Chapter 2: The genetic history of Australia

2.1 Introduction

The ancient continent of Sahul encompassed the present-day landmasses of Australia, New Guinea and Tasmania, which were separated by rising sea levels only ∼8 kya (Woodroffe, et al. 2000; Lewis, et al. 2013). There is strong evidence of human occupation in Australia and elsewhere in Sahul dating back to ∼50 kya (Clarkson, et al. 2015; O'Connell and Allen 2015; Veth, et al. 2017) implying humans arrived fairly soon after migrating out of Africa. While sea levels were lower at this time, reaching Sahul through Island Southeast Asia would still have required several long sea crossings. Key questions in the population history of this part of the world include: Are the present-day indigenous populations the descendants of these first settlers of the continent? Have there been additional migrations to the continent after that? What is the relationship, including genetic divergence time, between the populations of Sahul and those in the rest of the world? What is the relationship between Aboriginal Australians and Papua New Guineans? This chapter addresses these questions, many of which relate to the shared early history of Aboriginal Australians and Papua New Guineans, or "Australo-Papuans", as well as questions specific to the history of Aboriginal Australians. The next chapter focuses on Papua New Guinea.

There are two key components to the question of the relationship between Australo-Papuans and other human populations. The first is whether Australo-Papuans derive ancestry from a separate, perhaps earlier, migration out of Africa than the one giving rise to Eurasians. The second is, given at least some shared ancestry between Australo-Papuans and Eurasians, which lineage was the first to branch off from the others: Europeans (such that Australo-Papuans and East Asians are sister clades) or Australo-Papuans (such that Europeans and East Asians are sister clades)? Most genetic studies have placed Australo-Papuan populations within non-African variation, and furthermore as a sister group to East Asians, in clustering-type analyses (Li, et al. 2008; Hugo Pan-Asian SNP Consortium, et al. 2009), although if only a small proportion of their ancestry is from an earlier out-of-Africa event such analyses might not be sensitive to that. A study on the first Aboriginal Australian whole-genome sequence suggested that Australo-Papuans were an outgroup to Europeans and East Asians, diverging from them 62 to 75 kya, followed by some post-divergence gene flow between East Asians and Australo-Papuans (Rasmussen, et al. 2011). While this study is sometimes cited as also supporting an earlier exit from Africa for Australo-Papuan ancestors, it does not actually make this claim. On the question of later migration and gene flow into Australia, a number of hypotheses have been put forth. There are changes in the archaeological record of Australia starting around 4-6 kya, with the appearance of certain stone tools as well as the dingo, a wild dog (Brown 2013). While the dingo most certainly arrived with the help of humans, there is debate about the origin of the cultural changes. An early study of mitochondrial genomes suggested Aboriginal Australians had a closer relationships to southern Indian than to Papua New

Guinean populations (Redd and Stoneking 1999). Another study reported that Aboriginal Australian Y chromosomes in haplogroup C, which represents 44% of indigenous chromosomes in Australia (Nagle, et al. 2016), shared a common ancestor with chromosomes in southern India and Sri Lanka 195 generations ago (95% confidence interval = 49–532 generations), or 5,655 years assuming a generation interval time of 29 years (the study itself used 25 years) (Redd, et al. 2002). A later study based on genome-wide array genotyping of a population from northern Australia reported autosomal evidence for South Asian gene flow, estimating that 11% of Aboriginal Australian ancestry derives from sources related to present-day Dravidian speaking groups of Southern India, and dating this admixture to 4,230 ya (Pugach, et al. 2013). These studies suggested that this genetic ancestry arrived in Australia together with the cultural changes. Another study of uniparental markers however found no support for any South Asian gene flow to Australia (Hudjashov, et al. 2007).

Another debated topic in the history of Australia concerns the spread of the dominant pre-colonial language family of the continent, the Pama-Nyungan languages. These languages are spoken across 90% of Australia, absent only from the very northern parts, and are estimated on the basis of linguistic similarities to have spread during the Holocene, perhaps in the last ∼4-7 ky (Mcconvell 1996; Malaspinas, et al. 2016). No present-day languages outside of Australia are related to them, and while an internal expansion thus seems to be the most likely explanation for its spread, no known major cultural or technological processes in the history of Australia, other than the spread of small stone tools, correspond to this timeframe (Bowern 2010). It is therefore also not known whether or not this language expansion was associated with a spread of people and therefore genes, resulting in a reshaping of population structure in Australia.

In the last approximately 200 years following European colonization of Australia, large numbers of migrants from Europe and other parts of the world have entered the continent, resulting in widespread foreign admixture into the Aboriginal Australian population (McEvoy, et al. 2010) as well as disruption of the pre-colonial population structure.

In this chapter I primarily make use of two different data sets to address these various questions. The first is a dataset of Y chromosome sequences from 13 Aboriginal Australian and 12 Papua New Guinean men, generated at the Wellcome Trust Sanger Institute and published in 2016 (Bergström, et al. 2016). The second is a dataset of 83 Aboriginal Australian and 39 (inclusive of the aforementioned 12) Papua New Guinean whole-genome sequences, the former generated by external collaborators and the latter at the Wellcome Trust Sanger Institute, which were analysed in collaboration with a large consortium and published in 2016 (Malaspinas, et al. 2016). Many other results were obtained by collaborators working on that project, and where relevant I refer to those results in the text.

2.2 Analysis of Aboriginal Australian Y chromosomes

The Y chromosome has long served as a particularly useful part of the genome for the analysis of population relationships due the lack of recombination along most of its length. It is inherited in an unbroken lineage from father to son, such that the identification of a similar Y chromosome in two different men provides highly confident evidence for shared ancestry. With sequencing data from the chromosome and a measure of the mutation rate, the time at which that ancestor lived can also be estimated. Another advantage of the Y chromosome is that, in admixed populations such as Aboriginal Australians, an indigenous chromosome remains fully indigenous along its length and does not recombine and mix with foreign chromosomes. In Australia today, ∼30% of Aboriginal Australian males carry Y chromosomes of indigenous origin (Taylor, et al. 2012).

An earlier study genotyped a set of 144 self-reported Aboriginal Australian males at known Y chromosome variants to assign them to major haplogroups (Nagle, et al. 2016). Out of these, 13 individuals with indigenous chromosomes were re-contacted and consented to further study. Five of these were from haplogroup C, constituting 44% of indigenous chromosomes; six were from haplogroup K*, constituting 56% of indigenous chromosomes; and two were from haplogroup M (a branch of K*), constituting 2% of indigenous chromosomes (Nagle, et al. 2016). These samples were whole-genome sequenced and the reads mapping to the Y chromosome were extracted. These data were then analysed jointly with sequencing data from 12 Papua New Guinean males from the HGDP-CEPH panel as well as all the 1244 males from 26 worldwide populations from the 1000 Genomes Project (1000 Genomes Project Consortium 2015), calling genotypes at approximately 10 million sites on the chromosome deemed to be suitable for short read sequencing (Poznik, et al. 2013).

A maximum likelihood phylogeny constructed from these 1269 Y chromosomes recapitulates the known structure of human Y chromosome history, and reveals how Aboriginal Australian and Papuan chromosomes fit into this (Figure 2.1). Within both the C and the K*/M clades of the tree, the Aboriginal Australian and Papuan chromosomes form monophyletic clades with high bootstrap support (100% for the C and 97% for the K*/M, respectively). This is consistent with a shared origin of Aboriginal Australians and Papuans.

By counting the number of sites that have mutated between pairs of Y chromosomes, and the number of sites that have not mutated, an estimate of their divergence time can be obtained. These estimates were scaled to units of years by applying a 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 modern human (Fu, et al. 2014). This mutation rate is similar to that estimated from present-day Icelandic patrilines (Helgason, et al. 2015). This analysis resulted in divergence estimates of 54.3 ky (95% confidence interval (CI): 48.0–61.6 ky) between Sahul K∗/M and their closest relatives in the R and Q haplogroups, 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. These dates are consistent with an

early divergence of Australo-Papuan ancestors from populations in Eurasia, and with the generally-accepted archaeological record in Sahul, and thereby with the hypothesis that present-day people are the descendants of the first arrivals on the continent.

These results thus show no evidence for more recent arrival of Y chromosomes to Australia. In particular, they confidently refute the previous claim that Australian haplogroup C chromosomes arrived from South Asia in the Holocene (Redd, et al. 2002). A bootstrap analysis across the sites used on the Y chromosome expanded the confidence intervals of the haplogroup C split date to 44.9–65.9 kya, showing that technical uncertainty arising from read mapping or variant calling is not great enough to affect the conclusion of an old split. The greatly underestimated split date reported by the earlier study can be understood in terms of the technology used at the time. It used differences at short tandem repeats and a mutation rate at these repeats of 2.08×10^{-3} per generation to date the split, but it has later been shown that such analyses tend to massively underestimate older divergence times, largely due to saturation of recurrent mutations (Wei, et al. 2013). The results presented here thus represent a clear example of how the greater power and accuracy of direct sequencing compared to earlier methods for studying genome variation can resolve uncertainties about human population history. They do not by themselves definitely prove that there was no gene flow from South Asia, as it's still possible that the autosomal genome might harbour such ancestry, but they exclude the Y chromosome piece of the puzzle.

