Genomic insights into the human population history of Australia and New Guinea

Anders Bergström

Wellcome Trust Sanger Institute

Magdalene College

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

September 2017

Genomic insights into the human population history of Australia and New Guinea

Anders Bergström, Wellcome Trust Sanger Institute, Magdalene College, University of Cambridge

Abstract

The ancient continent of Sahul, encompassing Australia, New Guinea and Tasmania, contains some of the earliest archaeological evidence for humans outside of Africa, dating back to at least 50 thousand years ago (kya). New Guinea was also one of the sites were humans developed agriculture in the last 10 thousand years. Despite the importance of this part of the world to the history of humanity outside Africa, little is known about the population history of the people living here. In this thesis I present population-genetic studies using whole-genome sequencing and genotype array datasets from more than 500 indigenous individuals from Australia and New Guinea, as well as initial work on large-scale sequencing of other, worldwide, human populations in the Human Genome Diversity Project panel.

Other than recent admixture after European colonization of Australia, and Southeast Asian admixture in the lowlands of New Guinea in the last few millennia, the populations of Sahul appear to have been genetically independent from the rest of the world since their divergence ∼50 kya. There is no evidence for South Asian gene flow to Australia, as previously suggested, and the highlands of Papua New Guinea (PNG) have remained unaffected by non-New Guinean gene flow until the present day. Despite Sahul being a single connected landmass until ∼8 kya, different groups across Australia are nearly equally related to Papuans, and vice versa, and the two appear to have separated genetically already ∼30 kya. In PNG, all highlanders strikingly appear to form a clade relative to lowlanders, and population structure seems to have been reshaped, with major population size increases, on the same timescale as the spread of agriculture. However, present-day genetic differentiation between groups is much stronger in PNG than in other parts of the world that have also transitioned to agriculture, demonstrating that such a lifestyle change does not necessarily lead to genetic homogenization.

The results presented here provide detailed insights into the population history of Sahul, and suggests that its history can serve as an independent source of evidence for understanding human evolutionary trajectories, including the relationships between genetics, lifestyle, languages and culture.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the introduction of each chapter and/or specified in the text. None of the contents in this dissertation have been submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit for the Biology Degree Committee.

Anders Bergström

September 2017

Acknowledgements

First of all, I would like to thank Chris Tyler-Smith and Yali Xue for their supervision and guidance, and for making it possible for me to work on such exciting projects. Thanks also to the members of the Tyler-Smith group for all the feedback on my work and the stimulating and far-reaching scientific discussions. Thanks to Manjinder Sandhu, Richard Durbin, Shane McCarthy, Deepti Gurdasani and others at the Sanger Institute for fruitful collaborations, and the institute itself, its graduate program and the Wellcome Trust for giving me this opportunity. I would also like to thank the numerous people who supported me and gave me opportunities to learn and do stimulating science before embarking on this PhD, in particular Jonas Warringer and colleagues in Gothenburg and Gianni Liti and his team in Nice.

Secondly, I am grateful to the numerous collaborators who in one way or another have enabled the work described in this thesis, including Stephen Oppenheimer, Alexander Mentzer and colleagues in Oxford and at the Papua New Guinea Institute of Medical Research for our highly fruitful collaboration on Papua New Guinea, John Mitchell and colleagues in Australia, Eske Willerslev, Anna-Sapfo Malaspinas, David Reich, Jaume Bertranpetit, Aylwyn Scally and many others I have interacted with and learnt from during the course of this work.

Lastly, thanks to my friends and family, near and far, for their support and patience over these years. The biggest thanks goes to Giulia, for her endless support and encouragement.

Contents

Chapter 1: Introduction

1.1 Outline of this thesis

This thesis begins with the current introductory chapter, which outlines some basic principles of population genetics and experimental methodology, as well as the main features of human population history as currently understood. The second and third chapters describe my work on Aboriginal Australian and Papua New Guinean population history respectively, while the forth describes my initial work on the sequencing of a large set of diverse human genomes and associated technical aspects. Human population genetics is currently a rapidly moving field, and much progress has been made while the work described in this thesis was being conducted. On those questions that constitute the main focus of the thesis, I therefore try to present in the introduction the state of knowledge as it was while this work begun (mid-2014), and cover later findings in the results and discussion sections alongside my own work. On questions that fall outside that main focus, I however try to present in the introduction the state of knowledge up until the writing of this thesis.

1.2 Basic principles of population genetics

Genomes are composed of DNA molecules and contain the information needed for the life of an organism. They are shaped over time by evolutionary processes, some of which are driven by the more successful spread of some genome variants than others, i.e. natural selection (Darwin 1859), and some of which are driven by biological and chemical properties inherent to DNA and its transmission in populations (Lynch 2007). Evolutionary changes that are not influenced by natural selection are referred to as neutral. The genome of an organism constitutes a record of the processes that shaped it, both neutral and adaptive, and by studying genomes we can therefore learn about evolutionary history.

The ultimate source of all evolutionary change is mutation. Different mutational processes, most of which occur during the replication of DNA before cell division, lead to different types of changes in the genome sequence. The most abundant type of mutation is the substitution of a single nucleotide for another, which in a population become single-nucleotide polymorphisms (SNPs). Insertions and deletions of short segments of sequence, collectively referred to as indels, occur at a rate of about one order of magnitude lower than single-nucleotide substitutions. In the specific sequence context of short tandem repeats (STRs), i.e. repeated units of sequence a few base pairs long, slippage of the DNA polymerase enzyme during replication leads to mutation rates that can be several orders of magnitude higher than other indel mutation rates. In addition to these small-scale changes, genomes undergo larger-scale structural changes, where large parts of sequence are duplicated, deleted, inverted or translocated to a different part of the genome. In cases where there is a net change in the number of units of a given piece of the sequence in the genome, these are often detected as and referred to as copy number variants (CNVs) in a population.

In sexually reproducing organisms, the process of recombination results in a shuffling of genetic material between homologous chromosomes. Without recombination, sequences are passed on over generations as unbroken units on which all mutations that have occurred in the ancestors of that particular sequence accumulate. The mitochondrial genome and most of the mammalian Y chromosome are non-recombining. With recombination, pieces of sequence with different histories are joined together such that mutations that occurred in unrelated ancestors are brought together into the same genome. Recombination occurs during meiosis such that the chromosome passed on to the offspring is a recombined version of the two chromosomes carried by the parent, and the probability by which recombination events occur in any given part of the genome varies. The closer two variants are to each other along the chromosome, the less often a recombination event will occur between them. The resulting co-segregation of such variants in a population at a frequency greater than the random expectation is referred to as linkage disequilibrium (LD).

Basic principles underlie the fate of a mutation in a population. In the absence of natural selection, the frequency of a mutation in the population is entirely governed by chance and fluctuates from one generation to another due to the random sampling of alleles during meiosis. Under such conditions, the probability that a mutation will eventually reach fixation, i.e. to spread to all individuals in the population, is simply equal to its frequency in the population (Hartl and Clark 2007). This random change in allele frequencies over time is referred to as genetic drift.

If a mutation has an influence on the number of future descendants its carrier will have, the so-called fitness of its carrier, its frequency will also be affected by this selective process, in addition to random genetic drift. The larger the effect on fitness, the more rapidly it will either increase or decrease in frequency. Selection against mutations that decrease fitness is often referred to as purifying selection. A key parameter determining the strength of natural selection relative to drift is the effective population size (N_e), which can be thought of as the size of the subset of the population that is actually contributing genetic material to the next generation (Hartl and Clark 2007). When effective population size is small, random changes in allele frequency has a greater impact such that genetic drift occurs more rapidly, potentially overwhelming small fitness differences between alleles.

Across the tree of life, there is great diversity in the size and organization of genomes as a consequence of the varying importance of different evolutionary forces in the history of different types of organisms. Bacteria often have extremely large effective population sizes in which natural selection is very efficient relative to genetic drift, such that bacterial genomes are small (typically 3-10 Mb), have high protein coding content and little non-functional sequence (Lynch 2007). The smaller population sizes of large mammals result in less efficient selection and a greater importance of genetic drift and neutral processes. The large size of the human genome, at 3 Gb typical of large mammals, is mainly due to the largely neutral accumulation of a variety of so-called transposable elements, which are virus-like, self-replicating sequences present in hundreds of thousands to millions of copies throughout the genome. While much is still unknown about the approximately 98%

of the human genome that does not code for proteins, it appears that the majority of this sequence has little if any functional importance and so is evolving neutrally (Ponting 2017). This is a good thing from the point of view of the study of evolutionary history, as many analyses of this history are simplified if the changes in frequency of variants are primarily governed by drift, without the added complication of selection.

In humans, the rate of single-nucleotide mutation is approximately 1.25×10^{-8} per base pair per generation, though there is still considerable uncertainty about this number (Scally and Durbin 2012; Moorjani, Gao, et al. 2016). Across the genome, this means every human is born with somewhere between 10-100 new single-nucleotide mutations. The average inter-generational time has likely varied across time and between different human cultures, but has been estimated to somewhere between 25 and 30 years (Fenner 2005; Moorjani, Sankararaman, et al. 2016). The genetic effective size of human and ancestral hominin populations has likely also varied considerably throughout time, but is thought to typically have been on the rough order of 10,000.

1.3 Technology for studying variation in genome sequences

A range of technologies for studying variation in the genome sequences of individuals have been used throughout the history of genetics research. The earliest geneticists, going back to Gregor Mendel and his studies in pea plants in the mid-19th century and taking off more substantially through studies in fruit flies and other model organisms in the early 20th century, exploited the effects that certain genetic variants had on readily visible traits to track inheritance patterns. Experimental genetics thus started long before the discovery that the DNA molecule was the carrier of the genetic information, and the determination of its chemical structure. Later studies made use of differences in biochemical or immunological properties of proteins. These were thus still only assaying the genetic material indirectly, through the effects that variants might have on e.g. the shape or chemical charge of a protein, but extended the reach of genetics beyond variants that have large effects on the organism, even to variants that might be evolutionarily neutral.

The first experiments to assay variation in the DNA molecule directly made use of restriction enzymes discovered in bacteria. These enzymes introduce double stranded breaks in the DNA at specific target motifs typically around 4-8 base pairs long, which will occur every few thousand base pairs in a genome just by chance. The mixture of small DNA fragments resulting from these cuts can be separated by gel electrophoresis and hybridised to a sequence of interest to produce a visible profile of fragment lengths. Variation in the genome sequence between individuals, e.g. sequence differences within a potential enzyme target site, will sometimes lead to differences in the cuts being made and thereby differences in the fragment profiles. These restriction fragment length polymorphism (RFLP) studies enabled some early insights into the structure of genetic variation, as well as disease mapping and other practical applications in humans. Later studies also started making use of variation in the length of highly variable microsatellites, or short tandem repeats, which are amplified through a polymerase chain reaction (PCR) and then similarly visualized on a gel.

Direct sequencing of DNA molecules got started in the 1970's. The method that became predominant for the next several decades was the so-called chain termination method developed by Frederick Sanger and colleagues. This method makes use of dideoxynucleotides, modified versions of the standard deoxynucleotides that terminate elongation of the DNA molecule when incorporated by a DNA polymerase. In its mature form, by labelling the four different dideoxynucleotides with fluorescent dyes and separating the mixture of molecules terminated at different points by gel or capillary electrophoresis, the sequence of the molecule can be read off. The reads, continuous sections of sequence, are typically 400-800 base pairs long. This was the technology that was used to sequence the first whole genomes of bacteria and eukaryotes, including the human genome for which assemblies were published in 2001 (Lander, et al. 2001; Venter, et al. 2001). These early genome sequencing projects would typically proceed by sub-cloning the genome into sections on the size scale of $\sim$100 kb, sequencing and computationally assembling the reads into a consensus sequence for each one of these independently. This hierarchical sequencing approach reduces the complexity of the consensus sequence determination problem and the risk of sequence assembly errors caused by different parts of the genome having similar sequence, resulting in high quality assemblies. The sub-cloning process is labor-intensive and expensive, however, and most whole-genome sequencing projects are therefore no longer done in this way. The Sanger dideoxynucleotide method in itself, however, is still used, primarily for small-scale sequencing of targeted regions.

Another technological development that has been and remains very important to the study of genetic variation is the high-density microarray. These arrays can hold large numbers of short ($\sim$50-100 base pairs (bp)) oligonucleotide probes that target specific parts of the genome where some variant is known to segregate in the population. The relative intensity of DNA hybridization to two paired probes that differ only by the sequence corresponding to the two different alleles of a variant can then be used to determine the genotype at the variant site in a DNA sample. In this manner, the genotypes of up to millions of variants can be determined in a single, relatively cheap, experiment. These genotyping arrays enabled high-resolution studies of human genetic variation and population structure. However, the fact that the variants being assayed have to be known beforehand limits the scope of the technology, and the fact that for humans most variants have traditionally been ascertained in populations of European ancestry means that a bias might be introduced when studying more diverse populations. Nonetheless they remain an important tool in the study of genetic variation.

A major breakthrough came with the development of so-called next-generation sequencing technologies in the first decade of the 21st century (Metzker 2010). These encompassed several different technologies that could produce sequence reads at a much higher throughput and lower cost than the Sanger method. The Illumina technology emerged as the most cost-effective and is the most widely used today. Illumina machines perform so-called sequencing-by-synthesis, where DNA molecules are extended one fluorescently labelled, reversibly terminating, nucleotide at the time by a DNA polymerase and a photograph is taken to capture the identity of the incorporated nucleotide. The key to the high throughput of the technology is that this is done across billions of molecules in parallel, spread out over a flow cell. The reads produced currently range in size between 50 and 250 base pairs depending on the machine. The short read length, and the fact that most sequencing is done not in a hierarchical but rather through a so called "whole-genome shotgun" approach, means that determining the consensus sequence of the sequenced genome from the generated reads comes with greater computational challenges than the data generated by the early genome sequencing projects.

There are limits to what a given technology can tell us about a genome. The read length of a sequencing technology imposes an upper limit on the length of repetitive regions for which the sequence can unambiguously determined, such that it's not possible to say from which of two different but identical regions of the genome of length 100 base pairs an Illumina read of that length derives. The situation is somewhat improved by the application of so-called paired-end sequencing, which allows the two ends of a typically 300-600 base pair molecule to be sequenced. If the location of the first read in such a pair is unambiguously known, this can then help the determination of the second read's location. Large genomes that are rich in repeats, such as the human genome, will, however, still contain many complex and repetitive regions that are outside the reach of short read technologies. Another limitation when sequencing diploid organisms such as humans involves the determination of haplotype phase. At pairs of adjacent heterozygous sites in the genome, one of the alleles at each site will be physically located on the same chromosome, with the two other alleles located on the homologous chromosome. If a read or read pair spans two such adjacent heterozygote sites, this haplotype phase can be determined, but the physical scale at which this is possible is thus limited by the read length and read pair insert size. The average distance between heterozygote sites in humans is approximately 1000 base pairs, such that 100 base pair reads from the ends of 500 base pair molecules will provide only limited phase information. This limitation is shared by genotyping arrays, which only provide unphased genotypes. Methods based on sub-cloning single, long molecules into e.g. fosmids allow for longer-scale experimental phasing, but these methods are not widely used due to their labour intensity and high cost.

An independent technological development that has come to play an increasingly important role in the study of population history, particularly in humans in the 2010's, is the ability to extract and sequence DNA from ancient remains (Haber, et al. 2016; Slatkin and Racimo 2016). Sequence reads have successfully been obtained from human remains that are thousands and even tens or hundreds of thousands of years old. Data obtained in this way comes with a number of additional technical challenges. Firstly, ancient DNA tends to be highly fragmented, such that most molecules being sequenced are typically only 20-50 base pairs long. Secondly, the molecules suffer damage over time, and this damage is highly non-random – in particular, there is a very high frequency of C to T (or G to A on the complementary strand) substitutions caused by deamination of cytosine to uracil which is then read as thymine, particularly towards the ends of molecules (Orlando, et al. 2015). Lastly, ancient DNA samples are often contaminated with microbial DNA from the environment and/or human DNA from researchers or others that have handled the remains. The former is a mainly a financial problem, as large amounts of microbial DNA in the sample means that you need to sequence more in total to obtain a given amount of endogenous DNA, but the contaminating reads will be very different from any human genomes and so will not interfere with downstream analyses in any major way. The latter is potentially more problematic, as contaminant human DNA will be difficult to distinguish from endogenous human DNA, potentially leading to biased or artefactual interpretations. One way to get around this problem is to take advantage of the damage patterns of the reads, as modern contaminant DNA will lack the patterns characteristic of ancient DNA damage, and, if substantial contamination is detected, to restrict analyses to only the subset of reads that display such patterns (Fu, et al. 2015).

1.4 Methods for processing short read sequencing data

There are two principal ways in which to process short read sequencing data. The first is de-novo assembly, which is the assembly of the reads into a consensus sequence without the use of a pre-existing reference genome or other information external to the sequenced genome itself. This typically requires very high (>50x) Illumina read coverage for good results, and the computational algorithms in use to perform the assembly require very large amounts of memory. While the approach works very well for genomes less than 100 Mb in size (e.g. bacterial genomes), de-novo assemblies of human genomes are still highly fragmented and with the sequence of many regions incompletely determined. De-novo assembly of human genomes might be fruitful for particular use-cases, but it is not a commonly used technique in current analysis of sequencing data.

The second and most commonly applied approach to the processing of short read sequencing data is through the use of a reference genome sequence, to which the reads are mapped one-by-one. The human reference assembly is of high quality, and as genetic diversity is relatively low in humans, no sequenced genome will be very divergent from the reference genome. There are highly efficient algorithms for mapping short reads to a reference assembly and then aligning them against the reference sequence in the mapped location, allowing for some number of differences such that variants can be discovered. The most widely used software for this task is BWA, based on a string compression algorithm known as the Burrows-Wheeler transform (Li and Durbin 2009). Given the alignments of reads against the reference, the genotypes of the sequenced sample can be inferred from the reads covering a given position in the reference. Widely used software packages for performing genotype calling include samtools (Li 2011), GATK (McKenna, et al. 2010) and FreeBayes (Garrison and Marth 2012). A drawback of this approach is that it might introduce a reference bias – firstly, only parts of the genome present and correctly assembled in the reference sequence can be analysed, and secondly, if some individuals are more genetically divergent from the reference sequence than others, their reads might have a lower probability of being mapped successfully because they contain too many differences.

A key parameter determining the accuracy by which genotypes can be called against a reference genome using short read data is the sequencing coverage, the average number of reads covering a given site in the reference genome. With lower number of reads covering a site, genotype calling becomes increasingly susceptible to sequencing errors in individual reads (occurring at a rate of approximately 1% in Illumina reads) as well as sampling noise at heterozygous sites. A common target coverage in population history studies is 30x, with which very high accuracy genotypes can be obtained. Some study designs, particularly in medical genetics, favour lower coverage to instead afford sequencing of a larger number of individuals, and combine information across individuals to improve accuracy (Li, et al. 2011).

