The genomic history of Australia

The human population history of Australia remains contentious, not least because of a lack of large extensive genomic data. We generated high-coverage genomes for 83 geographically diverse Aboriginal Australians (all speakers of Pama-Nyungan languages) and 25 Papuans from the New Guinea Highlands. We find that Papuan and Aboriginal Australian ancestors diversified from each other 25-40 thousand years ago (kya), suggesting early population structure in the ancient continent of Sahul (Australia, New Guinea and Tasmania). However, all contemporary Aboriginal Australian studied descend from a single founding population that differentiated around 10-32 kya. We find evidence for a population expansion in northeast Australia during the Holocene (past c.10 kya) associated with limited gene flow from this region to the rest of Australia. This is broadly consistent with the spread of the Pama-Nyungan languages and cultural changes taking place across the continent in the mid-Holocene. We find evidence for a single out of Africa dispersal for all contemporary humans and estimate that Aboriginal Australians and Papuans shared a common ancestor with other Eurasians 60-100 kya, with subsequent admixture with different archaic populations. Finally, we report evidence of selection in Aboriginal Australians potentially associated with living in the desert.

During most of the last 100 ky, Australia, Tasmania and New Guinea formed a single continent, Sahul, which was separated from Sunda (the continental landmass including mainland and western island Southeast Asia) by a series of deep oceanic troughs never exposed by changes in sea level (the Wallacean region as defined by biogeographers). Colonisation of Sahul is thought to have required at least 8-10 separate sea crossings between islands, potentially constraining the occupation of Australia and New Guinea by earlier hominins. The age of the first occupation of Australia has been disputed. There are several archaeological sites in Australia dating to 40-45 kya (Figure 1), long argued to represent the age of first occupation despite a few sites dating to ≥ 50 kya. However, recent studies support the earlier dates, suggesting that Sahul was first settled by 47.5-55 kya. This is consistent with the earliest evidence for modern humans in Sunda at a similar time (Figure 1). Moreover skeletal remains that share morphological similarities with the ancestors of Aboriginal Australians and Papuans are found in South East Asia up until about 3,5 kya, suggesting that the ancestors of Aboriginal Australians and Papuans extended from Sahul to Sunda.
Historically, the morphological diversity among Aboriginal Australians was interpreted by some as indicating multiple ancestral migrations or descent from Javanese Homo erectus, with varying levels of gene flow from contemporaneous populations. However, statistical analyses indicate that Australian crania show no evidence of H. erectus admixture. Still, the distinctiveness of the Australian archaeological record has led to the suggestion that the ancestors of Aboriginal Australians and Papuans (hereafter referred to as Australo-Papuans), as well as a small number of other populations, left the African continent earlier than the ancestors of present-day Eurasians. Although such multiple dispersals from Africa are supported by some genetic studies, others have found support for only one out of Africa (OoA) event, with one or two independent founding waves into Asia, of which the earlier contributed to Australo-Papuan ancestry. Recent genomic results have also shown that both Aboriginal Australian and Papuan ancestors likely admixed with Neanderthal and Denisovan archaic hominins after leaving Africa.

Once in Sahul, contact among groups would have been affected by rising sea-levels that separated the Australian continent from New Guinea and Tasmania 7-14.5 kya through the formation of the Arafura Sea and Bass Strait (Figure 1). These events still appear to be part of the oral tradition of several Aboriginal Australian communities. Similarly, environmental variation accentuated during the last glacial maximum (LGM) 19-26.5 kya, leading to greater desertification of Australia and more challenging temperature gradients, appears to have had an impact on the number and density of human populations. In the same context, morphological and physiological studies find that Aboriginal Australians living in the desert areas today have unique adaptations, such as the absence of the increased metabolic rates observed in Europeans when exposed to the freezing night temperatures common in the desert.