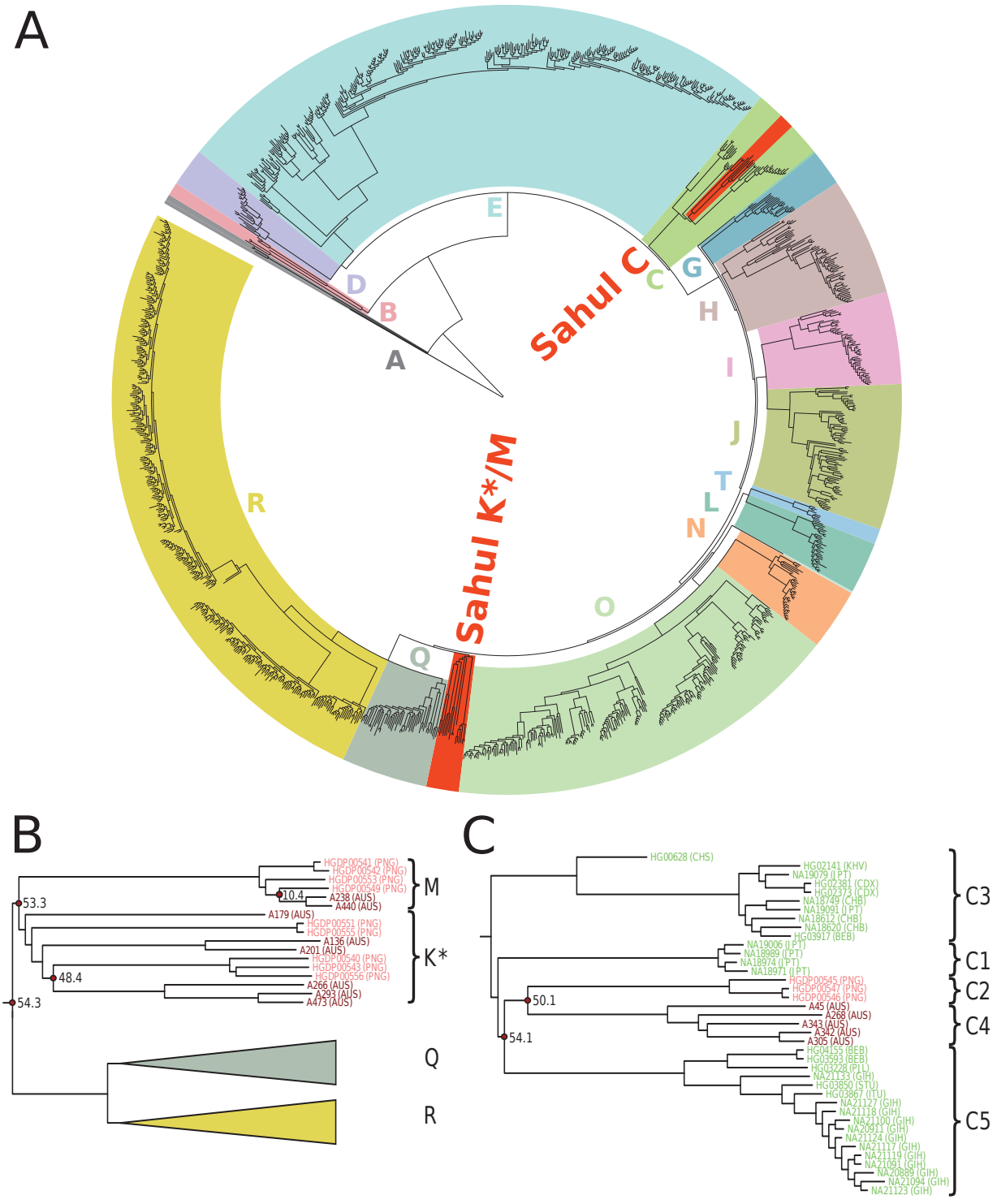

Figure 2.1: Phylogenetic History of Aboriginal Australian Y Chromosomes. *A) A maximum likelihood phylogeny of 1269 human Y chromosomes, including Aboriginal Australian and Papuan chromosomes highlighted in red, inferred using RAxML (Stamatakis 2014) from short-read sequencing data mapped to the ∼10 Mb of accessible sequence on the chromosome. High-level haplogroups are coloured and labelled along the tree. B) Detailed view of haplogroups K* and M. C) Detailed view of 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). 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.*

2.3 Foreign admixture in Australia

The dataset of 83 Aboriginal Australian whole-genome sequences from nine population samples across the continent allows foreign admixture to be studied using the autosomal genomes. The ADMIXTURE method (Alexander, et al. 2009) revealed widespread and variable non-indigenous ancestry (Figure 2.2A). Most of this ancestry appears to be of European origin, but there are also non-trivial amounts of East Asian ancestry, particularly in the north-eastern groups. The large variation in the overall amount foreign ancestry across individuals implies that the admixture is recent. Several individuals have no discernible foreign ancestry, especially those from the Western Central Desert area where all but one individual appear to have entirely indigenous ancestry. These differences between geographic regions likely reflect differences in the timing and impact of the European colonization of Australia.

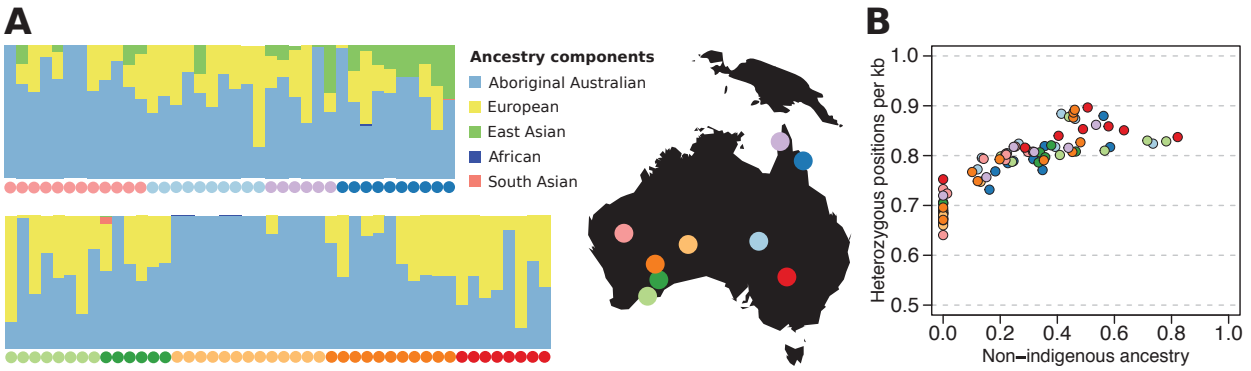

Figure 2.2: Genome-wide ancestry of Aboriginal Australian individuals. *Individuals are coloured according to their population origin, following the map in middle of the figure. A). The ADMIXTURE software was run with k=5 on Aboriginal Australians together with a worldwide panel including Europeans, East Asians, South Asians and Africans, which are not shown in the figure. At this k value, the ancestry components essentially correspond exactly to the ancestries of each of these continental groups, and are therefore labelled after them. The very small amounts of ancestry assigned to components other than the Aboriginal Australian, European and East Asian in a few individuals might just be noise in the estimation. The coloured circle under each individual denotes its population of origin, the geographical locations of which are indicated on the map to the right. B). Heterozygosity per individual versus the amount of non-indigenous ancestry. Population colours are as in A.*

The observation that several individuals lack non-indigenous ancestry is potentially important, not only because of what it says about the process of admixture since European colonization but also because such un-admixed individuals would greatly aid various analyses, especially those comparing Aboriginal Australians to worldwide populations. To further confirm the lack of admixture in these individuals, I conducted f_3 and D-statistic tests (Patterson, et al. 2012). Negative values in tests of the form f_3(Aboriginal Australian; Source A; Source B), where A and B were all possible pairs from a set of worldwide representative populations, indicate admixture in the history of the Aboriginal Australian sample from sources related to A and B. For 21 of the Aboriginal Australian individuals, none of the tests produced a negative value. In D-statistic tests of the form D(Yoruba, C; Aboriginal Australian, Papuan), where the west African Yoruba population serves as an outgroup and C was any of French, Han Chinese, Indian Gujarati, Indian Telugu, Sri Lankan Tamil, Pakistani Punjabi or Bangladeshi Bengali, 23 individuals did not show any significantly negative statistic (at $Z < -3$. At a threshold of $Z < -2$, the number was 18 individuals), meaning there is no evidence for any of these populations for C being genetically closer to Aboriginal Australians than

to Papuans. To test if the lack of evidence for admixture was due to low power in these single-sample tests, the same tests were performed on the pool of all Aboriginal Australian individuals who did not display any evidence of admixture, but there were still no negative f_3 or D-statistics. These results thus provide strong evidence that these individuals have not received any additional gene flow from outside Sahul since their separation from Papuans.