1.5 Methods for analysing population histories using genetic data

There are a variety of computational methods for learning about population histories from genetic data. They differ in what features of the genetic data they make use of, the amount of data they need for reliable results and the type of information they provide.

A large set of methods make use of allele frequencies at variant sites, under the simple assumption that populations or individuals that share ancestry will have similar allele frequencies. Principal components analysis (PCA) is one such commonly used method, allowing unsupervised clustering of individuals based on their ancestry (Patterson, et al. 2006; McVean 2009). The model-based clustering methods implemented by the STRUCTURE (Hubisz, et al. 2009) and ADMIXTURE (Alexander, et al. 2009) software are also commonly used, requiring the user to specify a number of ancestral components for the software to assign the ancestry of individuals into. The fixation index F_{ST} is a measure of the allele frequency differences between populations and thus is informative about the degree of genetic differentiation. A more recent development is a family of so-called f-statistics, comprising the f_2, f_3 and f_4 statistics and a variant of the latter known as the D-statistic (Patterson, et al. 2012). These can be used to conduct simple tests on allele frequency correlations, e.g. to test if allele frequencies are consistent with a simple tree topology or if there is admixture in the history of populations, and can be applied to single genomes or population data. As these methods make use only of allele frequencies, they can be applied to basically any type of genetic data without any particular requirements on the number or density of markers (though power will increase with the number of markers), except that markers are not in strong linkage disequilibrium e.g. in the case of PCA and ADMIXTURE.

Other methods additionally make use of genetic linkage information: the fact that physically nearby variants co-segregate on haplotypes in a population. The ChromoPainter and fineSTRUC-

TURE methods employ a framework where haplotypes are compared between individuals to assess similarity in a more high-resolution fashion (Lawson, et al. 2012). A set of methods aim to infer the local ancestry of haplotypes in admixed populations, by using reference panels related to the sources of admixture. These include the RFmix (Maples, et al. 2013), HAPMIX (Price, et al. 2009) and PCAdmix (Brisbin, et al. 2012) software packages. A requirement for these haplotype-based methods is that phase is inferred for the data, something which comes with some rate of error depending on the method for doing so. The most commonly used approach to haplotype phasing is to make use of a large reference panel of haplotypes, to which input genotypes are compared and the most likely phase inferred (Howie, et al. 2009; Delaneau, et al. 2011). In order to gain from the linkage information, haplotype-based inference methods also need a decent density of variants, such as that offered by whole-genome sequencing or high-density (e.g. $\sim$500,000 variants or more across the genome) genotyping arrays.

A recent development is the inference of population history using the distribution of coalescence times along a single genome, or between a small number of genomes, though the PSMC (Li and Durbin 2011) and later the MSMC (Schiffels and Durbin 2014) methods. As the rate of coalescence between haplotypes depends on the effective population size, this methodology can infer the history of effective population size of a population over time. Applied to multiple genomes from different populations, the relative rate of within- versus between-population coalescence can be used to infer the time scale of genetic divergence between two populations. These methods require genotype information not just at variant sites, but also at non-variant sites, and so can only be applied to whole-genome sequencing data. When applied to more than one genome, haplotype phase information is also required.

There is a disparate set of inference methods that are based on fitting explicit population history models to genetic data in one way or another. Such models might include a population topology, admixture events and rates, and effective population sizes. The parameter values are fitted typically by calculating the likelihood of the observed data, systematically searching for values that maximize this. Various features of the data can be used for such fitting, and a commonly used feature is the site-frequency spectrum (SFS). A related methodology which can be applied when it is not possible to calculate the likelihood of the data under a model is Approximate Bayesian Computation (ABC), in which summary statistics calculated from simulated data are compared to the observed data, in the search for simulation parameters that minimize the difference between these. These model-fitting approaches have the potential to infer highly complex and detailed population history models, but are often computationally very intensive, come with various technical issues related to efficient parameter search and overfitting, and due to their complexity their results can often seem opaque and difficult to evaluate.

1.6 A brief summary of current knowledge on human evolutionary history

1.6.1 Our closest hominin relatives

Anatomically modern humans evolved in Africa. We separated from our closest living relatives, the chimpanzees and bonobos, approximately 5-7 million years ago (Jobling, et al. 2004). Our closest known extinct hominin species were the Neanderthals and the Denisovans, both of which disappeared 40-50 thousand years ago (kya). Neanderthals inhabited large parts of Europe, the Near East and western Asia and much is known about their morphology and lifestyle from fossil remains. Additionally, genome sequencing data from Neanderthals, primarily a high-coverage sequence from a bone found in a cave in the Altai Mountains in southern Siberia, have provided insights into their genetic relationships to modern humans (Green, et al. 2010; Prufer, et al. 2014). The Denisovans were a completely unknown hominin group, discovered directly though the DNA sequencing of small bone fragments from the same cave in the Altai mountains, Denisova cave (Reich, et al. 2010). Teeth are the only diagnostic bone parts currently known from Denisovans, although it is possible that some bones previously thought to be from Neanderthals or other hominins might actually be from Denisovans. The genome sequence data obtained from these archaic hominins has revealed that the common ancestor of Neanderthals and Denisovans diverged genetically from modern humans approximately 600 kya (Prufer, et al. 2014), and thus likely left Africa at around this time. Neanderthals and Denisovans then diverged from each other approximately 450 kya, perhaps inhabiting primarily western and eastern Eurasia respectively. Analyses also indicated that both Neanderthals and Denisovan populations had very small effective population sizes throughout their histories after the split from modern humans, resulting in very low levels of genetic diversity.

1.6.2 Africa

There is no strong consensus on the time depth of the population structure of present-day modern human populations, but genetic analyses indicate that the deepest splits are between 100 and 200 thousand years (ky) old (Veeramah, et al. 2012; Kim, et al. 2014; Mallick, et al. 2016). These deep splits invariably involve the Khoe-San populations of southern Africa, which also appear to have had the largest effective population size of any human population throughout most of history (Kim, et al. 2014), though neither of these observations necessarily mean that southern Africa is the geographic origin of modern humans. Second to the Khoe-San, the most divergent populations appear to be certain rainforest hunter-gatherer groups of central Africa, with estimated split times to other populations on the order of at least 100 ky (Hsieh, Veeramah, et al. 2016; Mallick, et al. 2016). Population history within Africa appears complex, likely shaped by varying degrees of differentiation and admixture between different lineages over many tens of thousands of years (Gurdasani, et al. 2015). There have been suggestions that certain African populations might have experienced admixture from unknown archaic groups (Hammer, et al. 2011; Hsieh, Woerner, et al.

2016), but in the absence of actual genomic sequences from those archaic groups, these analyses are necessarily indirect and their conclusions are far from widely accepted. In the last few thousand years, African population structure appears to have undergone a major reshaping following the so-called Bantu expansion (Li, et al. 2014; Patin, et al. 2017). This was a major population expansion of agriculturalists from western Africa speaking languages in the Bantu family, quickly spreading to eastern and then southern Africa, likely leading to considerable genetic homogenization across these regions and thereby likely obscuring much of the older population structure. The last few thousand years has seen back-migrations of agriculturalist or pastoralist groups from the Near East into eastern Africa and later southern Africa, thereby bringing Eurasian ancestry to many present-day African populations in these regions (Pagani, et al. 2012; Pickrell, et al. 2014; Gallego Llorente, et al. 2015). Northern Africa has also seen major population turnover in recent times, such that a majority of the ancestry of present-day populations here can be traced to migrations from the Near East in the last few thousand years (Henn, et al. 2012).

1.6.3 Out-of-Africa

At a point which most studies estimate to be between 50 and 100 kya, with many pointing towards 60-80 kya, some humans diverged from their African relatives and migrated out of Africa (Macaulay, et al. 2005; Gravel, et al. 2011; Fu, Mittnik, et al. 2013; Schiffels and Durbin 2014; Pagani, et al. 2015). It should be noted that the date of genetic divergence from African populations does not necessarily coincide with the migration out of the geographical continent of Africa, as the migration could have been preceded by some period of genetic separation while still inside Africa. Once outside of Africa, modern humans seem to have dispersed widely and rapidly, with archaeological evidence from $\sim$50 kya appearing across Eurasia (Mellars 2006). There are some remains and archaeological evidence of anatomically modern humans outside of Africa that are considerably older than this, but their significance and in some cases their dating is disputed. It is widely acknowledged that some of these remains could very well represent earlier out-of-Africa migrations that had no discernible genetic impact on present-day non-African populations, e.g. because they left no or very few living descendants. This includes $\sim$100 ky old remains from the Levant (McDermott, et al. 1993), and potentially at least 80 ky old teeth from China (Liu, et al. 2015), though the latter date is controversial. It is important to be clear about what exactly is meant by ‘out-of-Africa’ in a given context, and in the context of the analysis of human genomes it typically refers to the genetic separation between Africans and the ancestors of present-day non-Africans, as such analysis necessarily cannot say anything about migrations that left no living descendants.

A key question is whether there was just a single group of humans (or in practice, perhaps a small number of groups with very similar ancestry) that migrated out of Africa and gave rise to all present-day non-Africans, or if there were multiple independent migrations. In the latter case, different present-day non-African populations could be derived from different subsets, or combinations of subsets, of African genetic diversity, and so have different genetic relationships to present-day Africans. The big picture consensus that has emerged is that all non-Africans likely derive most of their ancestry from the same single non-African ancestral group. This is anyhow the case for the uniparentally inherited mitochondrial genome and Y chromosome, the study of which provided some of the earliest evidence for the recent African origin of all humans (Cann, et al. 1987). However, these represent just two genetic lineages out of thousands in the genome, which additionally experience a higher rate of genetic drift, and so it is highly possible that any uniparental lineages deriving from earlier migrations could have been lost by chance during the last 60 or so ky (Nordborg 1998). The possibility of autosomal contributions from other out-of-Africa migrations has not been confidently excluded, and some non-genetic (Lahr and Foley 1994) and genetic (Reyes-Centeno, et al. 2014; Tassi, et al. 2015) studies have reported support for this scenario. In particular, it has long been hypothesized that certain populations in Southeast Asia and Sahul, namely Aboriginal Australians, Papua New Guineans, the indigenous populations of the Andaman Islands in the Indian Ocean, perhaps populations from southern India and Sri Lanka, and certain so-called Negrito groups in the Philippines and Malaysia, might harbour ancestry from such an earlier migration.

The migration of humans out of Africa was associated with a major bottleneck resulting in a decrease in genetic diversity in all non-Africans compared to Africans. This is perhaps the most striking feature of human population structure, and often has consequences for genetic analyses.

1.6.4 Archaic admixture

While little is known about the cause of extinction of our hominin relatives the Neanderthals and Denisovans, the proximity in time between their disappearance and the arrival of modern humans in their non-African habits raises competition with modern humans as at least a likely contributing factor. It's worth noting that a third hominin group, Homo floresiensis, likely much more distantly related to modern humans (Argue, et al. 2017), also seems to have disappeared from its habitat in island Southeast Asia within the same general timeframe, approximately 60 kya (Sutikna, et al. 2016).

A major finding from the analyses of the Neanderthal and Denisovan genome sequences, however, was that the genomes of these groups did not completely disappear. Firstly, it is estimated that all non-Africans carry approximately 2% Neanderthal ancestry (Prufer, et al. 2014), owing to admixture that occurred approximately 50-60 kya (Sankararaman, et al. 2012; Fu, et al. 2014), shortly after the migration of modern humans out of Africa. The largely uniform distribution of this Neanderthal ancestry across all non-African populations (Sankararaman, et al. 2014; Vernot and Akey 2014) suggests that the admixture occurred in a population that was ancestral to all of these, which is evidence in favor of a single out-of-Africa event. East Asians have, however, been found to harbour on the order of 10% more Neanderthal ancestry than Europeans (Wall, et al. 2013), reflecting either additional admixture events (Kim and Lohmueller 2015; Vernot and Akey 2015; Vernot, et al. 2016) or dilution in Europeans due to admixture with a lineage having less Neanderthal ancestry (Lazaridis, et al. 2014; Lazaridis, et al. 2016). Secondly, it is estimated that Aboriginal Australians, Papua New Guineans and some related populations in Melanesia and island Southeast Asia carry 3-5% Denisovan admixture (Reich, et al. 2011; Meyer, et al. 2012). East Eurasians and Native Americans seem to carry very low but non-zero (∼0.1%) levels of this ancestry (Skoglund and Jakobsson 2011; Qin and Stoneking 2015; Sankararaman, et al. 2016), while west Eurasians seem to have none.

There have been a number of studies of the functional and population-genetic consequences of archaic admixture. Analyses of ancient genomes from Europe have shown a gradual decrease in Neanderthal ancestry over time, consistent with purifying selection against this ancestry (Fu, et al. 2016). There is also less Neanderthal ancestry closer to functionally important elements in the genome (Sankararaman, et al. 2014; Vernot and Akey 2014). This could in principle be due to epistastic incompatibilities between Neanderthal variants and modern human genome backgrounds. However, theoretical studies have suggested that Neanderthal variants would have tended to be more deleterious in themselves, owing to the long period of less stringent purifying selection experienced by the very small Neanderthal populations, such that epistasis explanations might not be necessary (Harris and Nielsen 2016; Juric, et al. 2016). Observations that might still suggest some degree of genetic outbreeding depression are that Neanderthal ancestry is lower on the X chromosome, and in genes expressed in testis, which are known indicators of hybrid incompatibility (Sankararaman, et al. 2014). Patterns of Denisovan ancestry in modern genomes are largely similar to those of Neanderthal ancestry, implying they have been shaped by similar forces (Sankararaman, et al. 2016; Vernot, et al. 2016). While archaic ancestry seems to have been overall slightly deleterious, there are a number of specific variants that appear to have been beneficial and to have been positively selected in modern human populations (Gittelman, et al. 2016).

1.6.5 Europe

Europe was settled by modern humans approximately 45 kya (Mellars 2006; Higham, et al. 2011), and is currently the part of the world with by far the best-understood genetic history, owing to the large number of studies of both modern and ancient DNA from here. Some population changes within Europe during the Pleistocene have been described (Fu, et al. 2016), though much is still unknown about this period. Major changes to the genetic landscape of Europe occurred during the Holocene. Following the Neolithic transition from a hunter-gatherer to an agriculturalist lifestyle in the Near East, these early farmers expanded into Europe starting around 8 kya, partly replacing and partly admixing with the local hunter-gatherers. A second major transformation came during the Bronze Age starting around 5 kya, when pastoralists from the eastern European steppes migrated into Europe and admixed with local populations to replace as much as half of the ancestry in many regions (Allentoft, et al. 2015; Haak, et al. 2015; Lazaridis, et al. 2016). The result of these processes is that all present-day European populations are a mixture of these three divergent sources of ancestry, with different populations deriving different amounts of their ancestry from each of them. For example, Sardinians have mostly early farmer ancestry, while north-eastern European populations have high steppe components. The well-studied genetic history of Europe demonstrates that the expectation that patterns of genetic variation in present-day populations will largely reflect the initial peopling of a continent can be naïve, as the initial patterns are likely to be overwritten by later events. Furthermore, as major features of European genetic history were basically misunderstood or unknown until the advent of ancient DNA, it also serves as a cautionary example against overconfident conclusions drawn from modern DNA only.

1.6.6 East Asia

The genetic divergence between western and eastern Eurasians seems to go back to approximately 40 kya (Fu, Meyer, et al. 2013; Schiffels and Durbin 2014). Very little is currently known about the genetic history of East Asia after this point. Genetic differentiation between major East Asian groups, e.g. Chinese, Japanese and Vietnamese is relatively low (1000 Genomes Project Consortium 2015), which could reflect relatively recent shared ancestry between these groups, potentially compatible with a recent expansion and replacement following for example an agricultural transition. More studies, particularly of ancient DNA, will be needed to elucidate the history of this part of the world.

1.6.7 South Asia

The genetic history of South Asia is characterized by admixture between two very divergent sources of ancestry (Reich, et al. 2009; Moorjani, et al. 2013). The first is a branch of eastern Eurasian ancestry, and while it no longer seems to exist in an un-admixed form, it seems to be distantly related to the indigenous people of the Andaman Islands, who have been geographically isolated for perhaps 20 ky and so avoided later admixture (Mondal, et al. 2016). The second source is a branch of western Eurasian ancestry, not too distant from present-day European ancestry. While it still remains to be confirmed, it has been suggested that this ancestry made it into South Asia during the Bronze Age, perhaps through a population related to the eastern European steppe pastoralists who also migrated into Europe (Lazaridis, et al. 2016). The presence of Indo-European languages in both South Asia and Europe also suggests connections in relatively recent times. Present-day South Asian populations contain varying degrees of ancestry from these two sources, with a gradient of increasing western and decreasing eastern Eurasian ancestry going from south to north, and in some cases additionally some East Asian ancestry.

1.6.8 The Americas

The Americas were the last of the inhabited continents to be settled by humans, who reached them only $\sim$15-20 kya (Raghavan, et al. 2015; Skoglund and Reich 2016). The Siberian founder population appears to have been a mixture of around two thirds East Asian ancestry and one third of what has been referred to as "Ancient North Eurasian" ancestry (Patterson, et al. 2012; Raghavan, et al. 2014). The latter also later contributed to present-day European populations, which means that all Native Americans have a slightly closer affinity to Europeans than East Asians do. Once it had expanded out of Beringia, the population seems to have spread rapidly and colonized most of both North and South America very quickly. The entry into the Americas was associated with a bottleneck, such that Native American populations have the lowest levels of genetic diversity of any continental group today. Following European colonization in the last 500 years, there has been massive admixture from European and African sources, such that much of pre-Colombian population structure is likely obscured (Moreno-Estrada, et al. 2013; Moreno-Estrada, et al. 2014; Homburger, et al. 2015).

1.6.9 The Pacific

The last major part of the world, though not a continent, to be populated by humans was the large number of islands in the Pacific Ocean. An expansion of seafaring agriculturalists from Southeast Asia led to the peopling of the more remote islands only in the last few thousand years. The ancestry of these Polynesian peoples is thus largely East Asian. However, they also derive approximately 20% of their ancestry from Papuan or Melanesian sources, likely picked up by admixture during the expansion (Kayser, et al. 2008; Wollstein, et al. 2010; Skoglund, et al. 2016). It has still not been conclusively determined if the Polynesians reached the Americas prior to the era of European colonization. There has at least been no evidence for any genetic contribution to present-day Native American populations. However, the Polynesian population on Rapa Nui (Easter Island) has been found to harbour Native American admixture (Moreno-Mayar, et al. 2014), thus at least suggesting contact.

