of mutations on the Y chromosome relative to modern humans of an ancient human sample found near Ust'-Ishim in western Siberia, radiocarbon-dated to have lived ~45 KYA [S10]. This rate is similar to, but slightly lower than, that inferred from multi-generational genealogies of Icelandic males [S11] (~0.88×10^{-9} mutations per site per year), but we reason that the integration over generation times in diverse human populations over 45 KY inherent in the Ust'-Ishim rate estimate is more appropriate for the purposes of the present study. We also note that our main conclusions would not be affected by applying a slightly different mutation rate. We report 95% confidence intervals on the divergence times corresponding to the uncertainty of the Ust'-Ishim mutation rate (0.67×10^{-9} to 0.86×10^{-9}).

Details of novel SNPs and phylogenetic inferences within haplogroups C and K*/M

Our work provides new insights into the early and Sahul-specific differentiation within haplogroups C and K*/M, and we provide information on high-quality discovered SNPs that are informative for these branches in an accessible form in Table S1. This table lists the chromosomal position of each SNP in the GRCh37/hg19 coordinate system, the reference genome allele and the alternative allele (we note that for all these SNPs we have confirmed that the reference allele is the ancestral allele), any existing marker labels present in the January 18th, 2014 version (v9.05) of the ISOGG database (http://isogg.org/) and the genotype of all 25 Aboriginal Australian and Papua New Guinean samples included in this study (where "0" denotes the reference allele and "1" denotes the alternative allele). Of the variants listed in Table 1, M186 is absent from Table S1 because it is a deletion, while M130 and M526 are absent because they lie deeper in the phylogeny than the lineages included in Table S1.

Haplogroups C2 and C4 together form a monophyletic group and in our data are distinguished from the common ancestor with C5 by 25 high-confidence SNPs; similarly, Aboriginal Australian C4 and Papua New Guinean C2 samples form two distinct monophyletic clusters. Within the K*/M monophyletic haplogroup, which is distinct from QR, M forms the first branch, with the sampled Aboriginal Australian and Papua New Guinean chromosomes sharing a common ancestor within the last ~22 KY. The remaining K* chromosomes analysed fall into five deep clades (>48 KY), three Aboriginal Australian and two Papua New Guinean. This phylogeny is consistent with the most detailed previous phylogenetic analysis of haplogroup K [S12], resolving the multifurcation within the lineage there designated K2b1 (here K*/M). Since it seems likely that further complexity may be revealed in this part of the phylogeny by additional sequencing studies in the near future, we have refrained from assigning alphanumeric labels to the clades, and suggest that if necessary they are referred to using the haplogroup and first variant listed in Table S1, e.g. K*(GRCh37- 2,658,341-T) for the lineage defined by HGDP00540, HGDP00543 and HGDP00556.

SUPPLEMENTAL REFERENCES

S1. Nagle, N., Ballantyne, K.N., van Oven, M., Tyler-Smith, C., Xue, Y., Taylor, D., Wilcox, L., Turkalov, R., van Oorschot, R.A.H., McAllister, P., et al. (2015). Antiquity and diversity of Aboriginal Australian Y-chromosomes. Am. J. Phys. Anthropol. *online*.

S2. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

S3. The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68-74.

S4. Raghavan, M., Steinrucken, M., Harris, K., Schiffels, S., Rasmussen, S., DeGiorgio, M., Albrechtsen, A., Valdiosera, C., Avila-Arcos, M.C., Malaspinas, A.S., et al. (2015). Genomic evidence for the Pleistocene and recent population history of Native Americans. Science *349*, aab3884.

S5. Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint, arXiv:1207.3907 [q-bio.GN].

S6. Poznik, G.D., Henn, B.M., Yee, M.C., Sliwerska, E., Euskirchen, G.M., Lin, A.A., Snyder, M., Quintana-Murci, L., Kidd, J.M., Underhill, P.A., et al. (2013). Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science *341*, 562-565.

S7. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.

S8. Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res *39*, W475-478.

S9. Forster, P., Harding, R., Torroni, A., and Bandelt, H.J. (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet *59*, 935-945.

S10. Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L., Aximu-Petri, A., Prufer, K., de Filippo, C., et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. Nature *514*, 445-449.

S11. Helgason, A., Einarsson, A.W., Guethmundsdottir, V.B., Sigurethsson, A., Gunnarsdottir, E.D., Jagadeesan, A., Ebenesersdottir, S.S., Kong, A., and Stefansson, K. (2015). The Y-chromosome point mutation rate in humans. Nat Genet *47*, 453-457.

S12. Karafet, T.M., Mendez, F.L., Sudoyo, H., Lansing, J.S., and Hammer, M.F. (2015). Improved phylogenetic resolution and rapid diversification of Y-chromosome haplogroup K-M526 in Southeast Asia. Eur J Hum Genet *23*, 369-373.