The foreign, primarily European, admixture present in most Aboriginal Australians has a large impact on the properties of their genomes and complicates many population-genetic analyses, but knowing the amount of admixture in each individual can help. As an example, genome-wide heterozygosity varies widely across these individuals, but most of this variation is driven by differences in the amount of non-indigenous ancestry (Figure 2.2B), as the pairing in a genome of two chromosomes from divergent populations results in fewer segments that are identical by descent from recent ancestors. Heterozygosity in un-admixed Aboriginal Australians tends to be about 0.007 per base pair, slightly lower than East Asians but higher than Native Americans.

2.4 Testing for South Asian gene flow to Australia

None of the above analyses give any evidence for South Asian admixture in Aboriginal Australian genomes, unlike an earlier study based on genome-wide array genotypes (Pugach, et al. 2013). In ADMIXTURE analyses performed in that study, 11% of Aboriginal Australian ancestry was assigned to a component that was absent from Papua New Guineans and maximized in South Asian populations, with very homogenous levels of this component across Aboriginal Australian individuals. A single whole-genome sequenced Aboriginal Australian analysed with the same method also received about the same level of a component not present in Papuans, which in runs with different numbers of components corresponded to ancestry present in various combinations of South Asian, East Asian, and Philippine Negrito populations (Rasmussen, et al. 2011). The latter study, however, interpreted this as likely not reflecting actual South Asian admixture, but rather an inability of the method to accurately assign the Aboriginal Australian ancestry to the available components given only a single individual. A similar result was obtained in an ADMIXTURE analysis in a separate study of worldwide human diversity that included two Aboriginal Australian genomes (Mallick, et al. 2016), but was not commented on. An earlier array genotype study of a northern Aboriginal Australian group obtained results consistent with foreign admixture being only European in origin, using the conceptually similar methods of FRAPPE and STRUCTURE (McEvoy, et al. 2010).

In order to try to understand the discrepancies in ADMIXTURE results between these various studies, I performed a down-sampling analysis, running the software on an increasing number of Aboriginal Australian individuals. These individuals were selected on the basis of not displaying

any evidence from other analyses for European or other foreign admixture, and they were run in the context of Papuans, Europeans, East Asians and South Asians. I also performed the analogous analysis but instead varying the number of Papuan genomes, to see if a similar non-Sahul component could appear in these too at small sample sizes. These analyses showed that when the number of Aboriginal Australian individuals is low (1-3), they tend to be assigned 20% ancestry from the South Asian and East Asian components, but this is no longer the case when a larger number is included (Figure 2.3A). When down-sampling the number of Papuan individuals instead, the behaviour is partly but not entirely the same: with only a single Papuan individual, 2-5% of the East Asian component is sometimes assigned to it, but already with two individuals all of the ancestry is assigned to the Sahul component. So, while the situation is not entirely symmetrical between Aboriginal Australians and Papuans, these results still strongly suggest that the assignment of South and/or East Asian components to Aboriginal Australian genomes is an artefact of this type of method. While the array genotyping study (Pugach, et al. 2013) had a sample size (12 individuals) which according to this down-sampling analysis would appear to be sufficient to avoid this artefact, it is possible that other features of the data could also make it susceptible, i.e. array marker density and/or marker ascertainment. Lastly, a reanalysis in (Malaspinas, et al. 2016) of the same array data together with the whole-genome sequenced Aboriginal Australian genomes found no South or East Asian component in similar analyses, although it did observe 20-25% of components corresponding to ancestry present in New Guinean and Melanesian island populations (Extended Data Figure 2 of that study). It is therefore possible that these 12 array genotyped individuals, who are from the northern part of Australia, might harbour admixture from such more proximal sources, and that this ancestry is contributing to artefactual behaviour in model-based ancestry assignments.

In PCA and similar analyses, Aboriginal Australians also sometimes behave in a manner that might suggest higher similarity to South Asian populations. The array genotype study in (Pugach, et al. 2013) observed how Aboriginal Australians were shifted away from the ancestral pole defined by Papua New Guinean highlanders and interpreted this as reflecting South Asian admixture. In (Malaspinas, et al. 2016), all Aboriginal Australians, even those inferred to lack European or other foreign admixture, are shifted away from New Guinean highlanders in a similar analysis (Figure 2B of that study). However, caution is needed when interpreting this, as PCA and related results are dependent in complex ways on the ancestry composition and the sample sizes of the individuals analysed jointly. To address this in a way that should get around some of those potential complications, I performed a PCA analysis where only European, East Asian and South Asian individuals contribute to the calculation of the principal components, and Aboriginal Australian and Papuans are then projected onto a position within this space that reflects their relative similarity to these three. This reveals, in a manner highly consistent with the ADMIXTURE results, that many Aboriginal Australian individuals are drawn towards the European and East Asian corners, in some cases both, relative to the position of Papuans (Figure 2.3B). However, no individuals are

pulled towards the South Asian corner.

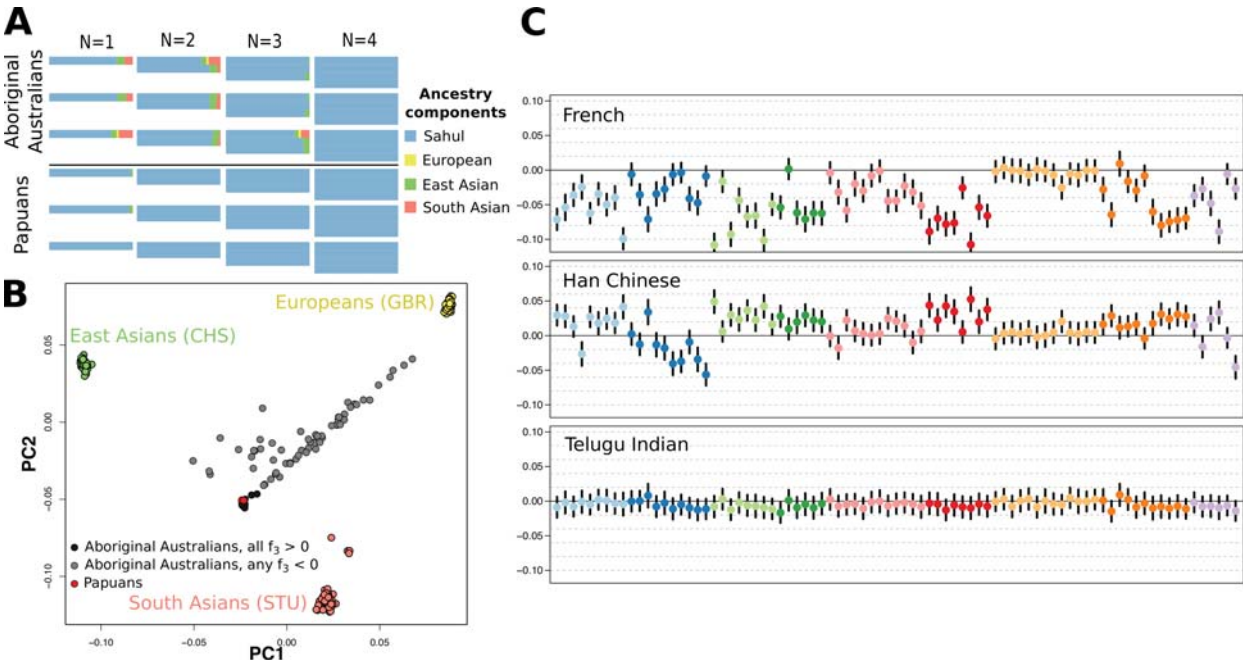

Figure 2.3: Testing for South Asian admixture in Aboriginal Australian genomes. *A) A down-sampling analysis of the effect of sample size on the assignment of non-Sahul ancestry components to Aboriginal Australian and Papuan genomes by the ADMIXTURE method. Down-sampling of the number of individuals shown at the top of each column was performed separately for Aboriginal Australians and for Papuans, and run in the context of 14 individuals from the other of these two populations, plus 30 individuals each from the British (GBR), Han Chinese (CHS) and Telugu Indian (ITU) populations. Three replicates with different sampled individuals were run for each sample size. In every run, the 14 Aboriginal Australian or Papuan individuals in the population not being down-sampled were assigned all of their ancestry to the Sahul component (not shown). B) A principal components analysis where the components were calculated using only the genotypes of British (GBR), Han Chinese (CHS) and Sri Lankan Tamil (STU) individuals, with Aboriginal Australians and Papuans then projected into the resulting space. Many Aboriginal Australian individuals are drawn towards Europeans and East Asians, relative to the position of Papuans, but none are drawn towards South Asians. Aboriginal Australians are coloured based on whether or not they have any negative values in admixture tests of the form* f_3*(Aboriginal Australian; Source A; Source B). C)* D*-statistics of the form D(Yoruba,C;Aboriginal Australian,Papuan) with C being French, Han Chinese or Telugu, displayed for each Aboriginal Australian individual separately, coloured by population as in Figure 2.2. Vertical lines correspond to ±3 standard errors.*

Finally, the formal f_3 and *D*-statistics tests reported in the previous section showed that there are several Aboriginal Australian individuals who display no evidence of admixture from any source, including South Asian populations. It would still possible that some of the individuals who carry e.g. European admixture could carry South Asian admixture too, with the former masking the effects of the latter on the test statistics. However, analysis of the patterns of *D*-statistics reveals that this is not the case – even those individuals with strong signals for other foreign admixture do not display signals for South Asian admixture (Figure 2.3C).