1.6.10 The effects of lifestyle on population history

Besides the basic questions of when and where different events in the population history of humans occurred, another key set of questions relate to the forces that were driving these events. On a very general level, it is clear that some parts of the world are more suitable for human occupation than others due to differences in geography and climate, with e.g. the cold of eastern Siberia explaining why the Americas remained unpopulated for so long. Some research utilizing data on past climate changes has suggested that climate is even a primary determinant of human migration patterns and timing (Eriksson, et al. 2012; Timmermann and Friedrich 2016). It is also worth noting however that hunter-gatherer humans appear to have been highly mobile once inside new continents, e.g. seemingly spreading across all of Australia in a timeframe too short to be tracked archeologically, and likewise reaching the very south of the Americas within at most a few thousand years after having entered this geographically and climatically very diverse region (Dillehay, et al. 2008).

Recent research, especially utilizing ancient DNA, is increasingly indicating that culture has also been a primary force shaping human population history, at least during the Holocene (the last $\sim$12 ky) (Gunther and Jakobsson 2016). The most well-studied examples of this come from Europe where, as mentioned above, the genetic landscape was dramatically reshaped by two major, culturally driven migration events, the first an expansion of agriculturalists and the second of Bronze Age pastoralists. The genetic histories of South Asia and sub-Saharan Africa, also as mentioned above, seem to have been massively affected by Holocene population movement and admixture as well. These studies are providing some answers to the long-standing question at the intersection of population genetics, archaeology, anthropology and linguistics about whether cultural practices, innovations as well as languages, particularly the transition from a hunter-gatherer to an agricultural lifestyle, spread through the horizontal transmission of ideas or through the movement and admixture of people: the "pots or people" debate. While these results are mainly favouring the spread of people, or "demic diffusion" (Moorjani, et al. 2013; Pickrell, et al. 2014; Allentoft, et al. 2015; Haak, et al. 2015; Patin, et al. 2017), some studies have also found evidence for lifestyle change without major genetic change in certain parts of the world (Siska, et al. 2017).

1.6.11 Adaptation to local environments

The selective pressures acting on human populations have varied over space and time, as groups in different parts of the world have settled in new environments or changed lifestyles. This has led to positive selection for genetic variants that confer adaptive advantages (Fan, et al. 2016). Variants conferring a lighter skin colour have been selected for in populations living at higher latitudes, possibly due to the need for UV radiation in vitamin D biosynthesis (Jablonski and Chaplin 2010). Populations living at high altitude in the Himalayas, Andes and the Ethiopian highlands have adapted genetically to the low oxygen levels of these environments (Bigham 2016). The short stature of certain central-African and Southeast Asian hunter-gatherer groups has been hypothesized to be an adaptation to life in a rainforest environment (Migliano, et al. 2013; Perry, et al. 2014). Pathogens appear to have constituted an important selective pressure during human evolution (Karlsson, et al. 2014), perhaps due Red Queen co-evolution dynamics (Siddle and Quintana-Murci 2014). Changes in diet also appear to have led to several instances of positive selection (Luca, et al. 2010), including adaptation to high fat diets in arctic populations (Fumagalli, et al. 2015) and to milk consumption in agriculturalist or pastoralists communities (Tishkoff, et al. 2007). There is also evidence for potentially widespread polygenic adaptation in human evolutionary history, in which large numbers of variants with small effects undergo small increases in frequency (Berg and Coop 2014; Field, et al. 2016).

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 $\sim$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, $\sim$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 $\sim$45-ky-old modern human (Fu, et al. 2014). This mutation rate is similar to that estimated from present-day Icelandic patrilines (Helgason, et al. 2015). This analysis resulted in divergence estimates of 54.3 ky (95% confidence interval (CI): 48.0–61.6 ky) between Sahul K$*$/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.

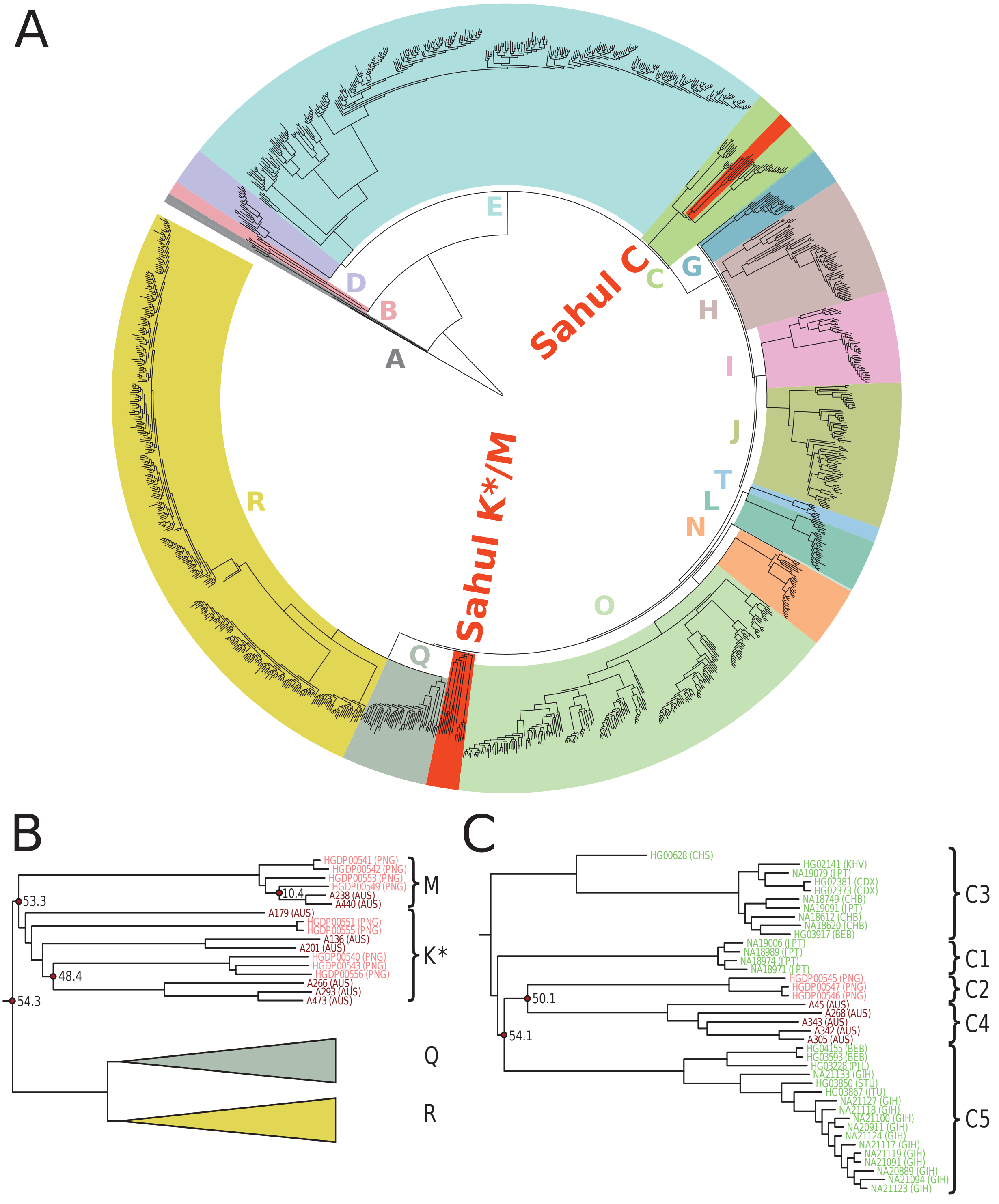

Figure 2.1: Phylogenetic History of Aboriginal Australian Y Chromosomes. *A) A maximum likelihood phylogeny of 1269 human Y chromosomes, including Aboriginal Australian and Papuan chromosomes highlighted in red, inferred using RAxML (Stamatakis 2014) from short-read sequencing data mapped to the ∼10 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.

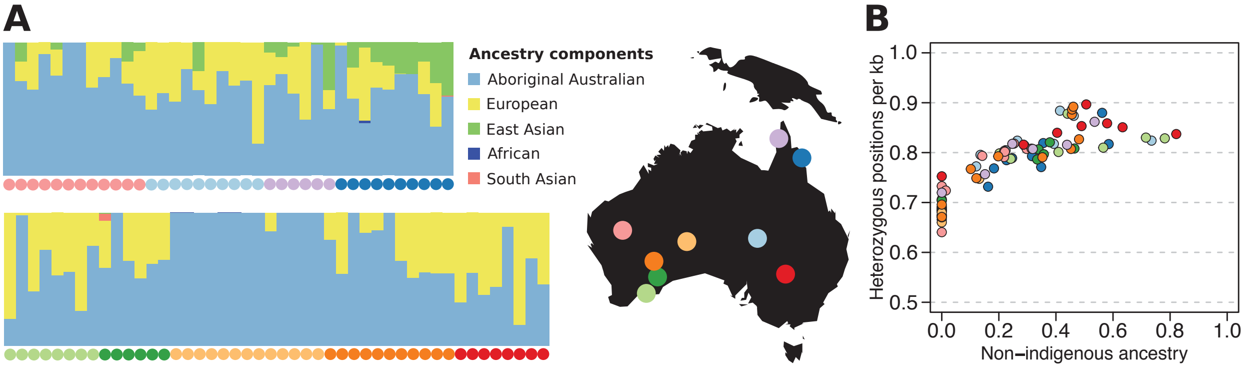

__Figure 2.2: Genome-wide ancestry of Aboriginal Australian individuals.__ Individuals are coloured according to their population origin, following the map in middle of the figure. A). The ADMIXTURE software was run with k=5 on Aboriginal Australians together with a worldwide panel including Europeans, 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.

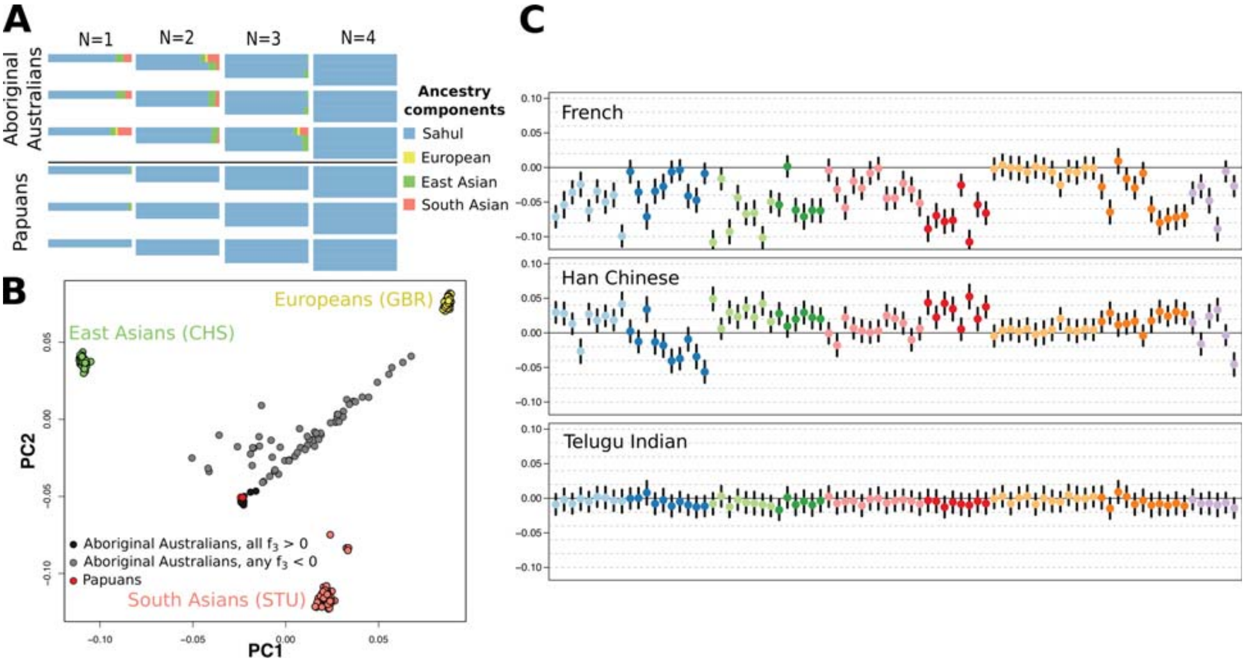

*Figure 2.3: **Testing for South Asian admixture in Aboriginal Australian genomes.** A) A down-sampling analysis of the effect of sample size on the assignment of non-Sahul ancestry components to Aboriginal Australian and Papuan genomes by the ADMIXTURE method. Down-sampling of the number of individuals 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

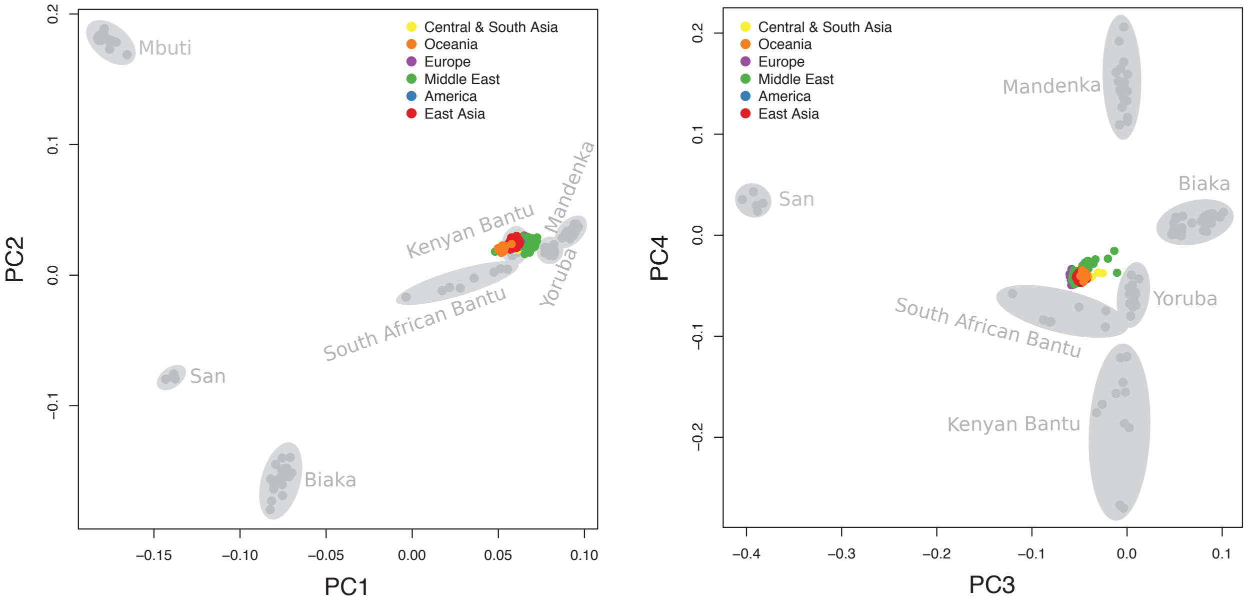

Figure 2.4: Relationships of non-Africans to Africans. *A PCA where the components were calculated only using African genotypes, and the non-Africans were projected into the resulting space, using Illumina 650K genotype array data from the HGDP-CEPH panel. Africans are displayed in grey, with population 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.

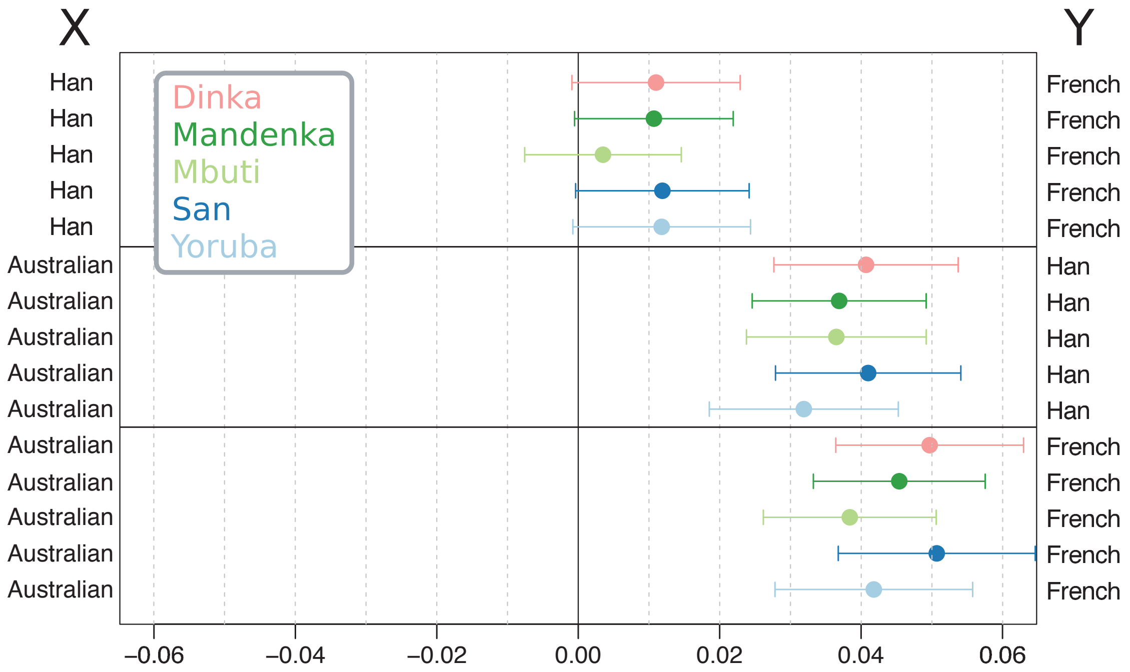

Figure 2.5: ***D-statistic tests of the relationships between Africans and non-Africans.*** *Tests of the form* D*(Chimp,African;X,Y), where African is each of the populations indicated in the grey box, and X and Y are the populations indicated on the left and the right side, respectively. A negative 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).