At the time of European contact, Aboriginal Australians spoke over 250 distinct languages, two-thirds of which belong to the Pama-Nyungan family. The place of origin of this language family, which covers 90% of the Australian mainland, has been debated, as has the effect of its extensive diffusion on its internal phylogenetic structure. The pronounced similarity among Pama-Nyungan languages, together with shared socio-cultural patterns, have been interpreted as the result of a recent, mid-Holocene, expansion. Other changes in the mid-late Holocene (~4 kya) include the efflorescence of backed blades (microliths) and the introduction of the dingo. The spatial distribution of microliths.
roughly correlates with the Pama-Nyungan languages. It has even been suggested that Pama-Nyungan languages, dingoes and backed blades all reflect a recent migration into Australia\cite{38}. Although an external origin for backed blades has been rejected\cite{36}, dingoes were certainly introduced, most likely via island south-east Asia\cite{37}. Rock art traditions also suggest contact between Sulawesi (Indonesia) and Australia\cite{38}. Intriguingly, a recent genetic study found evidence of Indian gene flow into Australia at the approximate time of these Holocene changes\cite{39}. Finally, substantial contact with Asians and Europeans is well documented in historical times\cite{40-43}, suggesting potentially complex admixture among present-day Aboriginal Australians.

After a century of research, the origins and evolutionary history of Aboriginal Australians continue to be debated. To date, only three whole genome sequences have been described - one deriving from a historical tuft of hair from Western Desert Australia\cite{20} and two others from cell lines with limited provenance information\cite{44}. In this study we report the first extensive investigation of Aboriginal Australian genomic diversity by reporting and analysing the high-coverage genomes of 83 Pama-Nyungan-speaking Aboriginal Australians and 25 Highland Papuans.

**Dataset**

We collected saliva samples for DNA sequencing in collaboration with Aboriginal Australian communities and individuals in Australia (S01). We sequenced high-depth genomes (average depth of 60X, range 20X-100X) from 83 Aboriginal Australian individuals representing a wide geographical distribution and a broad range of Pama-Nyungan languages (Figure 1, Extended Data Table 1, S02, S03, S04). Additionally, we sequenced 25 Highland Papuan genomes (38X-53X; S03) from five linguistic groups, and generated genotype data for 45 additional Papuans living or originating in the highlands (Figure 1). These datasets were combined with previously published genomes and SNP-chip genotype data, including Aboriginal Australian data from Arnhem Land and from a human diversity cell line panel from the European Collection of Cell Cultures\cite{44} (ECCAC, Figure 1, S04).

We explored the extent of admixture in the Aboriginal Australian autosomal gene pool by estimating ancestry proportions with an approach based on sparse nonnegative matrix factorization (sNMF)\cite{45}. We found that the genomic diversity of Aboriginal Australian populations is best modelled by a mixture of four main different genetic ancestries that can be assigned to four geographic regions based on their
relative frequencies: Europe, East Asia, New Guinea and Australia (Figure 2, Extended Data Figure 1, S05). The degree of admixture varies among groups (S05) with the Ngaanyatjarra speakers from central Australia (WCD) having a significantly higher “Aboriginal Australian component” (median value = 0.95) in their genomes compared to the median value of other Aboriginal Australian groups (median value = 0.64; Mann-Whitney rank sum test, one tail p-value = 3.55e-07). The “East Asian” and “Papuan” components are mostly present in northeastern Aboriginal Australian populations (Figure 2b, Extended Data Figure 1, S05), while the “European component” is widely distributed across groups. In most of the subsequent analyses, we either selected specific samples or groups according to their level of Aboriginal Australian ancestry, or masked the data for the non-Aboriginal Australian ancestry genomic component (S06).

**Colonisation of Sahul and diversification of Australians and Papuans**

The origins of Aboriginal Australians is a source of much debate, as are the nature of the relationships among Aboriginal Australians and between Aboriginal Australians and Papuans. Using $f_3$ statistics, estimates of genomic ancestry proportions and classical multi-dimensional scaling (MDS) analyses, we find that Aboriginal Australians and Papuans are closer to each other than to any other present-day worldwide population included in our study (Figure 2a, Figure 3a, S05). This is consistent with Aboriginal Australians and Papuans being derived from a common ancestral population, which initially colonised Sahul. Moreover, comparing outgroup $f_3$ statistics we do not find significant differences between Papuan populations (highland Papuan groups and HGDP-Papuans) in their genetic affinities to Aboriginal Australians (Figure 3b), suggesting that the Papuan groups share a common ancestor after or at the same time as the divergence between Aboriginal Australians and Papuans.