In summary, there does not appear to be any solid evidence for pre-historic South Asian gene flow to Australia, and previous reports of autosomal admixture most likely reflect technical artefacts.

2.5 Archaic ancestry in Sahul

Previous studies have found high levels of Denisovan ancestry in both Aboriginal Australian and Papuan individuals (Reich, et al. 2011; Prufer, et al. 2014), but these have been limited to small

numbers of individuals that have served as representatives for the whole continent. Extending the analysis to the more geographically diverse set of populations here, there is no detectable difference in the overall amount of Denisovan affinity between different Aboriginal Australian and Papuan subpopulations (*D*-statistic tests of the form *D*(Chimp, Denisovan; X,Y), largest |Z| across tests = 2.3, excluding individuals with foreign admixture). The same result is obtained if testing for Neanderthal affinity instead (*D*-statistic of the form *D*(Chimp, Altai Neanderthal; X,Y), largest |Z| across tests = 2.52). The results are consistent with the archaic admixture occurring in the common ancestor of all people in Sahul, and mirror the largely homogenous amount of Neanderthal ancestry found in Eurasia. Although Europeans are estimated to have slightly less Neanderthal ancestry than East Asians, this is likely due the effect of later dilution due to admixture with a proposed basal Eurasian lineage lacking Neanderthal admixture (Lazaridis, et al. 2014) (though alternative explanations suggesting additional admixture events in East Asian have also been proposed (Kim and Lohmueller 2015; Vernot and Akey 2015; Vernot, et al. 2016)), as discussed in Chapter 1. The same conclusion of a homogenous level of archaic ancestry across Sahul was reached using analyses of local archaic haplotypes in (Malaspinas, et al. 2016).

2.6 Out of Africa

The Y chromosomes of Aboriginal Australians and Papuans clearly have a shared, single origin with other non-African Y chromosomes, and the same is true of the mitochondrial genomes (Nagle, et al. 2017). However, it is still very much possible that the autosomal genome contains ancestry, perhaps just a very small amount, deriving from a separate out-of-Africa migration from the one giving rise to Eurasian ancestry.

In a PCA where the components are constructed using only African genotypes, using genotype array data from the HGDP-CEPH panel (Li, et al. 2008), all non-Africans, including Papuans, project approximately into the same part of the plot (Figure 2.4). This suggests there are at least no major differences in their relationships to present-day African populations.

D-statistics allow for more formal tests of the relationships between Africans and non-Africans. Tests of the form *D*(Chimp,African;X,Y) take values that are not significantly different from 0 if X and Y are from e.g. one European and one East Asian population (Figure 2.5). There is a slight trend towards higher African sharing with Europeans, though this might plausibly be due to small amounts of backflow to Africa from populations closer to Europeans than to East Asians (or alternatively, gene flow from Africa to Europe). Overall, these results are thus compatible with a single, shared non-African origin for Eurasians. However, when performing the same test setting one of X and Y as Eurasian and the other as Aboriginal Australian or Papuan, the results indicate stronger African sharing with the former (Figure 2.5). Exchanging African for the 45,000-year-old Ust'-Ishim individual (Fu, et al. 2014), who appears to be essentially an undifferentiated Eurasian, gives the same result, e.g.: *D*(Chimp, Ust'-Ishim; Aboriginal Australian, Han Chinese) = 0.0323, Z = 5.12. These test results are compatible with Aboriginal Australians and Papuans carrying some

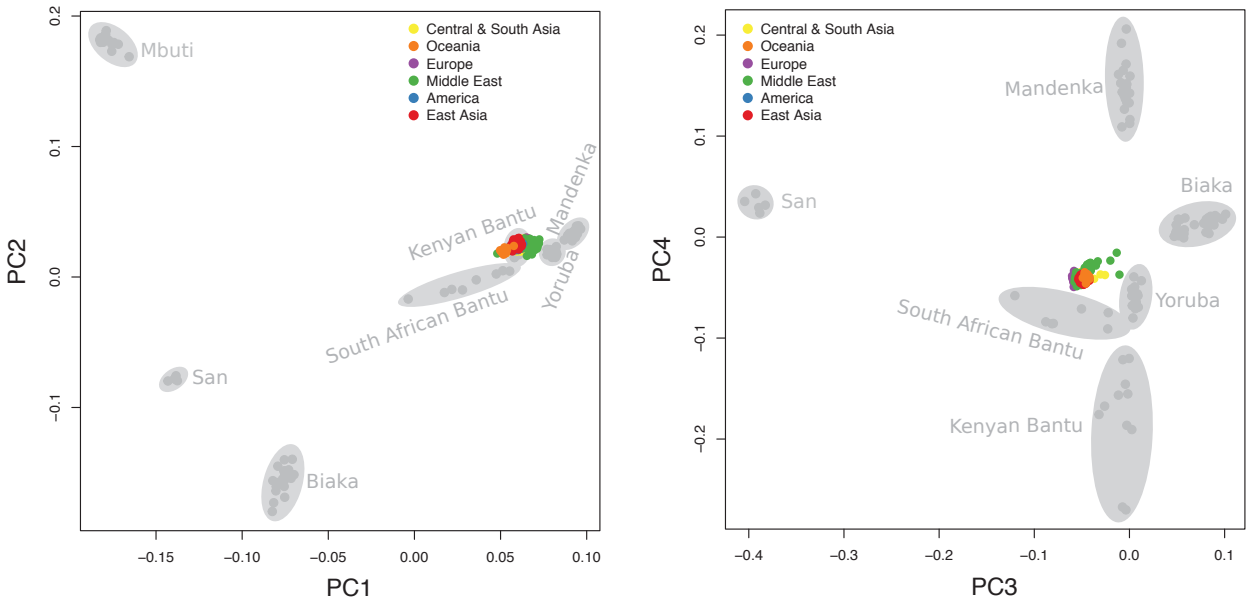

Figure 2.4: Relationships of non-Africans to Africans. *A PCA where the components were calculated only using African genotypes, and the non-Africans were projected into the resulting space, using Illumina 650K genotype array data from the HGDP-CEPH panel. Africans are displayed in grey, with population labels indicated next to the ellipse surrounding all individuals in the given population, while non-Africans are displayed in colour. The "Oceanian" label corresponds to individuals from Papua New Guinea and Bougainville Island. All non-Africans project into largely the same part of African variation, implying a shared origin. A few Middle Eastern and South Asian individuals harbour recent African admixture and depart from the main cluster of non-Africans.*

ancestry that derives from an earlier migration out of Africa, with that ancestry having lost genetic contact with Africans earlier and therefore sharing fewer African alleles relative to the ancestry that later became Eurasians.

However, *D*-statistics of the above form will also be affected by the Denisovan ancestry that is present in Aboriginal Australians and Papuans. Topologically, the Denisovan genome sits on the same branch as Chimp in this test, such that increased sharing between Denisovan and the Aboriginal Australian translates to a corresponding increased sharing between African and Eurasians. The test results thus likely reflect this, rather than ancestry from an earlier exit of modern humans from Africa (though it could theoretically reflect both of these simultaneously). This was demonstrated quantitatively in (Malaspinas, et al. 2016) by the direct calculation of a Denisova-admixture-corrected *D*-statistic, and in (Mallick, et al. 2016) by the use of admixture graphs. The uniform behaviour of different African populations in these *D*-statistics is also consistent with archaic admixture in Aboriginal Australians, but less so with earlier out-of-Africa ancestry. For example, unless the population leaving Africa earlier branched off from other Africans before the highly divergent San population did, which seems unlikely, these *D*-statistics would be expected to reach less positive values when involving the San than when involving other Africans. It thus appears that the allele frequencies of Aboriginal Australians and Papuans can be explained as deriving from the same out-of-Africa migration as Eurasians and subsequent admixture with Denisovans, although these allele frequency correlation approaches do not have power to rule out very small (e.g. a few percent or less) contributions to the ancestry of Aboriginal Australians and Papuans from an earlier migration from Africa.