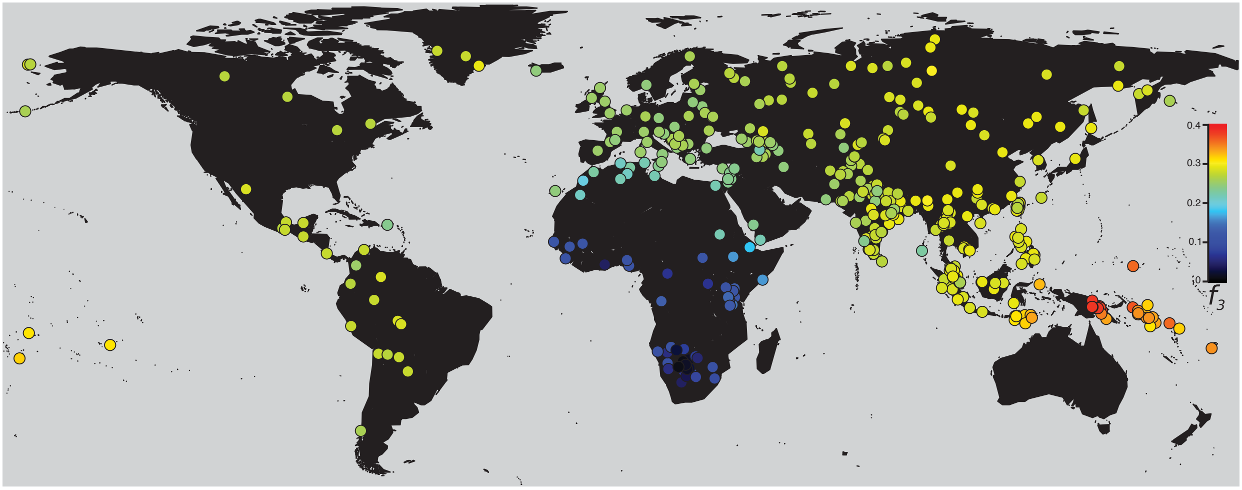

Figure 2.6: Genetic affinities of Aboriginal Australians to worldwide populations. The outgroup f_3*-statistic* f_3*(Aboriginal Australian,X;San) quantifies the amount of shared drift between Aboriginal Australians and worldwide populations X, relative to the southern African San population (red indicates 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).

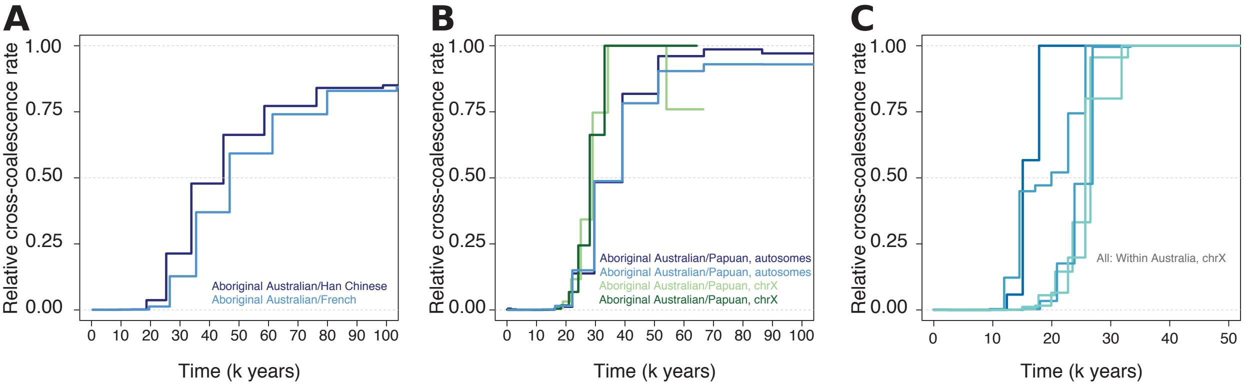

*Figure 2.7: **Timing of divergence between Aboriginal Australians from other populations.** A) Cross-coalescence curves between Aboriginal Australians and Eurasians suggests a split time of ∼45 kya from East Asians, and slightly older from Europeans. These were computed using MSMC2 on two genomes per 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).

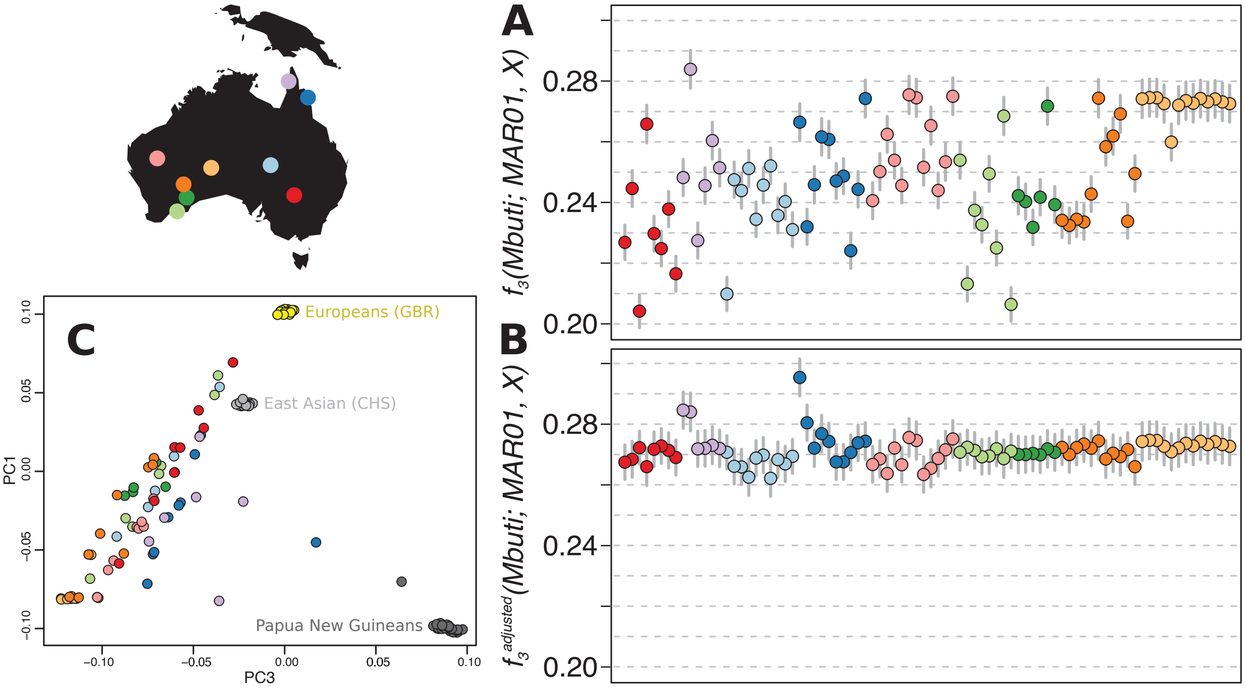

Figure 2.8: Aboriginal Australian genetic affinities to Papua New Guineans. *Individuals are coloured according to their population origin, following the map in the upper-left corner. A) Unadjusted outgroup* f$_3$*-statistics measuring genetic affinities of Aboriginal Australians to the arbitrarily chosen 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$_3$ ∼ *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 ∼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 ∼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 ∼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.

Chapter 3: The genetic history of Papua New Guinea

3.1 Introduction

The island of New Guinea constitutes the northern part of the Sahul continent, just south of the equator. Geographically and climatically it differs greatly from the very dry landscapes that make up much of Australia, being mostly covered in tropical rainforest, wetlands and grasslands. A mountain chain runs through the centre of the island, its highest peaks reaching over 4,000 meters but most of it situated at 1,000-2,000 meters. The archaeological evidence for human occupation in New Guinea is about as old as that for Australia, going back to $\sim$50 kya (Summerhayes, et al. 2010; O'Connell and Allen 2015), consistent with deriving from the same colonization event of the continent. Today, the island is politically split into two parts, the western half being part of Indonesia while the eastern part, plus a number of nearby islands of Melanesia, constitute the nation of Papua New Guinea (PNG). PNG is very culturally and linguistically diverse with approximately 850 different language groups, which is more than any other country and represents over 10% of all languages in the world. It is not known if this diversity is reflected in strong population structure on the genetic level.

New Guinea was one of the handful of places in the world where humans developed agriculture, in all cases in the Holocene, i.e. approximately the last 12 ky. The timing of this development in New Guinea is not very firmly established, but is estimated at approximately 10 kya (Denham and Barton 2006). Little is known about where within New Guinea agriculture originated, but the consensus is that it was somewhere in the highlands. Because of their elevation, the highlands are relatively cool and with plenty of rain and fertile soils. It has been hypothesized that the largest language family of New Guinea, the so called Trans-New Guinea (TNG) family, which is spoken across all of the highlands and large parts of the lowlands, spread alongside the spread of agriculture (Pawley 2005). While early influences from Southeast Asia cannot be completely ruled out, and there have been later external influences e.g. in the introduction of pigs and the sweet potato, it appears that the development of agriculture in New Guinea was indigenous and independent from developments in the rest of the world. There is therefore an opportunity to compare the population history of New Guinea with the histories of other parts of the world, in particular with respect to how it was shaped by the development of agriculture. Such a comparative view across independent histories might even allow us to say something about the "reproducibility" of human evolutionary trajectories.

In the last few thousand years, New Guinea saw an influx of people from Southeast Asia, leading to genetic admixture and the introduction of Austronesian languages. This is one of the world's largest language families, spoken by many different groups across Southeast Asia and Oceania, and was spread by massive seaborne migrations by Southeast Asian agriculturalists (Duggan and Stoneking 2014; Skoglund, et al. 2016). In New Guinea, Austronesian languages are fairly restricted to certain coastal areas. Genetic studies so far have documented Southeast Asian admixture in the lowlands of New Guinea, but not in the highlands (Stoneking, et al. 1990). However, most studies have been small both in terms of the number of samples and the number of markers assayed, most being limited to the uniparental chromosomes, such that the extent of Southeast Asian admixture is not understood in any greater detail. In particular, it has arguably not been conclusively established if the highlands indeed have been isolated from non-New Guinean gene flow.

In this chapter, I primarily make use of two different datasets to study the population history of Papua New Guinea, both generated at the Wellcome Trust Sanger Institute. The first is a set of 39 high-coverage, whole genome sequences from Papua New Guinea, also used in Chapter 2. The second is a dataset of genome-wide array genotypes from 381 individuals from Papua New Guinea, generated specifically for this study and published in 2017 (Bergström, et al. 2017).

3.2 Array genotyping of Papua New Guineans

381 individuals were selected from a set of ∼800 PNG DNA samples collected in the early 1980's and stored at the University of Oxford, and genotyped on the Illumina Infinium Multi-Ethnic Global array. This array contains 1.78 million markers and its marker content has been ascertained in a less European-centric manner than most arrays commonly used in the past; however, less than 600,000 of these were polymorphic in this set of PNG samples. This is, however, still sufficient for high-resolution population genetic analyses. After quality control at the individual and site levels, 378 individuals and 529,137 variants remained for downstream analyses. The sample set is very comprehensive, covering approximately 85 different and geographically diverse language groups within PNG (Figure 3.1).

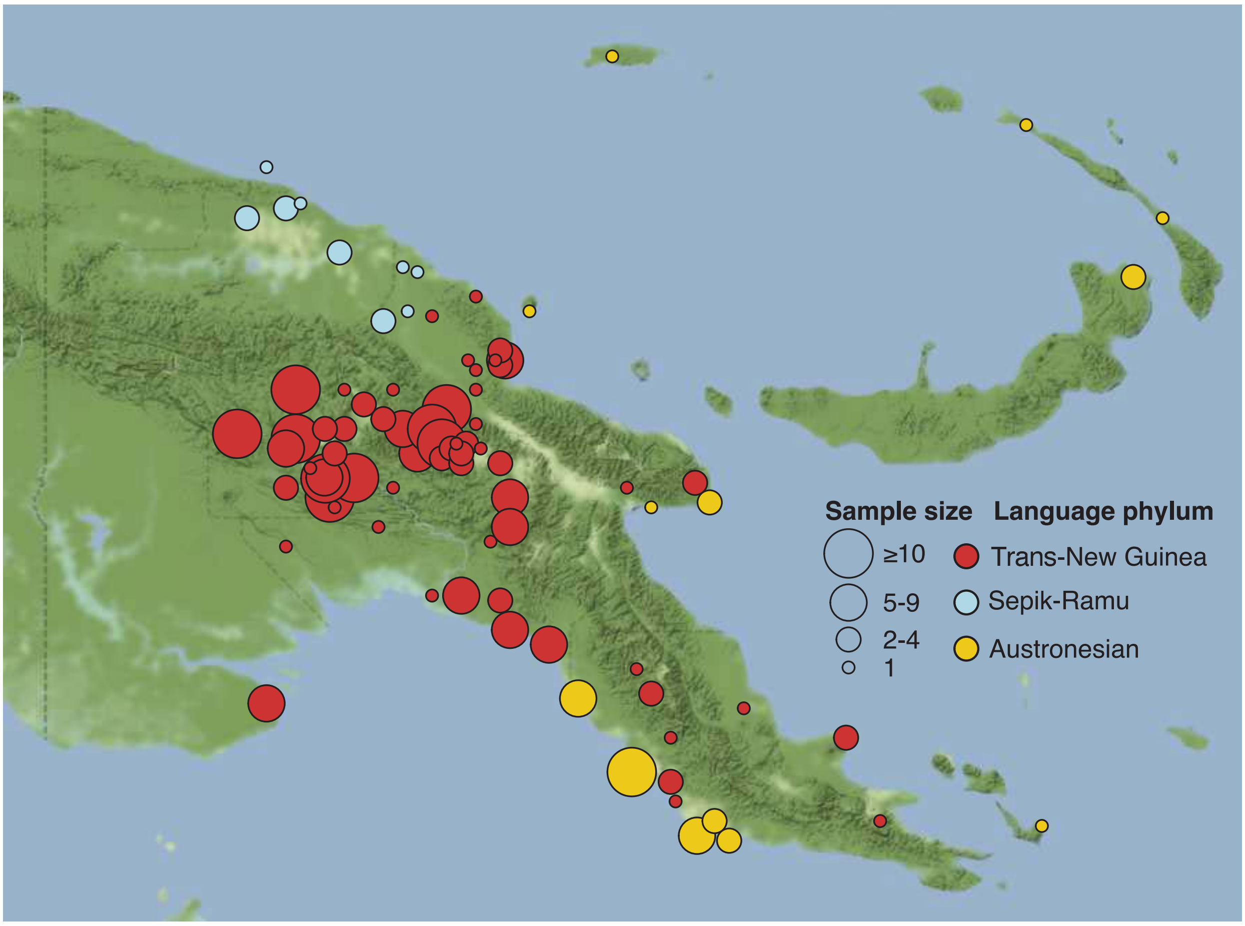

*Figure 3.1: **Genotyped samples in Papua New Guinea.** Each sampled language group is represented by a circle, the area of which indicates the number of individuals sampled in the group and the colour of which represents the top-level language phylum/family. An additional 39 individuals are not included in this figure as their exact language group is not known or their parents are known to come from different language groups. The base map was obtained from Stamen (http://maps.stamen.com/, under Creative Commons BY 3.0. Data by OpenStreetMap, http://www.openstreetmap.org/, under Open Data Commons Open Database License (ODbL)).*

3.3 Genetic identity of the HGDP-CEPH Papua New Guinean individuals

The widely used HGDP-CEPH collection of DNA samples from diverse, worldwide populations (Cann, et al. 2002) contains 17 individuals from Papua New Guinea. When analysed in the context of the comprehensive set of array genotyped PNG samples, it is clear that these 17 individuals actually consist of two subsets with very different genetic affinities. One set has affinities to groups in the East Sepik region of the northern PNG lowlands, consistent with the geographical coordinates provided in the HGDP-CEPH metadata, and the other set has affinities to groups in the eastern highlands. This finding has important implications for the use of these samples in analyses, particularly as I make use of whole-genome sequencing data from 14 of them. In what follows I refer to the lowlander subset as “HGDP_L” and the highlander subset as “HGDP_H”.

3.4 Southeast Asian admixture in PNG

The array genotype dataset reveals a very variable impact of Southeast Asian, or other non-New

Guinean, gene flow in PNG. Firstly, only two individuals displayed evidence of European ancestry, in sharp contrast with neighbouring Australia and consistent with the small impact of European colonialism on PNG. Secondly, there is a lack of Southeast Asian admixture in the highlands, as determined using ADMIXTURE as well as per-individual D-statistic tests of the form D(Yoruba, East Asian; Aboriginal Australian, PNG individual). Only four highlanders display East Asian ancestry in these analyses; however, principal component analyses of the relationships between highlanders and lowlanders show that all four of these individuals have lowlander affinities in the Papuan component of their genomes, such that they very likely reflect recent movement into the highlands from lowland groups. Consistently, analysis of the mitochondrial and Y chromosome variants included on the array showed that none of the sampled highlanders carried uniparental chromosomes of recent non-New Guinean origin. In summary, there is thus no evidence for Southeast Asian gene flow into the highlands, implying genetic independence of highlander ancestry from non-Sahul sources from the initial peopling of the continent until the present day (Figure 3.2).

Lastly, Southeast Asian admixture is present in all parts of the lowlands but in highly variable amounts across individuals and regions. Individuals speaking Austronesian languages have substantially higher amounts of Southeast Asian ancestry than those speaking non-Austronesian languages (mean of 38.7% vs 11.6%, $p = 1.4\times10^{-13}$, Wilcoxon rank sum test). Speakers of the Sepik-Ramu language family, which is an indigenous language family unrelated to the larger Trans-New Guinea family and spoken only in parts of the northern lowlands around the Sepik and Ramu rivers, display the lowest amounts of admixture (average of 4.3%). These results thus demonstrate that, while all coastal areas of New Guinea would have been accessible to the seafaring Southeast Asian migrants, the genetic impact was very variable.