To investigate the number of founding waves into Australia, we contrasted alternative models of settlement history through a composite likelihood method that compares the observed joint Site Frequency Spectrum (SFS) to that predicted under specific demographic models ($^{46,47}$Figure 4a, S07). We compared the HGDP-Papuans to four Aboriginal Australian populations with low levels of European admixture (Extended Data Figure 1) from both northeastern (CAI and WPA) and southwestern (WON and WCD) Australia. We compared one and two-wave models where each Australian region was either colonized independently, or by descendants of a single Australian founding population after its divergence from Papuans. The one-wave model resulted in a better fit to
the observed SFS, suggesting that the ancestors of the sampled Aboriginal Australians diverged from a single ancestral population. This scenario is also supported by MDS analyses, even when masking Eurasian tracts, as well as by estimation of ancestry proportion analyses where all Aboriginal Australians form a cluster distinct from the Papuan populations (Figure 2, S05). Additionally, it is supported by $f_3$ analyses where all Aboriginal Australians are largely equidistant from Papuans when adjusting for recent admixture (Figure 3c). Thus, our results based on 83 Pama-Nyungan speakers, do not support earlier claims of multiple ancestral migrations into Australia giving rise to contemporary Aboriginal Australian diversity$^{9-11}$.

The SFS analysis suggests that there was a bottleneck in the ancestral Australo-Papuan population ~50 kya (95% CI 35-54 kya, S07), which overlaps with archaeological evidence for the earliest occupation of both Sunda and Sahul, between 47.5-55 kya$^{4,5,48}$. We further infer that the ancestors of Pama-Nyungan speakers and Highland Papuans diverged ~37 kya (95% CI 25-40 kya, Figure 4a, S07), which is in close agreement with results of an MSMC analysis (Figure 4b, S08), a method estimating cross coalescence rates between pairs of populations based on individuals’ haplotypes$^{49}$. It is also in agreement with previous estimates based on SNP array data$^{39}$and the distribution of *Helicobacter pylori* strains$^{50}$. These results imply that the divergence between sampled Papuans and Aboriginal Australians is older than the disappearance of the land bridge between New Guinea and Australia about 8 kya, and suggest ancient genetic structure in Sahul. Such structure may be related to palaeoenvironmental changes leading up to the onset of the LGM. Sedimentary studies show that the vast Lake Carpentaria (500 x 250 km, Figure 1) began to form ~40 kya, when sea-levels fell below the 53m-deep Arafura Sill$^{51}$. Therefore, although Australia and New Guinea remained connected until the early Holocene, the flooding of the Carpentaria basin and its increasing salinity$^{51}$ may have promoted population isolation.

**Archaic admixture**

We characterised the number, timing and intensity of archaic gene flow events using three complementary approaches: SFS-based (Figure 4a, Figure 5c, S07), a goodness-of-fit analysis combining D-statistics (S09), and a method that directly infers putatively derived archaic ‘haplotypes’ (S11). Aboriginal Australians and Papuan genomes show an excess of putative Denisovan-derived
variants (Extended Data Figure 2d, S10), as well as substantially more putative Denisovan-derived haplotypes (PDH) than other non-Africans (Extended Data Figure 3). The number and total length of those putative haplotypes varied considerably across samples. However, the estimated number of PDH correlates almost perfectly ($r^2 = 0.96$) with the estimated proportion of Australo-Papuan ancestry in each individual (Extended Data Figure 3). We also estimated that the values of $F_{ST}$ between autosomal SNPs or PDHs assigned to WCD and Papuans were both around 0.12. Moreover, we found no significant difference in the distribution of the number of PDHs or the average length of PDHs between putatively unadmixed Australians and Papuans (Mann-Whitney U test, $p>0.05$). Taken together, these observations provide strong evidence for a single Denisovan admixture event that predates the population split between Australians and Papuans (see also 52) and widespread recent Eurasian admixture in Aboriginal Australians (Figure 2, S05). Furthermore, using the SFS-based approach and constraining Denisovan admixture to have occurred before the Aboriginal Australian-Papuan divergence results in an admixture estimate of ~4% (95% CI 3-5%, Figure 5c, S07), similar to the estimates using D-statistics (~5%, S09). The SFS analyses further suggest that Denisovan/Australo-Papuan admixture took place ~44 kya (95% CI 31-50 kya, S07). We note that the point estimate for the age of the bottleneck overlaps with the confidence interval for the age of admixture, and that a bottleneck could have occurred anywhere along the dispersal route of Australo-Papuan populations from the ancestral source.