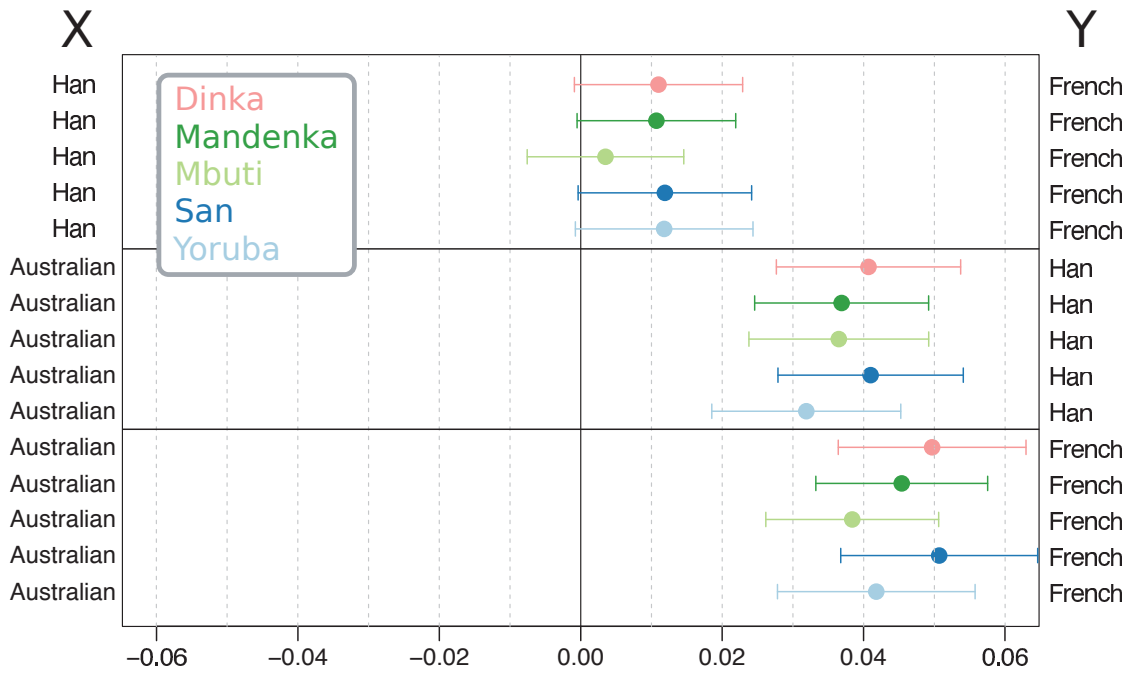

Figure 2.5: D-statistic tests of the relationships between Africans and non-Africans. *Tests of the form* D*(Chimp,African;X,Y), where African is each of the populations indicated in the grey box, and X and Y are the populations indicated on the left and the right side, respectively. A negative value indicates that the African population is more similar to X than to Y, and a positive value that it is more similar to Y than to X. Bars denote* ± *three standard deviations.*

A separate study analysing diverse human genomes reported evidence for at least 2% of Papuan ancestry deriving from an earlier out-of-Africa migration, in contrast to the above (Pagani, et al. 2016). This was primarily based on analyses of haplotype matching patterns between non-Africans, Africans and archaic genomes, as well as the observation that MSMC (Schiffels and Durbin 2014) cross-coalescence curves between Papuans and Africans indicated an earlier separation than those between Eurasians and Africans. The latter observation (as well as the same behaviour for Aboriginal Australian genomes) was also made in (Malaspinas, et al. 2016), but attributed to some combination of back-flow into Africa from Eurasian sources, technical uncertainty surrounding haplotype phasing and potentially effects of archaic admixture. Similar uncertainties exist around the haplotype matching approach of (Pagani, et al. 2016), but at present a proposed contribution of 2% from a population that left Africa earlier into Aboriginal Australian and Papuan genomes cannot be ruled out.

In summary, however, while there is disagreement about this very small amount of ancestry, all of these studies agree that the vast majority of Aboriginal Australian and Papuan ancestry derives from the same, single migration out of Africa as Eurasian ancestry. Such a small contribution, if real, is likely only detectable through sophisticated analysis of phased, whole-genome sequences, and is unlikely to have been visible to earlier genetic studies. Such studies reporting evidence for an earlier exit were likely instead confounded by the Denisovan admixture in Aboriginal Australian and Papuan genomes, which make them appear more divergent from Africans than what Eurasians do. The consensus agreement on the vast majority of non-African ancestry deriving from the same African source thus arguably represents an important milestone in human population genetics.

2.7 Relationship to Eurasians

Measuring genetic affinities to worldwide populations using the outgroup f_3-statistic reveals that the populations closest to Aboriginal Australians are Papua New Guineans and related populations of Melanesia and Polynesia, as expected (Figure 2.6). After this, there is higher affinity for East Asians and Native Americans than for Europeans. The gradient visible throughout island Southeast Asia likely reflects admixture in varying degrees between one ancestral component related to the populations of Sahul and one related to East Asians, as has been suggested by previous studies of these regions (Reich, et al. 2011; Lipson, et al. 2014).

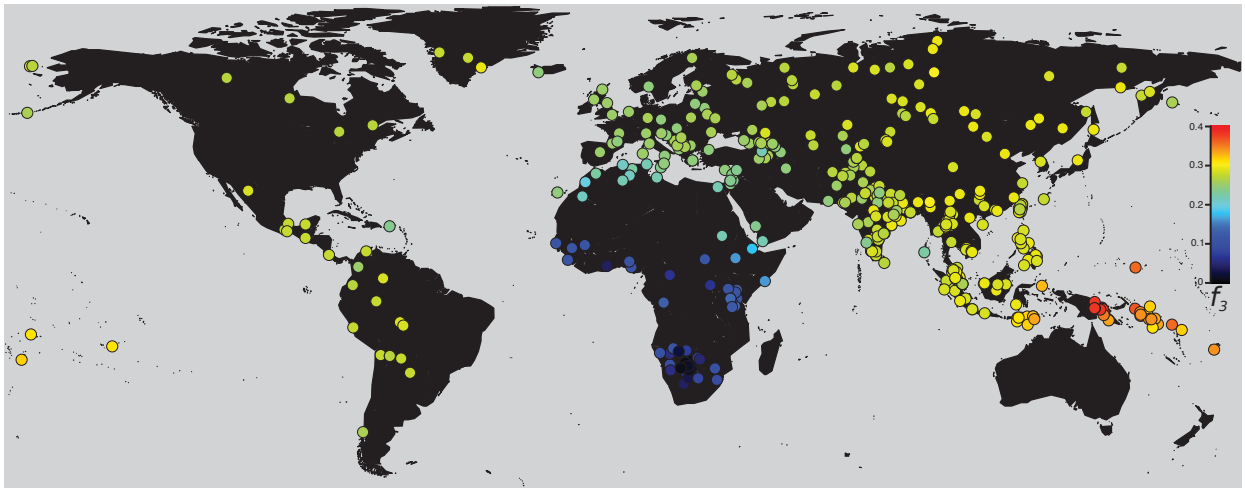

Figure 2.6: Genetic affinities of Aboriginal Australians to worldwide populations. *The outgroup f_3-statistic f_3(Aboriginal Australian,X;San) quantifies the amount of shared drift between Aboriginal Australians and worldwide populations X, relative to the southern African San population (red indicates higher affinity). Only unadmixed Aboriginal Australians were used. The data on the populations in the figure come from a range of genotyping array studies, compiled in (Malaspinas, et al. 2016). Note that as an African population is used as the outgroup, values for other African populations that might be related to that outgroup will have their values distorted.*

D-statistics confirm these general patterns more formally (Table 2.1). While I describe results for Aboriginal Australians in what follows, they are fully interchangeable with Papuans. Aboriginal Australians share more with East Asians than with Europeans, and also more with Native Americans than with Europeans. However, they share more with East Asians than with Native Americans, implying either gene flow between Aboriginal Australians and East Asians or between Native Americans and an outgroup, since the divergence between these lineages. Given that about a third of Native American ancestry is known to derive from a source related to West Eurasians rather than East Asians (Raghavan, et al. 2014), the latter scenario likely explains this statistic.

Both Europeans and the 45-ky-old Ust'-Ishim share more with East Asians than with Aboriginal Australians, however this statistic, like the statistics involving African populations above, is affected by Denisovan admixture. The strongly positive value of the statistic *D*(Chimp,Aboriginal Australian;French,Han) demonstrates that Aboriginal Australians cannot be a strict outgroup to Eurasians. This statistic will only be affected by the Denisovan ancestry in Aboriginal Australians if Han have substantially more Denisovan, or Neanderthal, ancestry than French – while Han seem

Test	*D*	Z
D(Chimp,French;Aboriginal Australian,Han Chinese)	0.0524	9.298
D(Chimp,Han;Aboriginal Australian,French)	-0.0109	-1.848
D(Chimp,Aboriginal Australian;French,Han)	0.0632	11.563
D(Chimp,Ust'-Ishim;Aboriginal Australian,Han Chinese)	0.0323	5.185
D(Chimp,Aboriginal Australian;Han,Karitiana)	-0.0165	-3.271
D(Chimp,Aboriginal Australian;French,Karitiana)	0.0493	9.686

Table 2.1:* D-*statistic tests on the relationships between Aboriginal Australian and Eurasian populations. *Karitiana is a Native American population from the Amazonian region of Brazil. Ust'-Ishim is a 45,000-year-old modern human from Siberia, approximately equally related to all present-day Eurasians (Fu, et al. 2014). Values are shown for Aboriginal Australians only, but they are very similar when using Papuans instead.*

to have on the order of 0.1% Denisovan (Sankararaman, et al. 2016) and slightly higher Neanderthal ancestry than French (Wall, et al. 2013), these small differences will not explain the very large value of the statistic. An earlier study proposed that Aboriginal Australians are in fact an outgroup to Eurasians, but that there had been gene flow between them and East Asians, after the latter separated from Europeans (Rasmussen, et al. 2011). However, this study did not take into account the effect of Denisovan admixture on these statistics, in particular how it will make Europeans share more with East Asians than with Aboriginal Australians.