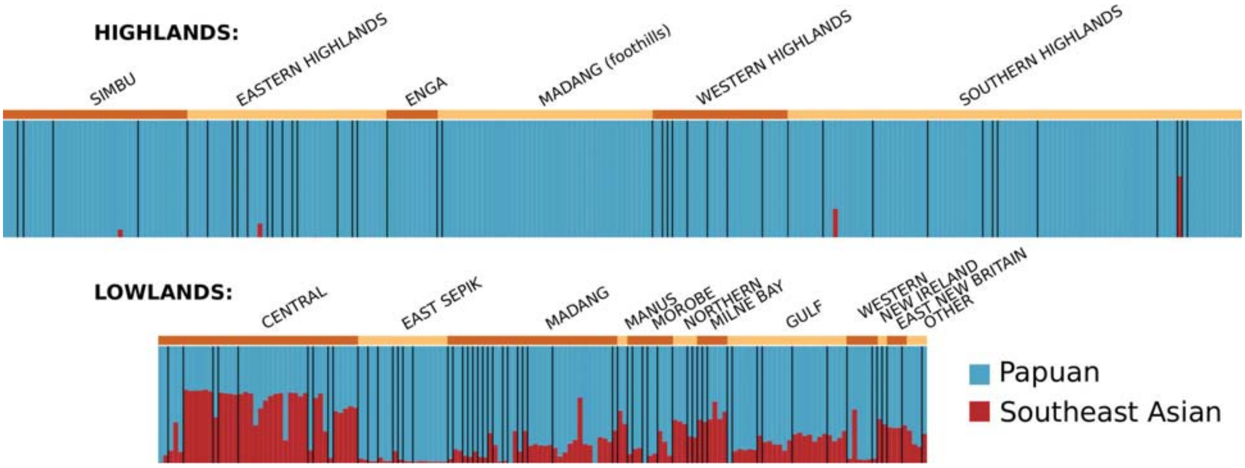

*Figure 3.2: **Southeast Asian admixture in PNG.** ADMIXTURE was run at K=2 together with the 504 East Asian individuals from the 1000 Genomes Project (not displayed) to estimate the Papuan and Southeast Asian ancestry proportions across sampled PNG individuals. Individuals are grouped by language group, separated by vertical black lines, and then by province. The estimated ancestry proportions correlate strongly with those estimated using* f_4*-ratios.*

3.5 The relationship to Aboriginal Australians

As discussed in Chapter 2, the genetic separation between Aboriginal Australians and Papuans appears to have occurred relatively early, long before the geographical separation of Australia and New Guinea following rising sea levels. Furthermore, Aboriginal Australians across different regions of Australia displayed a largely uniform relationship to Papuans, implying shared ancestry across Australia after the separation from Papuans, and at most limited gene flow from New Guinea after that point. However, it is not known what corresponding situation is on the other side of the Torres Strait, i.e. if all Papuans are uniformly related to Aboriginal Australians, or if, for example, groups on the southern coast have higher affinity, e.g. due to gene flow or ancestral population structure. In the array genotype dataset, no individual has a directly visible higher affinity to Aboriginal Australians than the highlanders do, i.e. there are no significantly negative values of the statistic *D*(Yoruba,Aboriginal Australian;X,Highlander). However, the effect of the widespread Southeast Asian admixture present in most lowlander genomes is to decrease affinity to Aboriginal Australians, thereby counteracting the ability of a test like this to detect a higher affinity. I therefore performed analyses aiming at assessing affinity to Aboriginal Australians while accounting for the Southeast Asian admixture.

First, in a PCA constructed using only PNG highlanders, Aboriginal Australians and East Asians (using 1000 Genomes Han Chinese as a proxy for the Southeast Asian ancestry present in New Guinea), the position of a sample projected into the resulting space should be informative about its relationship to these three ancestral poles. Most individuals line up along the arc between the highlander and East Asian corners, suggesting that the Sahul component of their ancestry is not any closer to Aboriginal Australians than highlander ancestry is (Figure 3.3C). Deviating from this trend are the individuals from the large islands of New Britain and New Ireland, as well as individuals from Western Province (the southernmost samples in our dataset). This could reflect a closer relationship to Aboriginal Australians in these groups, and at least superficially appears to mirror such a relationship reported for the island population of Tonga, further out into the Pacific (Skoglund, et al. 2016).

Second, data from each sampled individual in turn was used to fit an admixture graph (Patterson, et al. 2012) modelling their ancestry as a mixture of one source related to PNG Highlanders and a second source related to East Asians (Figure 3.3A). The key aspect of this graph in the context of the relationship to Aboriginal Australians is that it shows all the Sahul-related ancestry of the modelled individual coming from the same lineage as PNG highlander ancestry, with Aboriginal Australians being a strict outgroup to this lineage. Any extra affinity to Aboriginal Australians would lead to a rejection of this simple model. This graph fit the data with no outlier ($|Z|>3$) *f*-statistics for any except three individuals. Two of the individuals for which the model did not fit had European admixture, and the last is a highlander who is an ancestry outlier relative to their language group. A similar admixture graph was also tested using whole-genome sequencing data from the

Simons Genome Diversity Project (Mallick, et al. 2016), representing an independent dataset, with the Bougainville Islanders (who are not present in the array dataset) as the test population (Figure 3.3B). The simple graph with Aboriginal Australians being an outgroup to the Sahul ancestry also fits the data for Bougainville Islanders.

Last, *D*-statistics were directly plotted against each other to allow visual examination of how East Asian admixture impacts affinity to Aboriginal Australians (Figure 3.3D). The statistic *D*(Yoruba,Vietnamese;X,Chinese) measures the amount of Sahul ancestry, as opposed to East Asian ancestry, and the statistic *D*(Yoruba,Aboriginal Australian;X,Highlander) measures the affinity to Aboriginal Australians. If the Sahul component in some individuals has a stronger affinity to Aboriginal Australians, we would expect them to depart from the trend line (i.e. a lower value of the latter *D*-statistic than expected given their value of the former *D*-statistic). No individuals departed from the trend line, suggesting a uniform relationship to Aboriginal Australians.

The lack of signal in the admixture graph and *D*-statistics analyses suggest that the pull of New Ireland and New Britain groups towards Aboriginal Australians in PCA might be an artefact, perhaps related to the likely fairly deep divergence of these populations from mainland populations. The signal reported for Tongan islanders (Skoglund, et al. 2016) thus seems to be absent from the populations sampled here, closer to the New Guinea mainland. The PCA position of the Western Province Southern Kiwai individuals, while intriguing given their geographical proximity to Australia, also has no support in the formal statistics. In summary, there is no strong evidence for any of the sampled groups displaying a closer relationship to Aboriginal Australians than anyone else. This mirrors the uniform relationship to Papuans found among Aboriginal Australians, and sets Sahul apart from many other parts of the world where genetic gradients across geographical space tends to be the norm.

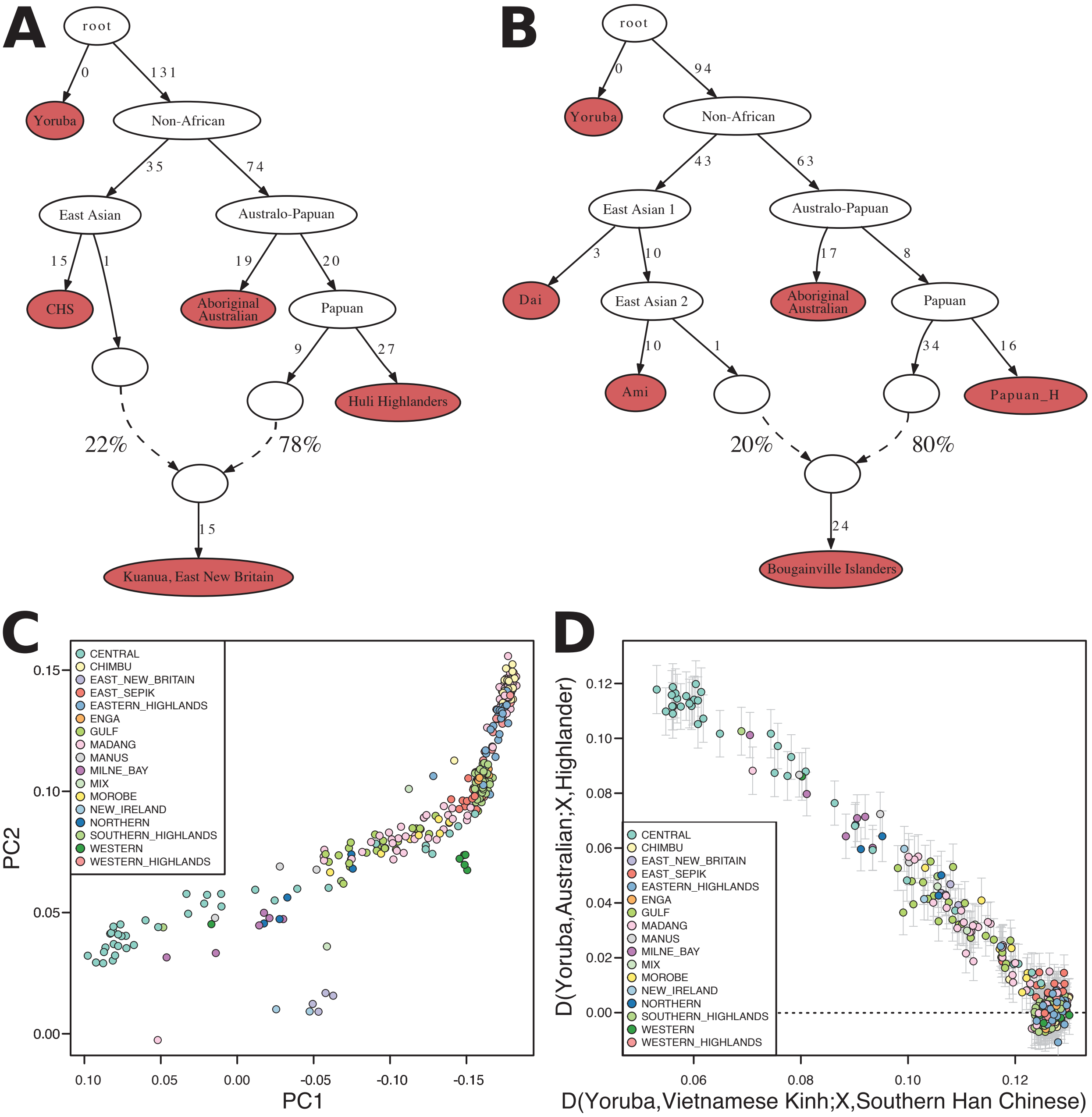

Figure 3.3: The relationship of PNG individuals to Aboriginal Australians. *A) An admixture graph, modelling the ancestry of an individual as a mixture of one source related to PNG highlanders and one source related to East Asians, fits the data for all except three outliers. B) A similar admixture graph using data from the SGDP fits the Bougainville Island population. C) A principal components analysis where only two PNG Highlanders, two Aboriginal Australians and two East Asian individuals are used to define the space, defining ancestral poles in the top-right, bottom-right and left part of the plot, respectively (but not displayed on the plot themselves). The rest of the individuals are then projected into this space. D) Two* D*-statistics plotted against each other, the one on the horizontal axis measuring the amount of Sahul, as opposed to East Asian, ancestry, and the one on the vertical axis measuring affinity to Aboriginal Australians.*

3.6 Local ancestry inference in PNG

The widespread Southeast Asian admixture present in the lowlands of PNG would confound many analyses, including comparisons between different groups within PNG. The local ancestry classification and subsequent masking of haplotypes has proven successful in recently admixed populations such as Latin- and African Americans where the ancestry blocks are still long (Brisbin, et al. 2012), but has not been widely applied in PNG or Melanesian populations where the admixture is likely to be at least 3,000 years old. Furthermore, current reference panels for haplotype phasing, which is a requirement for most local ancestry inference methods, do not include haplotypes from this part of the world. However, phasing haplotypes against the very large HRC panel (McCarthy, et al. 2016) and inferring local ancestry using RFMix (Maples, et al. 2013), using unadmixed PNG highlanders and the 1000 Genomes East Asians as the two reference populations, proved surprisingly successful in the PNG array genotype dataset. Validation experiments in which highlanders and East Asians were held out from the reference panels and then subjected to the same local ancestry inference indicated an ancestry misclassification rate of less than 0.5%. Furthermore, the overall ancestry proportions obtained by summing the lengths of all haplotypes of each ancestry correlated strongly with those obtained from global ancestry estimation methods (Pearson correlation between RFMix and ADMIXTURE = 0.997, between RFMix and f_4-ratio = 0.982). The masked genotypes were therefore used in many downstream analyses, thereby avoiding the confounding effects of Southeast Asian admixture to a large extent.

3.7 Population structure in PNG

The strongest genetic separation within PNG appears to be that between groups on the mainland and those on the large islands of the Bismarck Archipelago, New Ireland and New Britain, consistent with previous studies hinting at a potentially fairly old split between these (Friedlaender, et al. 2008). Within the highlands, there is very clear clustering into a western cluster, an eastern cluster and one cluster corresponding to a small set of Angan language groups living in the south-eastern highlands, the latter likely a case of genetic isolation. Within the lowlands, there is clear separation between the south coast and north coast (Figure 3.4).

I performed various analyses to determine the nature of the relationship between highlanders and lowlanders. When projected into a PCA space constructed using only highlander genotypes, all lowlander individuals project into largely the same part of the plot (Figure 3.4A). It is not the case that, for example, northern lowlanders project closer to northern highlanders, and southern lowlanders closer to southern highlanders. In the inverse experiment, all highlanders similarly project into largely the same part of the space constructed using only lowlander genotypes (Figure 3.4B). In both cases there are a handful of outlier individuals who are drawn towards particular parts of the space, but these likely reflect very recent admixture between the highlands and the lowlands (in some cases confirmed by the documentation on the sample origins). It thus appears that population structure in the highlands is largely independent of that in the lowlands.

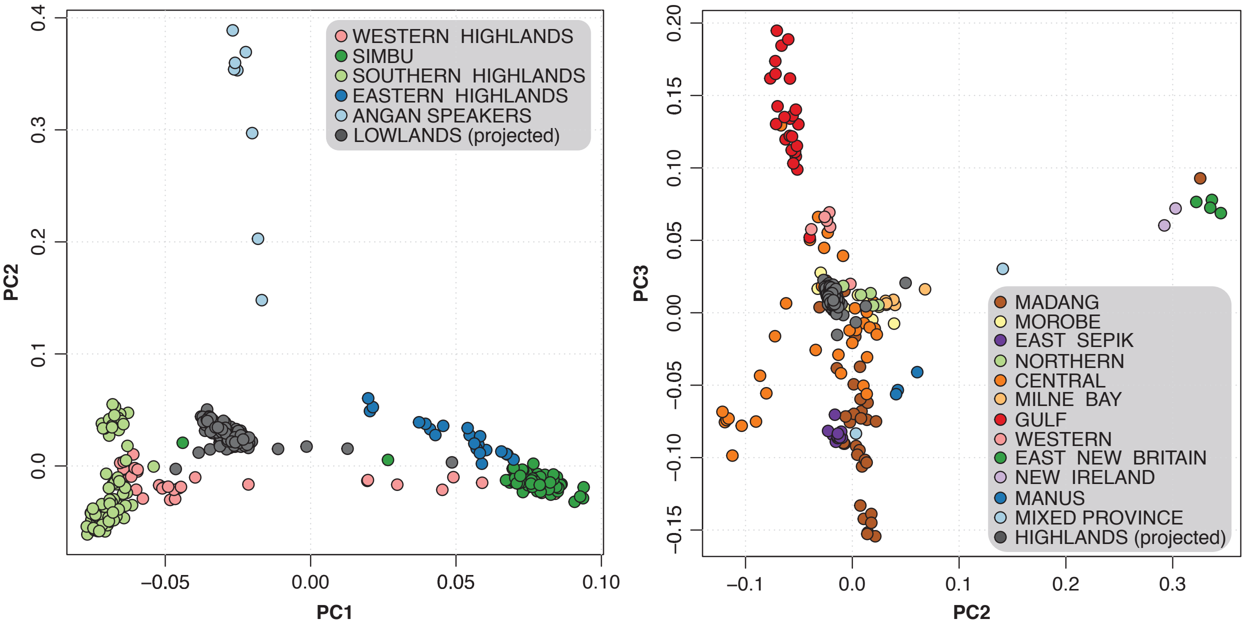

Figure 3.4: Population structure of the highlands and the lowlands of PNG. *A) A PCA constructed using only highlander genotypes reveals a strong division between eastern and western groups (PC1), as well as separation of Angan speaking groups (PC2). When projected into this space, all lowlanders group largely uniformly, except a few outliers. B) A PCA constructed using only lowlander genotypes reveals a strong separation between groups on the mainland and those on the large islands of New Britain and New Ireland (PC2), as well as separation between the north and the south coast (PC3). PC1 corresponds to Southeast Asian admixture (not shown). When projected into this space, all highlanders group largely together, except a few outliers.*

To further study the relationships between highlanders and lowlanders, four sets of all-against-all *D*-statistics were computed and visualized using Q-Q (quantile-quantile) plots, comparing the Z-scores obtained to those expected by chance under a normal distribution with mean 0, given the number of tests performed. These were calculated on data locally masked for East Asian ancestry (and excluding the PCA outliers mentioned above), in all cases with the African Yoruba as an outgroup:

- *D*(Yoruba,Lowlander;Highlander1,Highlander2) (Figure 3.5A): this statistic asks for each lowlander group if it shares more with one highlander group than another. While some values of this statistic reach the seemingly significant |Z|>4, the close fit to the diagonal line in the Q-Q plot demonstrates that this is expected by chance given the large number of tests performed. The results thus give no evidence for any lowlander group having a higher affinity to one highlander group over another.

- *D*(Yoruba,Highlander;Lowlander1,Lowlander2) (Figure 3.5B): this statistic asks for each highlander group if it shares more with one lowlander group than another. The deviation from the diagonal in the Q-Q plot indicates that this is sometimes the case. Most of the signal is driven by highlanders being closer to mainland lowland populations than to the divergent New Ireland and New Britain island populations, but some signal remains even when excluding these.

- *D*(Yoruba,Highlander1;Highlander2,Lowlander) (Figure 3.5C): this statistic asks for each highlander group if it is closer to another highlander group than to a lowlander group. The massive shift towards negative values indicates strong highlander genetic unity to the exclusion of lowlanders, and despite the large numbers of tests performed there is not a single test in which a highlander group is significantly closer to a lowlander than to another highlander group (largest Z=2.69).

- *D*(Yoruba,Lowlander1;Lowlander2,Highlander) (Figure 3.5D): this statistic asks for each lowlander group if it is closer to another lowlander group than to a highlander group. The deviation from the diagonal in the Q-Q plot indicates that this is not always the case. In other words, while there is highlander genetic unity, there is no lowlander genetic unity. Much of the signal is driven by most lowlanders being closer to highlanders than to the divergent New Ireland and New Britain island populations, but considerable signal remains even when excluding these. This is driven mainly by northern lowland groups being closer to highlanders than to southern lowland groups (example: *D*(Yoruba,Wagi Northern Lowlanders;Waima Southern Lowlanders,Aiya Highlanders) = 0.0095, Z = 3.074), and southern lowland groups being closer to highlanders than to northern lowland groups (example: *D*(Yoruba,Waima Southern Lowlanders;Wagi Northern Lowlanders,Aiya Highlanders) = 0.0105, Z = 3.576).

Lastly, following up on the observation that highlanders are not equally similar to all lowlander groups, outgroup f_3-statistics demonstrate that the set of groups that highlanders are the most similar to are those from the East Sepik area in the north-western lowlands (Figure 3.5F). This is surprising from a linguistic point of view, as it is the only sampled area where the widespread Trans-New Guinea language family is not spoken, instead being dominated by languages from the independent Sepik-Ramu family. There is, however, archaeological evidence for the transfer of items between the highlands and the Sepik region during the Holocene (Swadling, et al. 2008).