The SFS analysis also provides evidence for a primary Neanderthal admixture event (~2%, 95% CI 1-3%, Figure 5c, S07) taking place in the ancestral population of all non-Africans ~60 kya (95% CI 55-84 kya, Figure 5c, S07). Note that, although we cannot estimate absolute dates of archaic admixture from the lengths of PDHs and putative Neanderthal-derived haplotypes (PNHs), we can obtain a relative date. We found that for 20 putatively unadmixed Australians and 12 putatively unadmixed HGDP-Papuans, the average PNH length is 33.8 Kb and the average PDH length is 37.4 Kb. These are significantly different from each other ($p = 9.65 \times 10^{-6}$ using a conservative sign test), and suggest that the time since Neanderthal admixture was roughly 11% greater than the time since Denisovan admixture roughly in line with our SFS based estimates for Denisovan pulse (31-50 kya) versus the primary pulse of Neanderthal admixture (55-84 kya). The SFS analysis also suggests that the main Neanderthal pulse was followed by a further 1% (95% CI: 0.2-2.7%, Figure 5c, S07) pulse of
Neanderthal gene flow into the ancestors of Eurasians, and a smaller pulse into the ancestors of Asians (0.2%, 95% CI 0.1-1.0%, Figure 5c, S07), but there is little evidence for Neanderthal introgression private to Australo-Papuans, potentially limited to ~0.2% (95% CI 0.05-1.3%, Figure 5c, S07). In addition, the fact that the number of Neanderthal-specific introgressed sites increases from Europe to Australia (Extended Data Figure 2d, S10), and then decreases in Amerindians is consistent with recurrent Neanderthal (or Neanderthal-related archaic) gene flow during the waves of expansion into Eurasia. Our results are thus indicative of several pulses of Neanderthal gene flow into modern humans, as inferred previously\textsuperscript{53-55}. Note however, the apparent high levels in Neanderthal-specific introgressed sites in Australo-Papuans can be explained by the expected number of misclassified Neanderthal introgressed sites resulting from the shared ancestry of these two archaic hominins (S10).

Finally, using our SFS and haplotype based approaches, we explored additional models involving complex structure among the archaic populations. We found suggestive evidence that the archaic contribution could be more complex than a model involving discrete Denisovan and Neanderthal admixture pulses\textsuperscript{20,21} (S07, S11), supporting the view that the archaic contribution in Australo-Papuans is likely more complex than was previously assumed\textsuperscript{20,21} (S07).

**Out of Africa**

To investigate the relationship of Australo-Papuan ancestors to other world populations, we computed D-statistics\textsuperscript{56,57} of the form \((H1=\text{Aboriginal Australian}, H2=\text{Eurasian}), H3=\text{African}\) and \((H1=\text{Aboriginal Australian}, H2=\text{Eurasian}), H3=\text{Ust’-Ishim}\). Several of these were significantly positive (S09), suggesting that Africans and Ust’-Ishim – a ~45 kya modern human from Asia\textsuperscript{58} - are both closer to Eurasians than to Aboriginal Australians. These findings are in agreement with a model of Eurasians and Australo-Papuan ancestors dispersing from Africa in two independent waves. However, when correcting for a moderate amount of Denisovan admixture, Aboriginal Australians and Eurasians become equally close to Ust’-Ishim, as expected in a single OoA scenario (S09). Similarly, the D-statistics for \((H1=\text{Aboriginal Australian}, H2=\text{Eurasian}), H3=\text{African}\) becomes much smaller after correcting for Denisovan admixture. Additionally, a goodness-of-fit approach combining D-statistics across worldwide populations indicates stronger support for two waves OoA, but when taking Denisovan admixture into account, a one-wave scenario fits the observed D-statistics equally well (Figure 5a, S09).
To further investigate the timing and number of OoA events giving rise to present-day Australo-Papuan and Eurasians (Sardinians and Han Chinese) we used the observed SFS in a model based composite likelihood framework. When considering only modern human genomes, we find evidence for two waves OoA, with a dispersal of Australo-Papuans ~14 ky before Eurasians (S07). However, when explicitly taking into account archaic Neanderthal and Denisovan introgression into modern humans\textsuperscript{44,59}, the SFS analysis supports a single origin for the OoA populations marked by a bottleneck ~72 kya (95% CI 60-104 kya, S07). This scenario is reinforced by the observation that the ancestors of Australo-Papuan and Eurasians share a Neanderthal admixture event (95% CI 1.1-3.5%). Our analyses suggest that this single OoA ancestral population underwent two expansions at approximately the same time: one involving the ancestors of Australo-Papuan (51-72 kya) and the other, possibly slightly more recent, involving the ancestors of Eurasians (48-68 kya) (Figure 5c). Furthermore, modern humans have both an LD decay rate and a number of predicted deleterious homozygous mutations (recessive genetic load) that correlates with distance from Africa (S05, S10, and Extended Data Figure 2 a-c), again consistent with a single African origin. Aboriginal Australians also show levels of recessive load and LD that are intermediate between East Asians and Amerindians as expected if they all derive from the same OoA dispersal event.