A separate analysis of these data in (Malaspinas, et al. 2016), based on fitting models to the site frequency spectrum (Excoffier and Foll 2011), also favoured a model where Aboriginal Australians are the outgroup. The same conclusion, using Papuans in place of Aboriginal Australians, was reached in (Pagani, et al. 2016). Meanwhile, admixture graphs in (Mallick, et al. 2016) and (Lipson and Reich 2017) have been used to show that, when accounting for the Denisovan admixture, the allele frequencies of Aboriginal Australians and Papuans are consistent with being in a clade with East Asians, and thus with Europeans as the outgroup. Some insight into the cause for these opposing conclusions was provided in a recent study (Wall 2017), demonstrating that the model providing the best fit to the site frequency spectrum in (Malaspinas, et al. 2016) had such high rates of gene flow between Aboriginal Australians and the ancestors of Europeans and East Asians, until the split between the latter two, that in practice it essentially behaves like a model in which Europeans are the outgroup. The study also presented parsimony based allele sharing analyses that favour Europeans as the outgroup.

A recent study of predicted archaic haplotypes present in modern human genomes suggested an earlier split for the Aboriginal Australian (or Melanesian, in that study) lineage on the basis of the degree of difference between the local landscapes of Neanderthal ancestry across the genomes of different populations (Vernot, et al. 2016). However, the processes governing the loss of archaic haplotypes over time are arguably not well enough understood to take this as strong evidence. The overall balance of evidence at present, considering the results presented here and those in the literature discussed above, thus arguably points in the direction of the scenario where Aboriginal Australians and East Asians form a clade and Europeans were the first to branch off.

Application of the MSMC method to the question of the divergence time between Aboriginal Australians and East Asians leads to estimates of ∼45 kya, and a slightly older divergence to Europeans (Figure 2.7A). There is considerable technical uncertainty surrounding this analysis, in particular related to the likely poor phasing quality of Aboriginal Australian genomes. It has been shown that with inadequate statistical phasing the method tends to underestimate the age of population splits, perhaps due to reference bias arising from the haplotype panel used for the phasing (Song, et al. 2017). In any case, these estimates are similar to those obtained from the Y chromosome, and given the large uncertainty involved, compatible with the Sahul archaeological record and the notion that present-day Aboriginal Australians are the descendants of the first people to settle the continent ∼50 kya. MSMC analyses in a separate study produced estimates of ∼33 kya for the split between Papuans and mainland East Asians (Pagani, et al. 2016), which likely is an underestimate. Site frequency spectrum modelling analyses in (Malaspinas, et al. 2016) produced an estimate of ∼58 kya (95% confidence interval: 51–72 kya) for the age of the split under a model where Australo-Papuans are an outgroup to Eurasians; however, it was later noted that this is likely to be an over-estimate because the model contains high rates of gene flow to Eurasians after the split (Wall 2017).

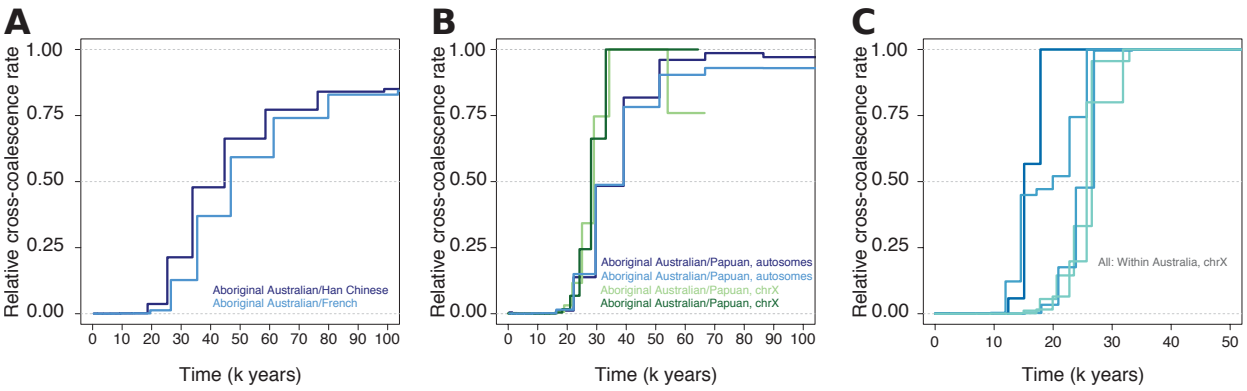

Figure 2.7: Timing of divergence between Aboriginal Australians from other populations. *A) Cross-coalescence curves between Aboriginal Australians and Eurasians suggests a split time of ∼45 kya from East Asians, and slightly older from Europeans. These were computed using MSMC2 on two genomes per population. B) Cross-coalescence curves between Aboriginal Australians and Papuans, suggests a split time of ∼30-35 kya. The autosomal curves were calculated using MSMC2 on two diploid genomes per population, while the X chromosome curves were calculated using MSMC on four chromosomes per population. In each case there are two curves corresponding to analyses performed with different sets of individuals. C) Cross-coalescence curves between different Aboriginal Australian sub-populations, calculated using MSMC on two male X chromosomes per population, using individuals that have relatively little foreign admixture (there were not enough such male individuals in different sub-populations to allow for these analyses to be run with four chromosomes). The curves displayed are those that go the deepest into the past, and these involve populations from the northeast against populations from the southwest of Australia. As some of these individuals still carry some level of foreign admixture, that might potentially push the curves towards artefactually older divergences.*

2.8 Relationships to Papua New Guineans

No individual from the Aboriginal Australian whole-genome sequencing dataset has a directly visible higher affinity to Papuans than the largely unadmixed individuals from the Western Central Desert do: i.e. there are no significantly negative values of the statistic

D(Yoruba,Papuan;X,Western Central Desert) (with the exception of one individual with a known parent from the Torres Strait islands). However, the effect of the widespread European and other foreign admixture present in most Aboriginal Australian genomes is to decrease affinity to Papuans, thereby counteracting the ability of a test like this to detect a higher affinity. I therefore performed analyses aiming to assess affinity to Papuans while accounting for the foreign admixture.

Similarly to the *D*-statistic above, an outgroup f_3-statistic of the form f_3(Mbuti; PNG highlander, X), where Mbuti is an African outgroup, gives highly variable values across Aboriginal Australian individuals (Figure 2.8A). However, there is a largely linear relationship between these values and the amount of foreign, non-Sahul (meaning neither Aboriginal Australian nor Papuan) ancestry estimated in each individual using ADMIXTURE, suggesting this is what drives the variation. Thus, after estimating the effect of foreign ancestry on the f_3-statistic by linear regression, the admixture-adjusted f_3-values are largely uniform across individuals (Figure 2.8B). This can also be seen by applying the same adjustment to *D*-statistics of the above form (not shown). Three individuals from the very north-east of Australia do show an increased affinity to Papuans even after adjusting for non-Sahul ancestry; however, at least two of these are known to have one parent from Papua New Guinea or the Torres Strait Islands, thereby representing admixture in very recent times. There are still small but significant differences overall between the nine sampled Australian populations in their adjusted f_3 values (Kruskal–Wallis test, $P = 0.0002$, after removing the three outliers). This is likely in part to be driven by imperfect adjustment, as the highest average adjusted f_3 is found in the group with the least foreign admixture, but after that the highest values are found in the two north-eastern groups. PCA analyses suggest there might be a slightly higher affinity to Papuans in the indigenous ancestral component in these groups (Figure 2.8C). This is further supported by other analyses in (Malaspinas, et al. 2016), which also suggest Papuan admixture potentially predated European colonization. A uniparental study has also reported one individual from northern Australia carrying a mitochondrial genome from haplogroup Q, otherwise only occurring in New Guinea and Melanesia (Hudjashov, et al. 2007). But with the exception of small amounts of admixture in the north-east, the big picture is that of a uniform relationship to Papuans across Aboriginal Australians, and the absence of any major gradient of Papuan affinity across Australia. The same conclusion was reached in an independent analysis of a smaller number of Aboriginal Australians individuals (Mallick, et al. 2016).