In summary, these results reveal a striking picture where all highlanders, regardless of geographic location, can be described as being a clade to the exclusion of lowlanders. As a telling example of this, Gende speakers who live on the very northern edge of the highlands are more similar to all other highlander people than to Sop speakers living in the lowlands just 40 km to the northeast (Figure 3.5E). This very sharp division between highland and lowland groups departs dramatically from the gradual isolation-by-distance patterns that are typically seen in human populations (Lao, et al. 2008; Novembre, et al. 2008). Possible explanations include a recent expansion-and-replacement episode that homogenized the ancestry of all highlanders, or possibly a more subtle, long-term process where gene-flow has remained high within the highlands but very limited between the highlands and the lowlands. Lastly, it can be noted that it's possible that there is a lack of statistical power with the current dataset to for example reject the clade-like status of highlanders (e.g. in tests of the form D(Outgroup,Lowlander;Highander 1;Highlander 2)), and that denser genotypes and larger population samples would potentially allow such rejections. However,

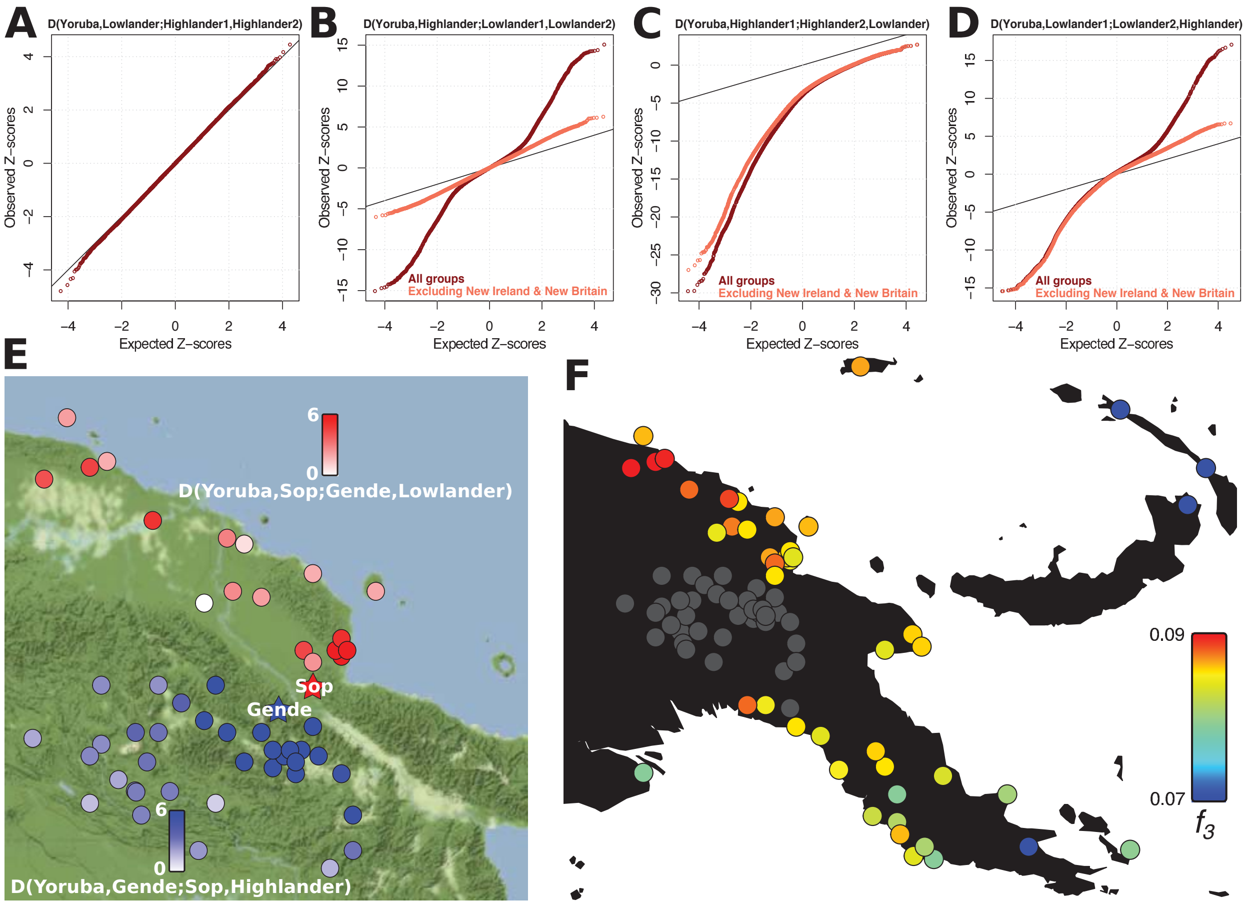

*Figure 3.5: **The genetic relationships between highlanders and lowlanders.** Lowlander genomes were locally masked for Southeast Asian ancestry. A-C: Quantile-quantile plots comparing distributions of Z-scores from D-statistics relating highlanders and lowlanders to those expected under a normal distribution. The African Yoruba population is used as an outgroup. A) Lowlanders are equally similar to different highlander groups. B) Highlanders have stronger affinity to some lowlander groups than to others. C) Highlanders are more similar to each other than to lowlanders. D) Lowlanders are not always more similar to each other than to highlanders. E) The Z-scores of two different* D*-statistics, the first measuring if the highland Gende speakers are more similar to the lowland Sop speakers or to other highlanders (blue meaning more highlander similarity), and the second if Sop speakers are more similar to Gende speakers or to other lowlanders (red meaning more lowlander similarity). Z-scores were capped at 6. F) Genetic affinity of highlanders (treated as a single group, marked in grey) to different lowland groups measured by the outgroup* f_3 *statistic* f_3*(Highlanders,X;Aboriginal Australian) (red meaning higher affinity).*

even so, the overall picture and what it tells us about the relationships between highlanders and lowlanders would still likely remain the same.

3.8 The time depth of population separation in PNG

While array genotypes allow for the similarity between individuals and groups to be assessed, their use is limited when trying to measure the timing of events. Whole-genome sequencing is better-suited for this purpose, and we have generated this for 7 different PNG groups; six in the highlands and one in the lowlands (the “HGDP_L” East Sepik subset of the HGDP-CEPH PNG samples). I applied the MSMC method (Schiffels and Durbin 2014) to these data in order to estimate pairwise split times between groups, scaling results using a point mutation rate of 1.25 × 10^{-8} per site per generation and a generational interval time of 30.5 years (the latter deriving from an anthropological study of a PNG highlander group (Wood 1987)).

The MSMC method requires phased haplotypes, and it is becoming increasingly clear that insufficient phasing accuracy can have large impacts of inferences of split times (Song, et al. 2017). This is of particular concern in populations like Papuans, the ancestry of which is not represented at all in current haplotype reference panels used for statistical phasing. I took two approaches to tackle this issue here. First, using 10x Genomics linked-read technology (Zheng, et al. 2016) to physically infer phase for a subset of eight genomes from four groups. This is a library preparation technology that links a small barcode sequence to all DNA fragments that derive from the same, longer (∼10-100 kb) molecule fragment, such that reads obtained from standard Illumina sequencing can then be computationally linked together and haplotype phase inferred. Second, using the haploid and therefore necessarily perfectly phased X chromosomes of males, as in Chapter 2.

The separation between highlanders and the East Sepik lowlanders appears to have occurred between 10 and 20 kya, perhaps 15 kya, on the basis of the genomes phased using linked-read technology (Figure 3.6A). With statistically phased data, the curves are substantially shifted towards older times and inferred splits would appear several thousand years older (and furthermore there is quite a large difference between results obtained using 4 haplotypes versus using 8 haplotypes, not shown), likely as a consequence of poor phasing. The X chromosome results generally suggest more recent split times, but the variable and sometimes non-monotonic behaviour of the curves makes these results more difficult to interpret.

Separations between groups within the highlands are more recent, with all appearing to have occurred ∼10 kya or more recently (3.6B). Similarly to above, physically phased genomes and X chromosomes result in more recent splits relative to statistically phased genomes. The oldest splits occur between groups in the eastern and western highland clusters, consistent with the population structure inferred from the array genotype dataset. A split between two groups within the eastern cluster (Gende and HGDP_H) appears to be less than 5 ky old.

In summary, the age of population structure in PNG does not date back to the initial peopling of Sahul. While there is considerable technical uncertainty surrounding the estimated dates, both methodological and related to mutation rate and generation time estimates, the overall picture is one where the separation between highlands and lowland groups occurred 10-20 kya, and highland groups then separated from each other within the last 10 kya. The use of physical phasing through the 10x Genomics linked-read technology greatly aided these analyses, providing one way to overcome the problem of haplotype phasing in diverse human populations.

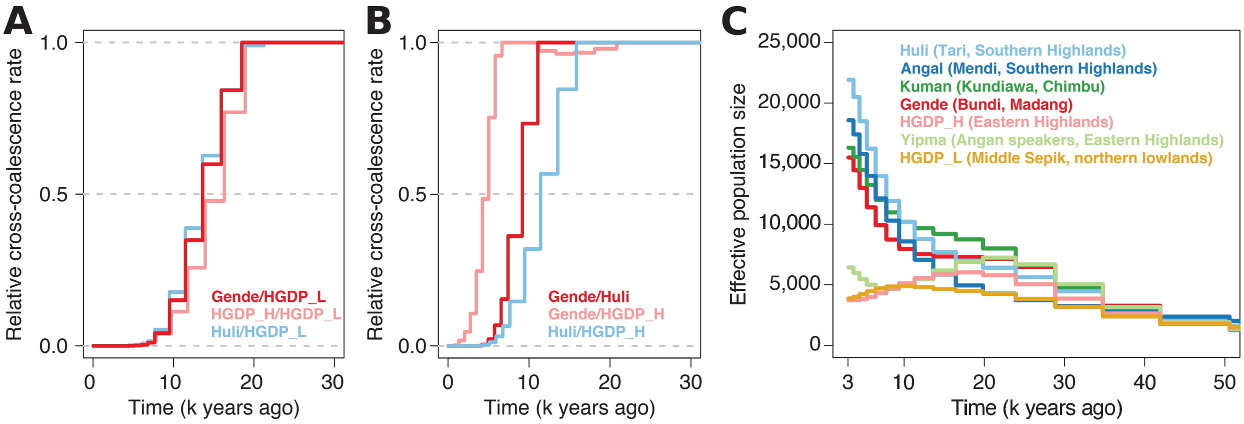

Figure 3.6: The time depth of population separation and growth in PNG. *(A) MSMC relative cross-coalescence curves between highland groups and a northern lowland (East Sepik) group suggests a split time between 10 and 20 kya. (B) MSMC relative cross-coalescence curves between highland groups suggest split times within the last $\sim$10 ky (Huli representing the western cluster, Gende and HGDP_H the eastern). These curves were inferred from genomes physically phased using linked-read sequencing. (C) Effective population size histories of different groups as inferred using SMC++ on five high-coverage genomes per group.*

3.9 Population size histories in PNG

The MSMC and the conceptually related SMC++ (Terhorst, et al. 2017) methods allow for effective population size histories to be inferred from whole genome sequences. The former requires haplotype phasing when run on more than one genome, while the latter utilizes only allele frequency information from additional genomes and therefore does not require phasing. Applied to PNG groups, the two methods give qualitatively similar results, though MSMC gives very high estimates of effective population sizes in the very recent times, which might be an artefact of the poor phasing quality of these genomes (not shown). The SMC++ results might thus be more reliable. They show that most highlander groups have experienced major growth of effective population sizes, especially in the last 10 kya (Figure 3.6C). Exceptions are the Angan-speaking Yipma, who in other analyses show signs of genetic isolation and heavy drift, and the HGDP_H group (the origin of which is not actually known, and also in contrast display a recent increase in the MSMC analysis). The Sepik lowlanders do not display the same growth, instead retaining a more or less constant effective population size. This is in line with anthropological records of lower population densities in the lowlands, and might be linked to widespread malaria in these regions (Riley 1983).

3.10 Genetic differentiation in PNG

Genetic differentiation between pairs of populations as measured using the F_{ST} statistic typically take values on the order of 1% or less between major, non-isolated populations within Europe, East Asia or South Asia (1000 Genomes Project Consortium 2015; Pagani, et al. 2016). Between European and East Asian populations, F_{ST} is approximately 10%.

Calculating F_{ST} between language groups in the PNG array genotype dataset after locally masking Southeast Asian ancestry (Figure 3.7), values between the western and eastern highlands clusters reach 2-3%, with all values within each of these clusters being below 2% but some being above 1%. Values between the Angan speaking groups in the south-eastern highlands and other highlander groups reach 4-5%, which is as high as between European and South Asian populations. This is all within a sampled area within the highlands that is approximately the same size as Denmark, or the Netherlands. Within both the southern and the northern lowland areas, levels of F_{ST} are as high as those in the highlands. This suggests that the mountainous terrain of the highlands, often cited as an explanation for the great cultural diversity of PNG, might not be what underlies differentiation between groups, instead raising the likely importance of cultural and linguistic factors. Between the highlands, southern lowlands and northern lowlands, F_{ST} values are even higher, with many over 5%. These results thus demonstrate that the great cultural and linguistic diversity of PNG is reflected in strong genetic differentiation.

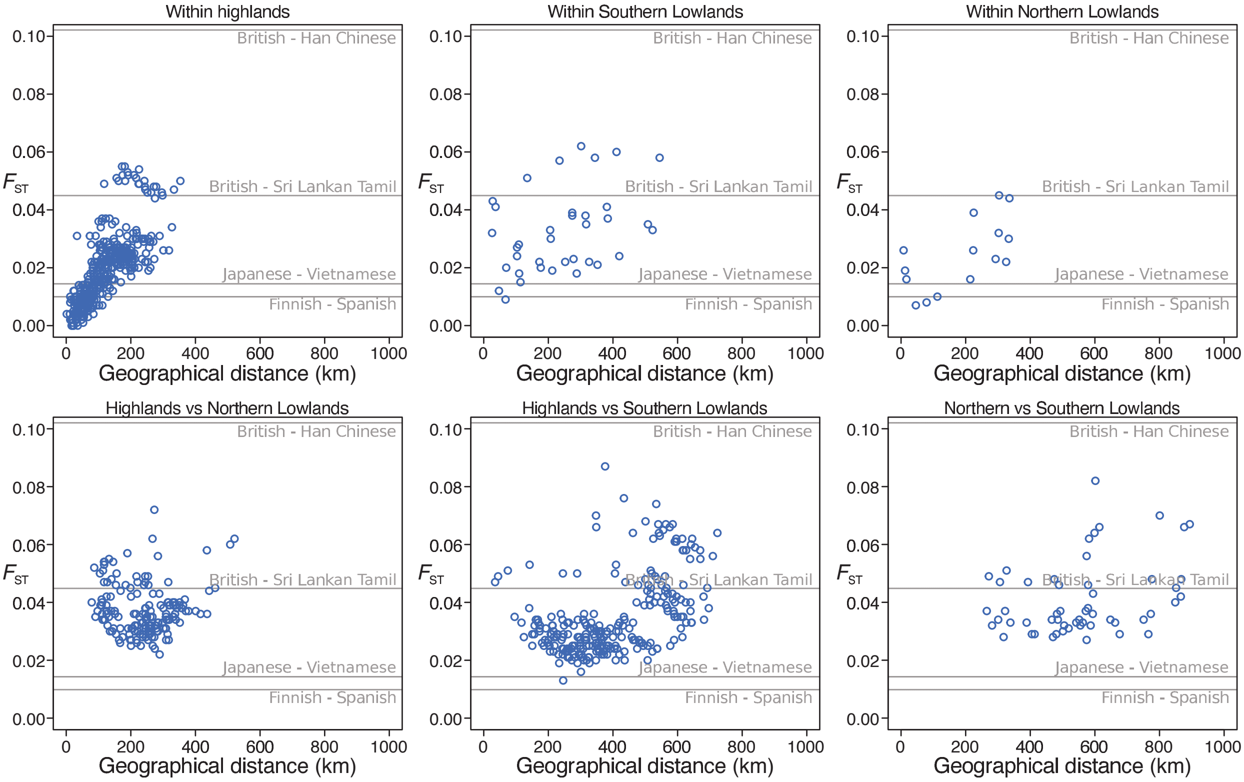

__Figure 3.7: Genetic differentiation in PNG.__ Geographic distance between groups is plotted against F_{ST}*, a measure of allele frequency differentiation. This was calculated after locally masking lowlander genomes for Southeast Asian ancestry, which otherwise has large effects on* F_{ST}*. Grey lines indicate a number of* F_{ST} *values between selected populations from the 1000 Genomes Project for comparison.*

Population differentiation is slightly stronger for the Y chromosome than for the mitochondrial genome (p = 0.0035, Wilcoxon signed rank test for difference in the mean of a F_{ST} metric for haploid loci). This implies that there is more female than male movement between groups and/or that male effective population sizes are smaller (e.g. due to larger variance in reproductive success), and is consistent with previous studies of New Guinean populations (Kayser, et al. 2003).

3.11 Diversity and isolation in PNG

The high differentiation between groups, the high linguistic diversity and the small present-day size of many PNG language groups raise questions about lifestyles and interaction patterns in PNG. While the coalescent-based approaches applied above (SMC++, MSMC) can reveal demographic histories deep into the past, they have limited resolution in last few thousand years. Patterns of runs of homozygosity (RoH) in a genome can inform on mating patterns in more recent history. A RoH reflects a genomic segment inherited from the same ancestor, and its length the number of generations that have passed since that ancestor lived. Recent small effective population sizes and inbreeding will therefore lead to an increase in the number of long RoHs in a genome. Inference of RoH in the genotype array dataset reveals levels that are generally higher than those of major populations in East Asia and Europe (Figure 3.8A), likely reflecting generally lower recent effective population sizes. There are also substantial differences between different groups within PNG (Figure 3.8B). The highest levels are found among the Angan-speaking groups from the south-eastern highlands. These are the highland groups that display the highest F_{ST} values to other highland groups (4-5%). The western highlands cluster has slightly lower amounts than the eastern highlands cluster. There is no overall difference between highlanders and lowlanders in their amount of long RoH (Wilcoxon signed rank test, p = 0.227).

As the rate at which F_{ST} increases is dependent on effective population sizes, the small sizes of PNG groups likely contribute to the high differentiation observed. F_{ST} to an outgroup can be taken as a proxy for the amount of genetic drift experienced by a given group. While neither the negative correlation between the estimated number of speakers in a language group and the F_{ST} to outgroup (Figure 3.8B), nor that between the number of speakers and the amount of RoH (Figure 3.8C), is significant, a handful of groups stand out in having small numbers of speakers, high F_{ST} to an outgroup and large amounts of RoH. These include the Angan speakers mentioned above as well as the Grass Koairi and the Keapara speakers, both from Central Province. These likely represent examples of recently culturally isolated groups that have experienced high levels of genetic drift in relatively short periods of evolutionary time. There is a significant positive correlation between the amount of RoH in a group and the F_{ST} to an outgroup (Figure 3.8D), consistent with small effective population sizes driving the genetic drift and high differentiation in PNG.