The model estimated from the SFS analysis also suggests an early divergence of Australo-Papuans from the ancestors of all non-Africans, in agreement with two colonisation waves across Asia\textsuperscript{20,21,39}. Under our best model, Australo-Papuans began to diverge from Eurasians ~58 kya (95% CI 51-72 kya, Figure5c, S07), whereas Europeans and East Asians diverged from each other ~42 kya (95% CI 29-55 kya, Figure5c, S07) in agreement with previous estimates\textsuperscript{19,39,60,61}. We find evidence for high levels of gene flow between the ancestors of Eurasians and Australo-Papuans, suggesting that, after the fragmentation of the OoA population (“Ghost” in Figure 5c) 57-58 kya, the groups remained in close geographical proximity for some time before Australo-Papuan ancestors dispersed eastwards. Furthermore, our results show multiple gene flow events between sub-Saharan Africans and Western Eurasians after ~42 kya. This supports previous findings of extensive contact between African and non-African populations\textsuperscript{60-62}.

Our MSMC analyses suggest that the Yoruba/Australo-Papuans and the Yoruba/Eurasians cross-coalescence rates are distinct, implying that the Yoruba and Eurasian gene trees across the genome
have on average a more recent common ancestor (Figure 5b, S08). We show through simulations that these differences cannot be explained by archaic admixture. Moreover, the expected difference in phasing quality is not sufficient to fully explain this pattern either (see S08). While a similar separation in cross coalescence rate curves is obtained when comparing Eurasians or Australo-Papuans with Dinka, we find that, when comparing the Australo-Papuans or the Eurasians with the San, the cross coalescence curves are overlapping (S08). We also find that the change in effective population size through time of Aboriginal Australians, Papuans, and East Asians is very similar until around 50 kya, including a deep bottleneck around 60 kya (Extended Data Figure 7). Taken together, these MSMC results suggest complex population structure in Africa preceding a split of a single non-African ancestral population, combined with gene flow between the ancestors of Yoruba or Dinka (but not San) and the ancestors of Eurasians, which is not shared with Australo-Papuans. These results are qualitatively in line with the SFS-based analyses (see e.g., Figure 5b).

**Genetic structure of Aboriginal Australians**

Uniparental haplogroup diversity in this dataset (Extended Data Table 1, S12) is consistent with previous studies of mitochondrial DNA (mtDNA) and Y chromosome variation in Australia and Oceania, including the presence of typically European, Southeast and East Asian lineages\(^{63-68}\). The combined results provide important insights into the social structure of Aboriginal Australian societies. Aboriginal Australian groups exhibit greater between-group variation for mtDNA (16.8%) than for the Y chromosome (11.3%), in contrast to the pattern for most human populations\(^{69,70}\). This result suggests higher levels of male than female migration between Aboriginal Australian groups and may reflect the complex marriage and post-marital residence patterns among Pama-Nyungan Australian groups\(^{71}\). Moreover, the inferred European ancestry for the Y chromosome is much greater than that for mtDNA (31.8% vs. 2.4%), reflecting male-biased European gene flow into Aboriginal Australian groups during the colonial era.