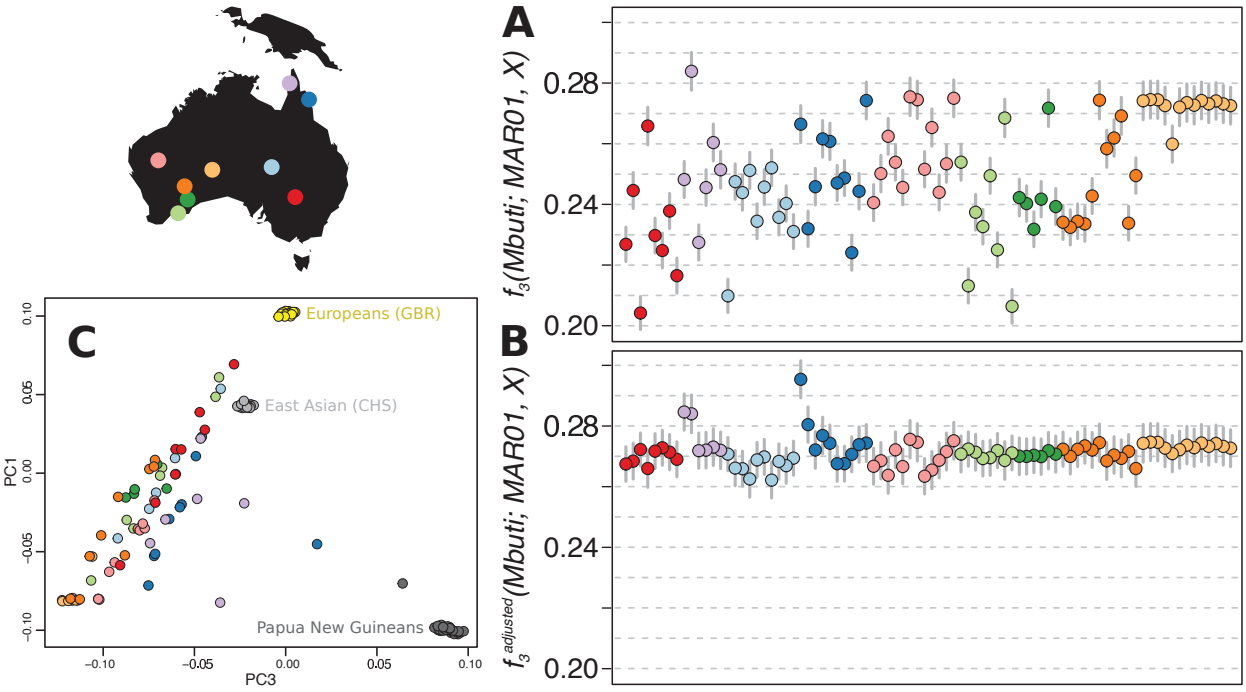

Figure 2.8: Aboriginal Australian genetic affinities to Papua New Guineans. *Individuals are coloured according to their population origin, following the map in the upper-left corner. A) Unadjusted outgroup* f_3*-statistics measuring genetic affinities of Aboriginal Australians to the arbitrarily chosen PNG highlander individual MAR01, using the African Mbuti as an outgroup. B) The same outgroup* f_3 *values after adjusting using the slope coefficient from a simple linear regression model of the form* $f_3 \sim$ *non-Sahul ancestry. Horizontal lines represent ±1 standard error. C) A principal components analysis with Aboriginal Australians, Papuans, Europeans and East Asians. PC1 separates Eurasians from populations in Sahul, with the admixture present in most Aboriginal Australians causing a dispersal along this component. PC3 separates Aboriginal Australians from Papuans. Some tendency for north-eastern groups to be shifted towards Papuans is visible, even ignoring the three outliers that are very clearly shifted (and likely represent admixture in the last generation).*

2.9 The time of separation between Aboriginal Australians and Papuans

The above f-statistics analyses indicate strong separation between Aboriginal Australians and Papuans, rather than a genetic continuum across Sahul. F_{ST}, a measure of allele frequency differentiation, is also high between the unadmixed Western Central Desert population of Australia and PNG groups, with values (0.11-0.13) as high or slightly higher than those between Europeans and East Asians (though this might be inflated because of excessive drift in this particular Australian group). However, these analyses are not directly informative about the timing of genetic divergence.

The Y chromosome analyses described above also pertain to the question of the split time between Aboriginal Australians and Papuans. Dating the most recent common ancestors between Aboriginal Australian and Papuan Y chromosomes (excluding the Australian individuals carrying haplogroup M chromosomes, likely reflecting recent admixture from the Torres Strait Islands) leads to estimates of 48.4 kya (95% CI: 42.8–54.9 ky) in the K*/M haplogroup (Figure 2.1B) and 50.1 kya (95% CI: 44.3–56.9 ky) in the C haplogroup (Figure 2.1C). Due to the limited number Y chromosomes sampled, it is, however, possible that there are other lineages which have not been sampled but which would reveal more recently shared common ancestors. I addressed this by constructing a

new phylogeny incorporating additional Y chromosomes from the novel whole-genome sequencing data from Papuans and Aboriginal Australians, as well as data from an early 20th century Aboriginal Australian hair sample (Rasmussen, et al. 2011), thereby increasing the sample size to 37 Papuan and 37 Aboriginal Australian (again excluding the M haplogroup) Y chromosomes of indigenous origin. The addition of these samples to the phylogeny did not lead to the appearance of any new branches that would constitute more recent common ancestors between Aboriginal Australians and Papuans. The results suggest that this sampling is comprehensive enough to have most likely identified the most recently separated linages, and thus that Aboriginal Australian and Papuan Y chromosomes really did separate from each other approximately 50 kya.

Application of the MSMC method (Schiffels and Durbin 2014) to two whole genome sequences per population results in a split time inference of approximately 30 ky (Figure 2.7B), in line with results obtained in (Malaspinas, et al. 2016). There is, however, uncertainty associated with these results owing to the likely poor phasing quality of Aboriginal Australian and Papuan genomes, to which MSMC is sensitive (Song, et al. 2017). The details of the statistical phasing strategy have also been shown to have large effects on MSMC split time analyses of Aboriginal Australian and Papuan genomes (Mallick, et al. 2016). One approach to get around the issue of phasing quality is to make use of X chromosome sequences from male samples, which due to their haploid state are necessarily perfectly phased. Population split times have previously been estimated indirectly from pairs of male X chromosomes by inferring cross-population effective population size using the PSMC method (Li and Durbin 2011), however this approach has limited resolution on the time-scale of the last $\sim$30 kya. Using multiple X chromosomes per population with the MSMC method allows for the direct study of more recent splits. While many cross-coalescence curves computed in this way between Aboriginal Australian and Papuan populations exhibit non-monotonic and difficult-to-interpret behaviour, the curves that allow for more straightforward interpretation suggest splits of $\sim$30 kya, similar to the autosomal curves but perhaps slightly more recent (Figure 2.7B). An independent analysis based on the site frequency spectrum in (Malaspinas, et al. 2016), estimated a split time of $\sim$37 kya. A previous study also estimated a split time in the same date range, at 36 kya, on the basis of LD correlations (Pugach, et al. 2013).

In summary, the data suggest a relatively old split time between Aboriginal Australians and Papuans, on the order of 30 kya. This is approaching the age of the split between European and East Asian populations, which is at least 40 ky old (Fu, Meyer, et al. 2013; Schiffels and Durbin 2014), though technical uncertainty still exists. In any case the split is likely substantially older than the geographical separation of Australia and New Guinea following rising sea levels only in the last 10 ky. The split times between the Y chromosomes of these populations are even older, at approximately 50 kya, and methodologically much more reliable. However, it's important to note that uniparental chromosome splits do not necessarily correspond closely to population splits – it's possible that lineages that shared more recent common ancestors have been lost due to drift (examples of which have been demonstrated through ancient DNA in other parts of the world (Posth,

et al. 2016)). The Y chromosome results still, however, reinforce the overall picture from the autosomal data of a surprisingly early population separation in Sahul.

2.10 Population structure and its time depth within Australia

The widespread European and other foreign admixture in the Aboriginal Australian whole-genome sequencing dataset unfortunately makes analysis of population structure within Australia challenging. However, some structure is still discernible, with the most notable aspect being differentiation between southwest and northeast. Analyses in (Malaspinas, et al. 2016) using the MSMC method for purposes of dating population splits within Australia found the results to be too unreliable, likely as a consequence of poor haplotype phasing quality. Applying MSMC to pseudo-diploid male X chromosomes, as described above, restricting to individuals that have relatively low levels of foreign admixture, allows for at least some basic assessment of the time depth of population structure. While the resulting curves are noisy and difficult to interpret, the oldest splits between different Aboriginal Australian populations appear to date back to 20-25 kya, and involve groups from the southwest and the northeast (Figure 2.7C). There is, however, substantial methodological uncertainty surrounding these estimates, including how the foreign admixture that is present in some of the individuals might make divergences appear older than they are. An independent analysis in (Malaspinas, et al. 2016) based on fitting models to the site frequency spectrum, and restricting only to the small number of individuals displaying very little or no foreign admixture at all, obtained an estimate of $\sim$31 kya (95% confidence interval: 10-32 kya) for the divergence between south-western and north-eastern groups, though with some gene flow after the split. Overall, these results suggest that population structure within Australia might be relatively old, but more data and further analyses is needed to obtain higher confidence results on this question.