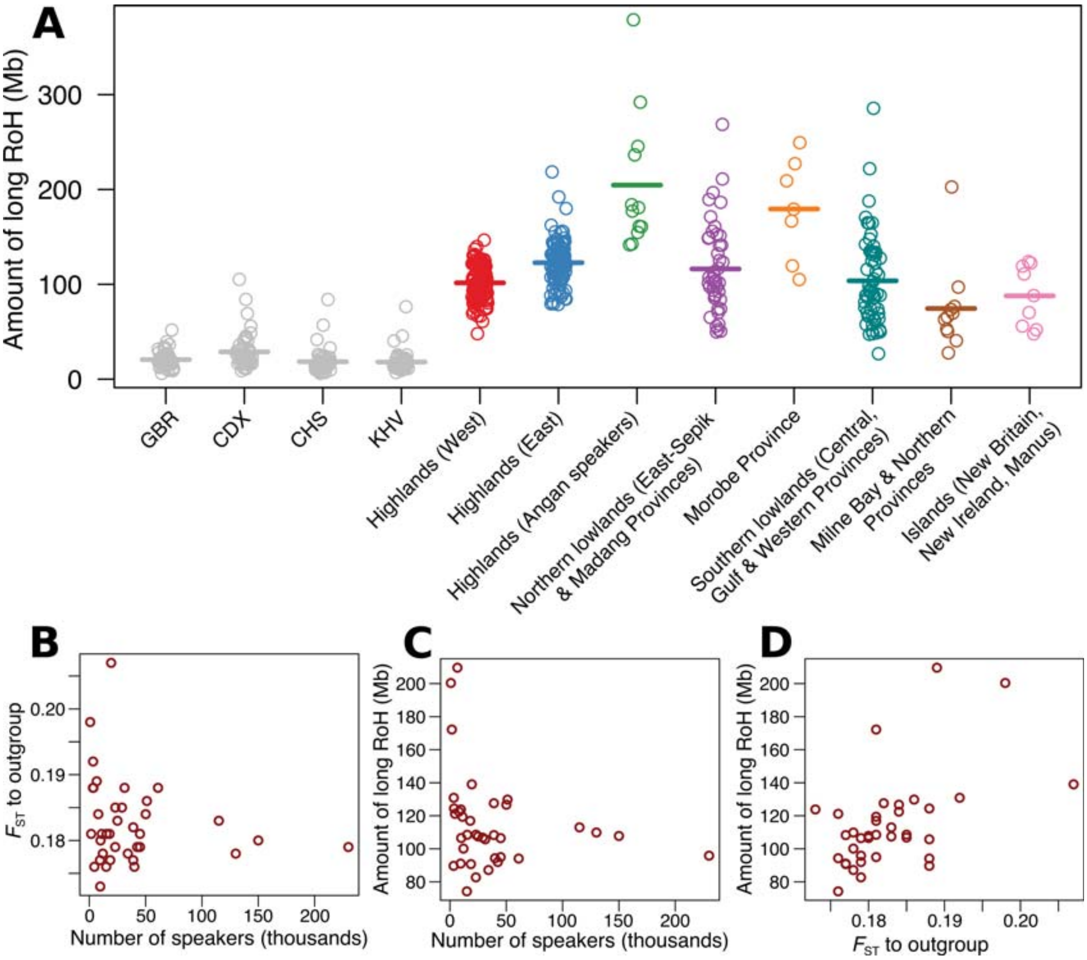

Figure 3.8: Patterns of runs of homozygosity and genetic drift. *A) The sum of lengths of RoH longer than 1Mb in each individual genome, grouped by major region. Horizontal lines indicate the mean per region. Four populations from the 1000 Genomes project are included for comparison: GBR = British in England and Scotland; CDX = Chinese Dai in Xishuangbanna, China; CHS = Southern Han Chinese; KHV = Kinh in Ho Chi Minh City, Vietnam. B) The estimated number of speakers in a group plotted against* F_{ST} *to an outgroup (KHV). Pearson correlation = -0.176, p = 0.2967. C) The estimated number of speakers in a group plotted against the group mean sum of lengths of RoH longer than 1 Mb. Pearson correlation = -0.233, p = 0.1656. D)* F_{ST} *to an outgroup (KHV) plotted against the group mean sum of lengths of RoH longer than 1 Mb. Pearson correlation = 0.520, p = 0.0009662. In B-D, only groups with at least two individuals are included. Estimated speaker numbers for each language group were obtained from Ethnologue (Lewis, et al. 2016).* F_{ST} *to KHV was calculated after locally masking lowlander genomes for Southeast Asian ancestry, and restricted to groups with at least two sampled individuals. Results are very similar if using Aboriginal Australians as outgroup instead of KHV.*

3.12 A model for population history in Holocene PNG

Rather than supporting a naïve view where population structure was established shortly after the peopling of New Guinea and maintained since then, the results obtained here indicate that more recent processes have overwritten earlier structure. Ultimately, ancient DNA will likely be needed to determine with more certainty what exactly these processes were, but at present one can at least speculate. One model which would be compatible with present-day patterns of genetic variation is an expansion of a single agriculturalist group in the highlands, starting approximately 10 kya. If all highlanders derive most of their ancestry from this same source population, this could explain their clade-like relationship to lowlanders. It is also compatible with the split time estimates between highlanders and lowlanders and those within the highlands, and the observation of major increases in highlander effective population sizes within the same time period. This agriculturalist expansion could also have spread the languages that have now developed into the Trans-New Guinea language family.

If this model is correct, then it means that, similarly to west Eurasia and sub-Saharan Africa, and likely East Asia too, agriculture spread through the movement of people in PNG. Its independent history would thereby provide another data point in favour of people in the “pots vs people” debate. At present, a model where agriculture spread without genetic admixture or replacement cannot be completely ruled out – perhaps the current divergences were established while groups where still hunter-gatherers, reflecting either some pre-agricultural population transformation (the likes of which have been described in Europe (Fu, et al. 2016)) or some kind of upper limit on the divergence achievable between hunter-gatherer groups living in a fairly small geographical area. However, the close correspondence between the genetic time estimates and the archaeological evidence for the emergence of agriculture arguably lends support to an agriculturally-driven restructuring.

A striking result which sets PNG apart from other regions of the world that have also undergone agricultural transitions is the very strong present-day genetic differentiation. The high F_{ST} values across all of PNG differ dramatically from the relatively homogenous genetic landscapes of Europe, East Asia, and most populations of sub-Saharan Africa. Areas of comparable size to PNG in Europe have been found to have close to no population differentiation, e.g. Great Britain (Leslie, et al. 2015), the Netherlands (Genome of the Netherlands 2014) and Denmark (Athanasiadis, et al. 2016). An important insight into this comes from ancient DNA studies in Europe and the Near East, which have documented a dramatic but gradual decline in F_{ST} values in this part of the world over the last 10 ky (Lazaridis, et al. 2016), as well as higher differentiation between hunter-gatherer groups than between farmer groups (Skoglund, et al. 2014). Before the agricultural transition in west Eurasia, some hunter-gatherer groups were as different as present-day Europeans and East Asians (F_{ST} of $\sim$10%). These studies show that the homogenous landscape of present-day Europe actually only emerged in the last few thousand years. While an agricultural transition might be necessary to achieve this high level of genetic homogenization, the history of PNG demonstrates that it is not sufficient: PNG also went through an agricultural transition, but present-day genetic differentiation is high.

This thus calls for another explanation for the dramatic differences in genetic structure between PNG and west Eurasia or other genetically homogenous regions. A hypothesis is that the key difference is the absence in PNG of a Bronze age, Iron age or other similar post-agricultural, cultural transformation. In west Eurasia, the Bronze Age started $\sim$4.5 kya and was driven by an expansion of herders from the Pontic–Caspian steppe who had domesticated horses, metal technology and perhaps a different social structure. This expansion resulted in dramatic genetic admixture and replacement across all of Europe, as well as replacement of the vast majority of the indigenous languages with Indo-European languages (Allentoft, et al. 2015; Haak, et al. 2015). The presence of Indo-European languages, and steppe-related ancestry (Lazaridis, et al. 2016), in South Asia suggest that a similar expansion shaped genetic structure here too. In sub-Saharan Africa, the expansion of Bantu-speaking farmers from west Africa was associated with an Iron Age (Huffman 1982), and resulted in genetic homogenization across a large geographical area (Li, et al. 2014; Patin, et al. 2017). PNG might then be similar in genetic structure to pre-Bronze Age Europe, with small, sedentary, agriculturalist groups and relatively little gene flow between them. The genetic results also align with, and provide some insight into, the enormous linguistic diversity of PNG, and perhaps suggest that many other parts of the world also harboured greater linguistic diversity in the past, before being homogenized by large-scale cultural expansions.

These expansions have often been associated with the rapid proliferation of particular Y chromosome lineages, e.g. R1b with the steppe migration into Europe, E1b with the Bantu expansion in sub-Saharan Africa, and likely R1a in South Asia (Karmin, et al. 2015; Poznik, et al. 2016), giving rise to 'star-like' phylogenies as the rapid spread leaves little time for new mutations to accumulate on the basal branches. The explanation for this is likely a social structure with higher variance in male reproductive success, where whatever Y chromosomes are carried by a small number of high-status males expand dramatically. An extreme case of this from more recent times might be the Y chromosome lineage thought to be carried by the central Asian ruler Genghis Khan, today present in an estimated 8% of East Asian males (Zerjal, et al. 2003). Consistent with the genetic landscape of PNG being unperturbed by large-scale, post-agricultural expansions, the phylogeny of PNG Y chromosomes does not feature any star-like bursts of expansion (not shown). One striking case of Y chromosome structure can be noted, however: the Gende highland group carry a Y chromosome from haplogroup C at a frequency of 86%, but this chromosome is absent from the rest of the highland samples examined. Analysis of whole-genome sequencing data from five randomly selected Gende males indicates a common ancestor for their Y chromosomes only $\sim$1 kya. It thus seems likely that this chromosome has risen in frequency due to a recent, small-scale instance of high male reproductive variance in the Gende speakers. However, consistent with the high degree of population differentiation in PNG, this Y chromosome expansion appears to have been contained to only this single language group, without spreading more widely.

The proposed model of an agriculturalist expansion in the highlands is simple, and therefore likely to be inaccurate in the finer details. There are also aspects that it fails to explain, including the fact that lowland groups also practice agriculture, and in many areas also speak languages from the Trans-New Guinea language family. If this was the result of the same expansion of people spreading down from the highlands, more recent divergences and greater genetic continuity between highland and lowland groups would have been expected. More work is thus needed to further the test the basic outline of the model and refine it with additional detail.

3.13 Conclusions

The people living in the highlands of PNG appear to have been unaffected by Southeast Asian gene flow, thereby making them as genetically independent from Eurasian sources as Aboriginal

Australians, i.e. for ~50 ky. While the notion that agriculture was an independent development in the PNG highlands largely serves as a background assumption for genetic studies, derived from archaeology, these genetic findings themselves also provide evidence for this notion.

The first detailed view of population structure in PNG, made possible by the comprehensive array genotyping dataset in combination with whole-genome sequences, hints at a complex population history. The age of present-day population structure in the highlands is not older than the development of agriculture, suggesting a reshaping following this lifestyle transition. The highlands-lowlands contrast constitutes a major barrier to gene flow, and all people across the highlands appear to form a clade relative to lowlanders. Population differentiation is high, despite divergences not being particularly ancient, instead implying that relatively recent isolation between groups is responsible. Population genetics theory also demonstrates that F_{ST} will increase more rapidly in groups with smaller effective population size. A parallel can perhaps be drawn between PNG and some communities in South Asia, where founder effects and cultural isolation between groups even living in the same geographical areas has led to strong genetic differentiation (Reich, et al. 2009; Nakatsuka, et al. 2017). A hypothesis for why PNG has such strong population structure while regions such as Europe, East Asia and parts of sub-Saharan Africa do not, is that the latter regions have been recently homogenized by large-scale, technology-driven expansions such as Bronze and Iron ages. The history of PNG therefore demonstrates that human population histories are not deterministic, but can take quite different trajectories in different parts of the world.

There is currently no consensus on where within the highlands agriculture was first developed. The population-genetic analyses presented here provide little if any insight into the matter – relationships between present-day groups tell us very little about the geographical origins of their ancestors. However, the results can at least be superficially compared to predictions implied by a hypothesis put forth on anthropological and archaeological grounds, and which states that agriculture likely started in the western parts of the PNG highlands and then spread from there to the eastern parts (Feil 1987). The basis for this hypothesis is several observations suggesting higher present-day agricultural productivity in the western parts, as well as larger group sizes. The genetic results give some support for this. Firstly, the levels of runs of homozygosity are slightly lower in the west (Figure 3.8A), suggesting slightly larger population sizes. Secondly, although this might not be statistically significant given the resolution offered by the methods, inferred historical effective population sizes in the last few thousand years are also slightly higher in the western groups (Figure 3.6C).

It is worth noting that, despite the strong differentiation and low gene flow between groups, Papua New Guinea has undergone one quite substantial cultural transition in the last 500 years or so. Before this, the prominent agricultural crops in PNG were taro and yams; however, today the non-indigenous sweet potato dominates (Bourke, et al. 2009). It is not entirely clear how the sweet potato, which originated in the Americas, reached New Guinea: whether it was brought by

European explorers or by Polynesian seafarers shortly prior to the era of European colonialism. Genetic analyses of sweet potato samples seem to favour the latter hypothesis (Roullier, et al. 2013). In any case, the use of the sweet potato seems to have spread rapidly across New Guinea without signs of any substantial genetic restructuring, showing that such cultural innovations can disseminate even through human societies that are highly genetically and linguistically fragmented.

3.14 Materials and Methods

Array genotyping and 10x Genomics linked-read sequencing was carried out by the Wellcome Trust Sanger Institute sequencing and genotyping facilities. Analyses of these data and the 39 whole genome sequences are described in further detail in (Bergström, et al. 2017). A brief description of the methods used for the analyses described here follows.

Array genotypes were filtered for markers with ambiguous chromosomal location, indel status, high rate (>5%) of missing genotypes, low minor allele frequency (<1%), and for individuals with high rate (>10%) of missing genotypes, to produce a final variant set consisting of 529,137 variants and 378 individuals. These genotypes were combined with those from previously whole-genome sequenced PNG, Aboriginal Australian (Prufer, et al. 2014) and worldwide populations (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 multi-allelic after merging. Data from the Simons Genome Diversity Project (Mallick, et al. 2016) was downloaded and analysed in isolation.

Uniparental genotypes were handled separately to assign each individual to haplogroups. Array genotypes were merged with whole-genome sequencing data using Y chromosome genotypes previously called, and mitochondrial genotypes called using FreeBayes v0.9.18 (Garrison and Marth 2012) with the arguments "--ploidy 1" and "--report-monomorphic". For the Y chromosome, the patterns of genotypes across individuals were hierarchically clustered and variants defining haplogroups manually identified and checked against the ISOGG database (http://isogg.org/tree/, April 21 2016 release). Individuals were assigned to the C, M, S, O and K haplogroups. For the mitochondrial genome, the same approach was applied, using information in the PhyloTree database (build 17) (van Oven and Kayser 2009), alongside an independent, higher-resolution haplogroup prediction by the HaploGrep 2 software (Weissensteiner, et al. 2016). The two approaches agreed on the (lower-resolution) haplogroup in all cases. Individuals were assigned to the P, M, E, B4, B5b and N13 haplogroups. A uniparental F_{ST} metric was calculated as (H_T – H_S) / H_T, where H is haplotype diversity for the total (H_T) and the subpopulations (H_S), calculated as $1 - \sum p^2$, where p is allele frequency.

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. f_3- and D-statistics,

f_4-ratios and admixture graph fits were calculated using ADMIXTOOLS 4.1 (Patterson, et al. 2012). F_{ST} between pairs of language groups was calculated using EIGENSTRAT 6.0.1 (Patterson, et al. 2006), restricted to groups that have at least two sampled individuals. Local ancestry assignment was performed using RFMix v.1.5.4 (Maples, et al. 2013), using PNG highlanders and 1000 Genomes East Asians as reference panels, after phasing haplotypes on the Sanger Imputation Server (McCarthy, et al. 2016) using the EAGLE algorithm (Loh, et al. 2016) and the Haplotype Reference Consortium (release 1.1) reference panel. Runs of homozygosity were inferred using PLINK v1.07 (Purcell, et al. 2007) with the arguments "--homozyg-window-kb 1000 --homozyg-window-snp 50 --homozyg-density 50 --homozyg-window-het 0".

Effective population size histories were inferred on five genomes per population using SMC++ (Terhorst, et al. 2017) and on four genomes per population using MSMC (Schiffels and Durbin 2014) (as well as MSMC2), in both cases using the mappability mask recommended for the latter. MSMC was also used to estimate the relative cross-coalescence rate between populations, applying the "—skipAmbiguous" argument to exclude unphased segments and the "—fixedRecombination" argument. To make use of 10x Genomics experimental phasing data, phase information from the LongRanger v2.1.2 VCFs was lifted into the existing VCFs, setting to unphased any position where the genotypes disagreed between the two VCFs. Analyses on the X chromosome were performed as described in Chapter 1. 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 30.5 years (Wood 1987).

To infer the genetic identify of all the 17 Papua New Guinean individuals in the HGDP-CEPH panel, array genotype data on these (Li, et al. 2008) was merged with the newly generated PNG array genotypes, and their relationships to the latter was inferred using PCA and outgroup f_3-statistics.