Based on the genome sequences, we find genetic relationships within Australia that mirror geography, with a significant correlation \((r_{\text{GEN,GEO}} = 0.59, \text{p-value} < 0.0005)\) when comparing the first two dimensions in an MDS analysis (S14). This correlation is higher when genomic regions of putative
recent European and East Asian (i.e., Han Chinese) origin are “masked” ($r_{GEN, GEO} = 0.77$, p-value < 0.0005, Extended Data Figure 5). The main axis of genetic differentiation in the masked Aboriginal Australian genomes was determined using the Bearing correlogram approach. We found that an axis of angle = 65° compared to the equator (i.e., in the southwest to northeast direction) explains most of the genetic differentiation (S14).

Populations from the centre of the continent occupy positions genetically intermediate to this axis (Extended Data Figure 5). A similar result is observed with an $F_{ST}$-based tree for the masked data (Figure 6a, S05) as well as in analyses of genetic affinity based on the $f_3$ statistic (Figure 3b), suggesting a population division between northeastern and southwestern groups. Such structure is further supported by the SFS analyses showing that populations from southwestern desert and northeastern regions diverged as early as ~31 kya (95% CI 10-32 kya), followed by limited gene flow (estimated $2Nm$<0.01, 95% CI $2<Nm$< 11.25). The analysis of the major routes of gene flow within the continent supports the idea that the Australian interior has acted as a barrier to gene flow. Indeed, using a model inspired by principles of electrical engineering where gene flow is represented as a current flowing through the Australian continent and observed $F_{ST}$ values are a measure of connectivity, we find that gene flow occurred preferentially along the coasts of Australia (Extended Data Figure 6, S14). These findings are consistent with a model of expansion followed by population fragmentation when and the extreme aridity in the interior of Australia formed barriers to population movements during the LGM.

We used MSMC based on autosomal data and mtDNA Bayesian Skyline Plots (BSP) to estimate changes in effective population sizes within Australia. The MSMC analyses show evidence of a population expansion starting ~10 kya in the northeast, while both MSMC and BSP suggest a bottleneck in the southwestern desert populations taking place during the past ~10 kya (Extended Data Figure 7, S08, S12). This is consistent with archaeological evidence for a population expansion associated with significant changes in socio-economic and subsistence strategies in the Holocene.

European admixture almost certainly had not occurred before the late 18th century, but earlier East Asian and/or Papuan gene flow into Australia could have taken place. We characterized the mode and tempo of gene flow into Aboriginal Australians using three different approaches (S06, S07, S13). We used approximate Bayesian computation (ABC) to compare the observed mean and variance among
Aboriginal Australian individuals in the proportion of European, East Asian and Papuan admixture, to
that computed from simulated datasets under various models of gene flow. We estimated the European
and East Asian admixture to have occurred on the order of ten generations ago (S13), consistent with
historical and ethnographic records. Consistent with this, the local ancestry approach based on RFMix
suggests that the European and East Asian admixture is more recent than the Papuan admixture
(Extended Data Figure 4a). In addition, both the ABC and SFS analyses suggest that the best fitting
model for the Aboriginal Australian-Papuan data is one of continuous but modest gene flow, mostly
unidirectional from Papuans to Aboriginal Australians, and geographically restricted to northeast
Aboriginal Australians (2Nm=0.4, 95% CI 0.0-20.4, Figure 4a, S07).

To further investigate Papuan gene flow, we conducted follow-up analyses on the Papuan ancestry
tracts obtained from the local ancestry analysis. We inferred local ancestry as the result of admixture
between four components: European, East Asian, Papuans and Aboriginal Australian (S06). We chose
WCD as the representative of Aboriginal Australian ancestry, because it is the least admixed
population among our Australian samples (Figure 2, S05). Papuan tract length distribution show a clear
geographic pattern, with “younger tracts” (higher median length and variance) in individuals closer to
New Guinea and “older” (lower median length and variance) in individuals closer to WCD (Extended
Data Figure 4b); there is a strong correlation of Papuan tract length variance with distance from WCD
to other Aboriginal Australian groups (r=0.64, p-value<0.0001). The prevalence of short ancestry tracts
of Papuan origin, compared to longer tracts of East Asian and European origin, suggests that a large
fraction of the Papuan gene flow is much older than that from Europe and Asia, which is consistent
with the ABC analysis (S13). We also investigated possible South Asian (Indian related) gene flow into
Aboriginal Australian, as reported by a recent study