2.11 Conclusions

In contrast to many other parts of the world where population histories are appearing fairly complex, the broad outlines of Sahul history so far seem relatively simple: a single colonization event $\sim$50 kya giving rise to all present-day inhabitants, no additional gene flow into the continent after that until recent times, and an early (perhaps at 30 kya) divergence between Aboriginal Australians and Papuans.

The long-standing debate regarding an earlier exit from Africa for the ancestors of Aboriginal Australians and Papuans appears largely, though not completely, resolved in the light of recent studies, which all at least agree that any ancestry contribution from such a migration must be very limited. The Denisovan admixture is likely responsible for making these populations look more divergent in some earlier studies. In line with this, recent studies of the Andamanese islanders, distant relatives of populations in Sahul but lacking the large amounts of Denisovan ancestry, and also previously hypothesized to carry ancestry from the same earlier migration from Africa,

have found their ancestry to be fully compatible with coming from the same source as Eurasians (Mallick, et al. 2016; Mondal, et al. 2016).

While uncertainties still exist regarding the timing of genetic divergence of populations in Sahul from those in Eurasia, the estimates are compatible with the accepted archaeological record of the earliest human activity in the continent, and with the notion that present-day Aboriginal Australians and Papuans are the descendants of the first settlers. The Y chromosome analysis arguably provides the methodologically most reliable estimates, indicating that the divergence from Eurasians can at least not be older than 47.8-61.6 ky, or 44.9-65.9 ky if factoring in additional technical uncertainty. The autosomal estimates are also compatible with a date of around ∼50 kya, but are subject to more methodological caveats. Another piece of evidence with relevance to the question of divergence time is the Denisovan admixture, which, given the absence of any archaeological evidence for the presence of archaic hominins in the continent, most likely predated the colonization of Sahul. This has been dated to 44-54 kya on the basis of the length of Denisovan ancestry blocks (Sankararaman, et al. 2016), calibrated assuming a Neanderthal admixture event in all non-Africans at 50-60 kya (Fu, et al. 2014). Similar results were reached independently in (Malaspinas, et al. 2016), indicating that the Neanderthal admixture in the ancestors of Aboriginal Australians and Papuans event occurred 11% earlier than the Denisovan admixture event.

Recently, new archaeological work from a site in northern Australia reported evidence of human occupation at 65 kya (conservatively 59.3 kya), far earlier than previously thought (Clarkson, et al. 2017). This date is only just compatible with the Y chromosome divergence estimates, but cannot confidently be said to be incompatible with autosomal MSMC estimates, given the large uncertainties in phasing, mutation rate and generation times. However, the date is arguably not compatible with the Denisovan admixture date estimates: if Denisovan admixture occurred at 65 kya, then Neanderthal admixture would need to have occurred ∼71.5 kya, which would be too early according to current understanding. Importantly, the dating of archaic admixture is not dependent on the autosomal mutation rate assumed (though does depend on generation times). It therefore seems unlikely that any such early inhabitants of Australia are the ancestors of present-day people. It is not impossible that earlier, small groups of humans made it to Australia without leaving descendants (or so few that their ancestry is not detectable in present-day genomes) – similar scenarios, though potentially also technical artefacts in dating, might explain very early findings in China (Liu, et al. 2015) and North America (Holen, et al. 2017).

The absence of gene flow to Australia (until European colonization) makes it unusual in a world-wide context, as most populations in other major parts of the world have come into genetic contact with other populations particularly following the several large expansions of the last 10 ky. The earlier reports of Indian gene flow to Australia appear to have been incorrect. While gene flow from Southeast Asia affected the coastal areas of New Guinea, there is currently no evidence for this in Australia. Given the size of the Australian landmass it seems unlikely that Southeast Asian seafarers would have missed it, and there is evidence of pre-European colonization visits from Makassan

sea cucumber collectors to northern Australia at least since the early 18th century (Macknight 1986). The most likely explanation for the arrival of the dingo is also arguably through Southeast Asian seafarers, a suggestion supported by genetic analyses of uniparental dingo chromosomes (Ardalan, et al. 2012; Oskarsson, et al. 2012). It thus seems likely that Southeast Asian people did reach Australia, though without admixing with the local people. However, as current sampling of Aboriginal Australians is geographically limited and analyses are complicated by recent colonial admixture, it cannot be ruled out that future studies might uncover Southeast Asian admixture in particular sub-populations in Australia.

2.12 Materials and methods

The Y chromosome phylogeny inference and dating are described in greater detail in (Bergström, et al. 2016). Sequencing of Aboriginal Australian and Papua New Guinean whole-genomes was carried out by the Wellcome Trust Sanger Institute sequencing facilities, and only reads mapping to the Y chromosome were then analysed. The expanded phylogeny incorporating additional samples was constructed using the same technical protocol. Briefly, variants were called from read alignments using FreeBayes v.0.9.18 (Garrison and Marth 2012) with the arguments "—ploidy 1" and "—report-monomorphic", jointly across samples, restricted to $\sim$10 million sites on the Y chromosome that are suitable for short read mapping (Poznik, et al. 2013). Sites were further filtered on the basis of unusually high or low coverage across samples, an excess of reads with mapping quality 0 (meaning another mapping location in the genome is equally good), an excess of missing genotypes or an excess of samples having reads that do not agree with the called genotype. Additionally, genotypes were set to missing for a given sample that had less than two reads overall at a given site, more than one allele supported by more than one read, or a fraction of reads supporting the called genotype of less than 0.75. A maximum likelihood phylogeny was then constructed using RAxML 8.1.15 (Stamatakis 2014). The age of a give node in the tree was estimated using the ρ statistic (Forster, et al. 1996), i.e. averaging over all possible pairwise divergences between the chromosomes in the two descendant branches (pooling data across low-coverage samples when appropriate), each estimated as the number of derived mutations separating the two sequences divided by the number of sites with an ancestral genotype called. Ancestral state was inferred by aggregating genotype calls across the 12 samples in haplogroups A and B from the 1000 Genomes Project. Divergence times were converted to units of years by applying a mutation rate of 0.76×10^{-9} mutations per site per year (95% confidence interval: 0.67×10^{-9} to 0.86×10^{-9}), inferred from the amounts of missing mutations relative to present-day individuals on the Y chromosome of the 45,000-year-old Ust'-Ishim sample (Fu, et al. 2014).

The whole-genome sequencing dataset and analyses of it is described in greater detail in (Malaspinas, et al. 2016). The sequencing of Papua New Guinean whole genomes was carried out by the Wellcome Trust Sanger Institute sequencing facilities. A brief description of the methods used for the analyses described here follows. The data were merged with data from the 1000

Genomes Project (1000 Genomes Project Consortium 2015) using the "merge" command from the bcftools software (https://samtools.github.io/bcftools/) with the "-m" argument, excluding sites that became multiallelic after merging. Genotypes for chimpanzee were extracted from the UCSC hg19-panTro4 axtNet alignments and merged in the same way. PCA analyses were performed using EIGENSTRAT 6.0.1 (Patterson, et al. 2006), using the "-w" flag to restrict the calculation of principal components to a subset of individuals. ADMIXTURE 1.23 (Alexander, et al. 2009) was used for model-based ancestry assignment. Per-individual heterozygosity was calculated directly from the number of heterozygous genotype calls and the number of homozygous genotype calls. f_3- and D-statistics were calculated using ADMIXTOOLS 3.0 (Patterson, et al. 2012). MSMC (Schiffels and Durbin 2014) (as well as MSMC2) was used to estimate the relative cross-coalescence rate between populations, using the recommended mappability mask and applying the "—skipAmbiguous" argument to exclude unphased segments and the "—fixedRecombination" argument. To run MSMC analyses on male X chromosomes, genotypes were first called per-sample using FreeBayes v0.9.18 (Garrison and Marth 2012), with the arguments "-- ploidy 1" and "--report-monomorphic". Positions were excluded if the read depth was below one third or above double the average X chromosome read depth, if the $\log_{10}$ genotype likelihood of the second-best genotype was higher than -30 or if an indel genotype was called. Haploid genotypes from different males were then combined into synthetic diploid chromosomes, which were used as input for MSMC as above. All demographic results were scaled to units of years using a mutation rate of 1.25×10^{-8} per site per generation and a mean generation interval time of 29 years (Fenner 2005). For X chromosome results, this rate was scaled down by a factor of 0.75, inferred from the relationship between the Y chromosome and the autosomal mutation rate estimates reported from the depletion of mutations relative to modern genomes of a 45,000-year-old modern human (Fu, et al. 2014). Statistical tests including linear regression were carried out in R. Data was plotted onto maps using the R maps and mapdata packages.