To investigate if linguistic relationships reflect genetic relationships among Aboriginal Australian
populations, we built a Bayesian phylogenetic tree for the 28 different Pama-Nyungun languages

**Pama-Nyungun languages and genetic structure**

To investigate if linguistic relationships reflect genetic relationships among Aboriginal Australian
populations, we built a Bayesian phylogenetic tree for the 28 different Pama-Nyungun languages
represented in this sample (Figure 6b, S15). The linguistic and \( F_{ST} \)-based genetic trees obtained (Figure 6) share several well-supported partitions. For example, both trees indicate that the northeastern (CAI and WPA), and southwestern groups (ENY, NGA, WCD and WON) each form a cluster, while PIL, BDV and RIV are found between them. A distance matrix between pairs of languages, computed from the language-based tree, is significantly correlated with geographic distances \( (r_{GEO, LAN} = 0.83, \) Mantel test two-tail p-value on 9,999 permutations = 0.0001). This suggests that differentiation among Pama-Nyungan languages in Australia follows geographic patterns, as observed in other language families elsewhere in the world\(^{15,76}\). Furthermore, we find a correlation between linguistics and genetics \( (r_{GEN, LAN} = 0.43, \) Mantel test p-value < 0.0005) that remains significant when controlling for geography \( (r_{GEN, LAN, GEO} = 0.26, \) Mantel test p-value < 0.0005). This is consistent with language differentiation after populations lose (genetic) contact with one another\(^{77}\). The correlation between the linguistic and genetic trees is all the more striking given the difference in time scales: the Pama-Nyungan family is generally accepted to have diversified within the last 6 ky\(^{78}\), while the genetic estimates are two to five times that age. The linguistic tree thus cannot simply reflect initial population dispersals, but rather reflects a genetic structure that has a complex history, with initial differentiation 10-32 kya, localised population expansions (northeast) and bottlenecks (southwest) ~10 kya, and subsequent limited gene flow from the northeast to the southwest. The latter may be the genetic signature that tracks the divergence of the Pama-Nyungan language family.

**Selection in Aboriginal Australians**

To identify any selection specific to Aboriginal Australians, we used two different methods based on the identification of SNPs with high allele frequency differences between Aboriginal Australians and other groups, similar to the often used Population-Branch Statistics\(^{79}\) (PBS, S16). First, we scanned the Aboriginal Australian genomes for loci with an unusually large change in allele frequencies since the divergence from Papuans, taking recent admixture with Europeans and Asians into account. Among the top ranked genomic regions (Extended Data Table 2), we identified candidate loci that might be related to cold tolerance and dehydration resistance. One peak of high differentiation (the 7\(^{th}\) highest peak) is located near the \( NETO1 \) gene, which harbours alleles that have previously been shown to be associated with thyroid hormone levels. Interestingly, it has been suggested that thyroid hormone levels are associated with Aboriginal Australian specific adaptations to desert cold\(^{80}\). We investigated this
potential thermoregulatory adaptation further by identifying genomic regions showing high
differentiation associated with different ecological regions in Australia (S16). The top candidate gene
in this scan is KCNJ2, encoding a potassium channel protein harbouring alleles associated with
thyrotoxic periodic paralysis\textsuperscript{81}. 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\textsuperscript{80}.

Another locus of interest close to the 8\textsuperscript{th} highest peak of differentiation, SLC2A12, is associated with
serum urate levels\textsuperscript{82}. The pathophysiology of dehydration includes elevated serum urate levels.
Therefore, these results are suggestive of a locus that may be involved in tolerance to dehydration in
Aboriginal Australians. Although further studies are needed to associate putative selected genetic
variants in Aboriginal Australians with specific phenotypic effects, the current selection scan provides
candidate genes for such future efforts.

**Discussion**

Our findings shed light, but also raise new questions, concerning on the population history of
Aboriginal Australians. They suggest an early population structure in Sahul likely dating back ~37 kya
(25-40 kya), when the ancestors of Highland Papuans and Pama-Nyungan Aboriginal Australians
diversified. Intriguingly, despite this, our results also indicate that the population that diverged from
Papuans was the ancestor of all the Aboriginal Australian groups sampled in this study; yet,
archaeological evidence shows that by 40-45 kya, humans were widespread within Australia (Figure 1).
Three non-exclusive demographic scenarios can account for this observation: 1) the Aboriginal
Australian ancestral population prior to the divergence from Papuans was widespread, maintaining
gene flow across the continent; 2) it was deeply structured, and only one group among the early settlers
survived to give rise to Aboriginal Australians; and 3) other groups survived, but the descendants are
not represented in our sample. Additional modern genomes, especially from Tasmania and the Non-
Pama-Nyungan regions of the Northern Territory and Kimberley (both regions highly distinct
linguistically \textsuperscript{83} and not represented in our sample), as well as ancient genomes pre-dating European
contact in Australia and other expansions across South East Asia\textsuperscript{38}, should help resolve these questions
in the future.
To add to this already complex picture, our estimates of ~44 kya (31-50 kya) for the time of admixture between the Australo-Papuan ancestors and an archaic hominin distantly related to Denisovans are very young. In the absence of paleontological evidence that archaic hominins crossed the Wallace Line, combined with evidence of much lower levels of Denisovan ancestry across East Asia and the Americas\(^{52,88}\), it is likely that the admixture occurred in Southeast Asia or even further to the west, constraining the age when the ancestors of living Australo-Papuan colonised Sahul and/or the actual timing of Denisovan admixture. In this context, it is noteworthy that our SFS based time estimates relies on the use of recently suggested molecular clock (1.25\(\times\)10\(^{-8}\), see\(^{84}\)) and generation time for humans (29 years\(^{85}\)). Should any of these parameters change, our genetic-based time estimates will need revisions too.

Interestingly, our results also show that southwestern and northeastern Pama-Nyungan populations diverged 10-32kya. Together with the evidence for selection in genes that may have provided an advantage in extreme desert environments, such as those experienced in Western Desert populations during the LGM, these results point to a long-standing genetic structure among Pama-Nyungan Aboriginal Australians that survived post-glacial demographic changes. In other parts of the world, including South East Asia, Pleistocene demographic patterns were overlaid by post-glacial and Holocene expansions that left both genetic and linguistic regional signatures\(^{87}\). In Australia, the archaeological record also shows post-glacial expansions\(^{73,74}\), while the spread of Pama-Nyungan languages across the continent is generally accepted to be mid-to-late Holocene\(^{35}\). Our genetic findings indicate an early Holocene demographic expansion localized to northeast Aboriginal Australians, as well as gene flow spreading from the northeast across the continent. These observations are consistent with a possible origin and spread of the Pama-Nyungan languages from the northeast of Australia to the rest of the continent. Thus, evidence from genetics may add to the linguistic and cultural evidence - such as the spread of large ceremonial gatherings, trade and exchange intensification, broad alliance networks, cross-group male ritual induction, new plant foods, among several others\(^{35}\) – that the dispersal of Pama-Nyungan languages has been driven by both cultural diffusion and demic expansion.

**Data access**

The whole genome sequence data and SNP array data generated in this study are available upon request from E.W (ewillerslev@snm.ku.dk) and D.M.L. (d.lambert@griffith.edu.au). The Papuan whole
genome sequence data generated in this study are also available under managed access through the EGA database (https://www.ebi.ac.uk/ega) under study accession number EGAS00001001247.

**References for the main text**


44. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505, 43–49 (2014).


51. Holt, S. Palaeoenvironments of the Gulf of Carpentaria from the last glacial maximum to the present, as determined by foraminiferal assemblages. (2005).


78. Evans, N. & Jones, R. in *Archaeology and linguistics: aboriginal Australia in global perspective* (Oxford University Press Australia, 1997).


Supplementary Information (see annex)

S01 Ethical approvals in relation to sampling in Australia

S02 Ethnography and linguistics for the Aboriginal Australian individuals

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

S04 Reference panels, relatedness and runs of homozygosity

S05 Linkage disequilibrium (LD) and population structure within Australia

S06 Local ancestry

S07 Demographic inferences

S08 MSMC analysis

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

S10 Mutation load analysis
S11 Archaic gene flow
S12 Uniparental markers
S13 ABC analysis to characterize recent European, East Asian and Papuan gene flow
S14 Spatial analyses
S15 Computational phylogenetics: Pama-Nyungan languages
S16 Scan for positive selection