Chapter 4: Large-scale sequencing of worldwide human populations

4.1 Introduction

Access to data from many different and diverse human populations is clearly key to efforts to understand human genetic diversity and population history. A number of different datasets have been produced and served as useful resources for the human genetics research community. The HapMap project was an early effort that provided array genotypes from populations of mainly European, East Asian and African ancestry (International HapMap Consortium 2005). The Human Genome Diversity Project (HGDP) and the CEPH foundation established a collection of cell lines from a large number of diverse human populations (Cann, et al. 2002), and array genotyping of this collection has resulted in widely used datasets (Jakobsson, et al. 2008; Li, et al. 2008; Patterson, et al. 2012). The POPRES resource released genotype array data on close to 6,000 individuals (Nelson, et al. 2008), with particularly good geographical coverage of Europe; however, its wider usage has been limited as the data are available only under managed access rather than open access. The 1000 Genomes Project released data based on a combination of low-coverage whole-genome and high-coverage exome sequencing data from worldwide human populations, initially from seven populations (1000 Genomes Project Consortium 2010), then 14 populations (1000 Genomes Project Consortium 2012) and finally 26 populations (1000 Genomes Project Consortium 2015). The aggregation of data generated in many different studies, particularly large medical genetics studies, has led to the establishment of resources based on data from tens of thousands of individuals, such as the Exome Aggregation Consortium (ExAC) (Lek, et al. 2016) and the Haplotype Reference Consortium (McCarthy, et al. 2016). While the donor consents for the constituent studies typically preclude open access to the individual genotypes, these resources enable accurate assessments of allele frequencies and/or high-quality genotype phasing and imputation. Recently, projects with sampling strategies focused explicitly on population history have provided high-coverage whole-genome sequencing data from very large numbers of diverse populations, but limited to typically 2-4 individuals from each of them; the Simons Genome Diversity Project (SGDP) with 300 individuals from 142 populations (Mallick, et al. 2016), and the Estonian Biocentre Human Genome Diversity Panel (EGDP) with 483 individuals from 148 populations (Pagani, et al. 2016).

A number of different factors influence the utility of a given dataset or resource to different areas of study within human genetics. The primary utility of population-genetic datasets to medical genetics studies is providing information on the frequency of particular alleles in populations of broadly-defined ancestry and in constituting a resource for phasing and imputation, such that the sample size of the dataset is key. For population history studies, the diversity of the sampled population is very important, as different populations might provide different information into history.

The HGDP-CEPH collection, mentioned above, has several attractive features as a resource for human population genetics. It contains individuals from about 52 different populations, with typically around 20 individuals from each of these. While projects such as HapMap and the 1000 Genomes Project sampled individuals mainly from major continental populations (sampling criteria included being non-vulnerable and relevant to medical-genetic studies), the HGDP sampling encompasses many smaller populations of particular anthropological, linguistic, historical or genetic interest. As a few examples, from Africa, it contains the Khoe-San, believed to represent the earliest branching modern human population, and central African rain forest hunter-gatherers. From Europe, it contains the Basque population, one of the few European groups to speak a language that is not part of the Indo-European family, and the isolated Orkney islanders. From the Middle East and South Asia, it contains the Druze ethno-religious group and the isolated Kalash group from the western Himalayas. From East Asia, it contains several minority ethnic groups from China, including the Turkic speaking Uyghurs from western China. From Oceania, it importantly contains Papua New Guineans. From the Americas, it contains several Native American groups without European admixture, which otherwise is ubiquitous in the majority populations. As the resource consists of cell lines, an unlimited amount of DNA can be obtained. Additionally, while data obtained for population history studies sometimes has restrictions on how it can be used or shared, the policies of the HGDP-CEPH resource is such that all data can be analysed without restrictions and distributed openly. To date, array genotype and other data generated from this collection have been used in hundreds of population genetics studies.

Whole-genome, high-coverage sequencing of most of this panel was undertaken at the Wellcome Trust Sanger Institute. This project is still ongoing at the time of writing, but in this chapter I describe some initial work on this sequencing data, focusing on the more technical aspects rather than population genetics analyses.

4.2 Sequencing of the HGDP-CEPH panel

The full HGDP-CEPH panel contains over 1000 samples; however, some of these are close relatives, duplicates, have ancestry deviations or other complications, such that a core set of 952 unrelated samples has been established (Rosenberg 2006). Out of these, approximately 135 had been sequenced to high-coverage (here meaning at least approximately 30x) as part of another project, the Simons Genome Diversity Project (Mallick, et al. 2016). These were sequenced mostly as PCR-free libraries on the Illumina HiSeq2000 machines with paired-end 100bp reads. In addition, 10 samples had been sequenced in an earlier study (Meyer, et al. 2012). The remaining samples were sequenced at the Wellcome Trust Sanger Institute (WTSI). An initial batch of 178 samples was sequenced as PCR-based libraries (PCR-free libraries were not available for large-scale production sequencing at WTSI at the time). The remaining approximately 650 samples were sequenced as PCR-free libraries. These were all sequenced on the Illumina HiSeqX machines with paired-end 150 bp reads. A number of samples were deliberately sequenced more than once across technologies (SGDP versus WTSI versus Meyer, PCR libraries versus PCR-free libraries), to allow for assessments of reproducibility and batch effects. All reads were mapped to the GRCh38 reference genome.

4.3 Characterization of cell line chromosomal abnormalities

The HGDP-CEPH DNA samples derive from lymphoblastoid cell lines (LCLs) collected and established by a number of different laboratories (Cann, et al. 2002). Culturing of cell lines always comes with some risk of new mutations being introduced. In particular, large-scale chromosomal changes, e.g. loss, gain or rearrangements of large segments or entire chromosomes can occur, often without impact on the viability of the affected cells (Shirley, et al. 2012). As the lines are maintained as populations of cells, a de-novo variant might be present in just some subset of the cells in this population, or in all of them. When sequencing DNA from cell lines, these abnormalities could interfere with the accurate determination of the germ-line genome sequence of the donor.

To determine the extent of chromosomal abnormalities in the HGDP-CEPH cell lines, I analysed patterns in the depth of sequencing reads mapped against the reference genome. The excess, or depletion, of reads relative to the genome-wide average should be proportional to the frequency of the chromosomal abnormality variant in the cell population, reaching up to 1.5 in the case of gains or down to 0.5 in the case of losses, if all cells carry the change. Measuring the total number of reads mapping to a chromosome overall will likely be enough to identify high-frequency, whole-chromosome events, but to allow for lower-frequency and partial events I plotted coverage along the length of each chromosome and manually inspected these for deviations (Figure 4.1). This identified approximately 50 events across samples, though the majority constituted only very slight deviations from the genome-wide average. Most events affect whole chromosomes or large segments, e.g. half, of chromosomes, but there are also smaller events on the scales of tens of megabases. Chromosomes 9 and 12 were subject to a larger number of whole-chromosome gains than other chromosomes. There appeared to be no correlation between ancestry of the donor and the rate of chromosomal abnormalities.

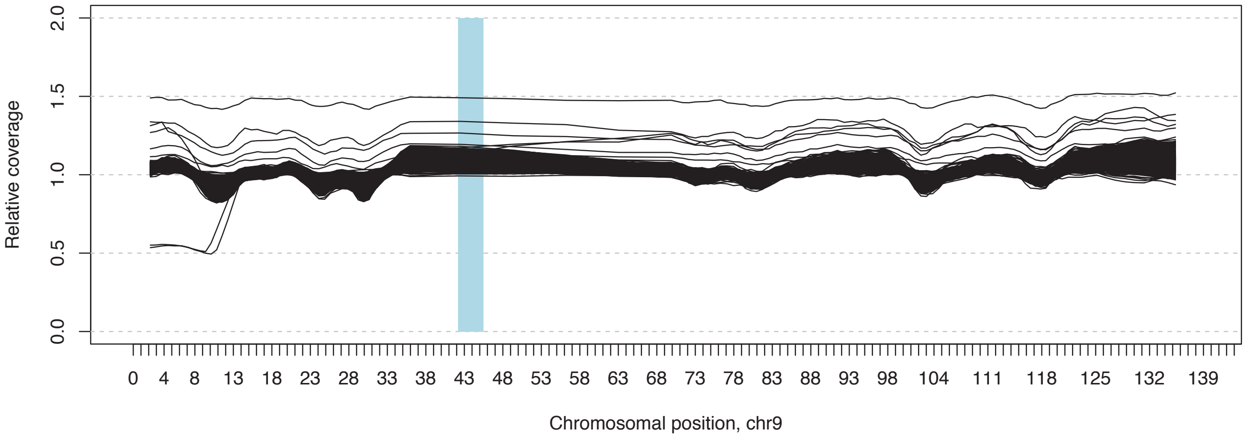

Figure 4.1: Identification of chromosomal abnormalities in the HGDP-CEPH cell lines from high-coverage whole-genome sequencing data: an example from chromosome 9. *Each black line represents a single sequenced individual, displaying the average coverage relative to the genome-wide median in a rolling window across the chromosome. The blue rectangle indicates the location of the centromere. Several samples containing copy-number gains of the whole chromosome, in varying proportions of the cells in the sequenced cell population, are visible. Additionally, two samples containing a ∼10 Mb deletion at the beginning of the chromosome, seemingly carried by close to all cells in the populations, are visible.*

Chromosomes X and Y have sex-dependent expectations for the sequencing coverage relative to that of the autosomes. Furthermore, in contrast to the autosomes where any larger-scale chromosomal abnormalities in-vivo are invariable associated with developmental disorders, copy-number deviations on the X and Y often do not have major phenotypic effects (other than affecting fertility). A large population sample will therefore contain healthy individuals with such deviations with a non-trivial probability; for example, a XXY configuration (diagnosed as Klinefelter syndrome) occurs at a rate of approximately 1 in 1000 males (and one example was detected in the 1000 Genomes Project (1000 Genomes Project Consortium 2015)). Analysing coverage on the X and Y chromosomes in the HGDP samples revealed a number of individuals with deviating copy numbers (Figure 4.2). Two males display substantially reduced coverage on the Y chromosome, likely reflecting loss of Y in the cell line, but potentially also reflecting actual loss of Y in the healthy donor, a phenomenon known to occur in certain tissues, particularly in older men (Forsberg 2017). 10 or so females display varying degrees of loss of X, likely on the cell line level. One male displays coverage consistent with a XXY configuration, thereby likely representing a naturally occurring example of this karyotype.

In addition to these copy number variations, one sample (HGDP01097, from the Chinese Tujia population) was found to be completely homozygous across the entire length of chromosome 1 (confirmed using array genotype data from this sample (Li, et al. 2008)), but the copy number of the chromosome is normal. This is likely a cell line artefact; however, the phenomenon does occur at low frequencies in-vivo owing to a process known as uniparental disomy. While this is typically associated with genetic disorders due to increased risk for a homozygous state at recessive disease variants (King, et al. 2014), it has been described in healthy individuals as well (Field, et al. 1998).

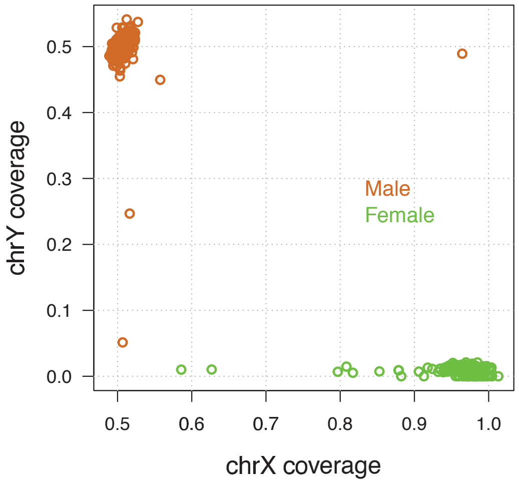

Figure 4.2: Sex chromosome sequencing coverage in the HGDP samples. *Depth of coverage on the X and the Y chromosomes relative to the genome-wide median are plotted against each other for each individual sample. In males, the X and Y chromosomes are haploid, and coverage is therefore expected to be half of the genome-wide median.*

4.4 Effects of chromosomal abnormalities on sequencing and genotyping

The HGDP samples have been genotyped on arrays in the past (Jakobsson, et al. 2008; Li, et al. 2008; Patterson, et al. 2012), data that have been used in hundreds of studies. An unequal copy number of the two chromosomes could conceivably interfere with accurate array genotyping, e.g. leading to some heterozygous genotypes incorrectly being called as homozygous for the allele carried by the chromosome in higher abundance. The genotyping of homozygous sites should not be affected. I investigated the effects in the most heavily used dataset, generated with the Illumina 650K array (Li, et al. 2008). Coverage per chromosome in the whole-genome sequencing data for a given sample is associated with reduced heterozygosity in the array genotypes (Figure 4.3). This thus shows that these cell line abnormalities have affected the array genotypes. It also shows that the abnormalities themselves have been present in the cell lines for a long time, and have not arisen recently or just locally in a subset of cells that was used to extract DNA for this particular whole-genome sequencing project.

I next examined the effects that chromosomal copy number abnormalities might have on the calling of variants from whole-genome sequencing data. While changes affecting large chromosomal regions would have obvious confounding effects on the identification of germ line copy number variation in the donors, it’s less clear how it might affect the identification and genotyping of small nucleotide variants. To address this, I performed a down-sampling experiment using the necessarily haploid X chromosomes in male samples. By pooling reads from two different male X chromosomes at different relative abundance and calling variants across these reads, treating them as coming from a single sample, the sensitivity in calling heterozygote variants can be assessed. The calling of homozygous variants should not be affected by changes in chromosomal copy numbers. The results show that as the copy number of the second chromosome decreases, the ability

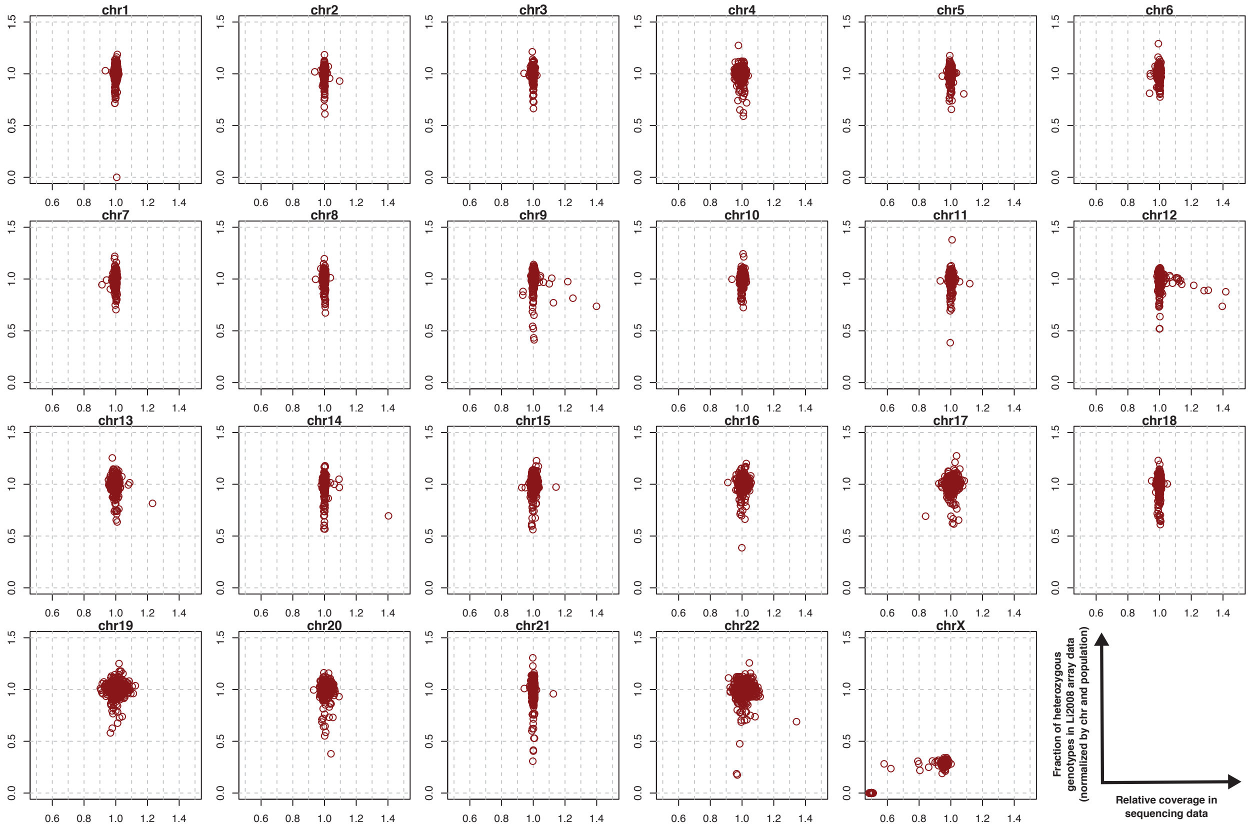

Figure 4.3: The effects of cell line chromosomal copy number abnormalities on array genotyping in the HGDP panel. *For each chromosome separately, sequencing coverage in the whole-genome sequencing data (normalized by the genome-wide median) is plotted against the fraction of heterozygous genotypes in the array data (normalized by chromosome and by population) for each individual sample. Deviations from a balanced chromosomal copy number is associated with reductions in heterozygosity; see for example chromosomes 9 and 12.*

to call heterozygote genotypes decreases quite rapidly (Figure 4.4). However, the effect is not as strong with increases in the copy number of the second chromosome. This is expected, as a variant caller is more likely to incorrectly call a heterozygous site as homozygous if the read count for the less frequent allele is low. Also as expected, the negative effects were more pronounced with lower levels of overall coverage.

In summary, the cell line artefacts affect genotypes called both from genotype arrays and from whole-genome sequencing data. The latter are affected more by chromosomal losses than by gains. On the basis of this, we decided to exclude nine samples that carried large, high frequency losses (as well as one sample that had gains on four separate chromosomes, and the sample with a homozygous chromosome 1) from most downstream analyses.

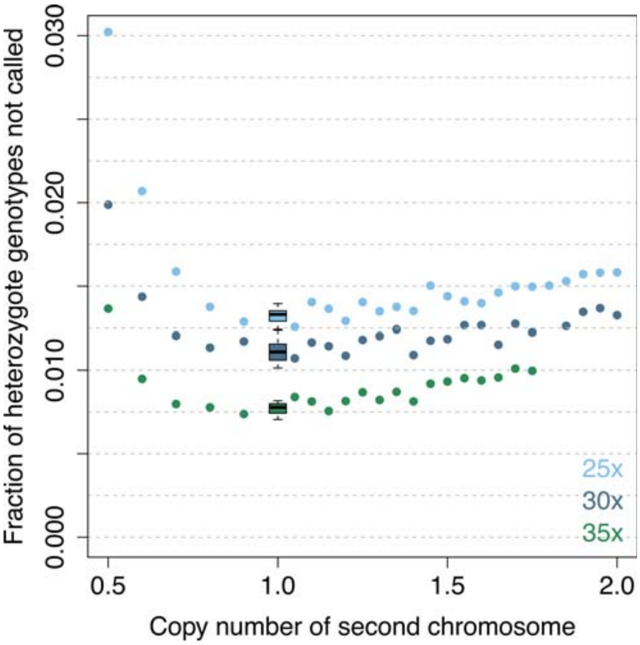

Figure 4.4: The effects of chromosomal copy number abnormalities on the ability to call heterozygote genotypes in high-coverage whole-genome sequencing data. *Data from the X chromosome of two male samples (both from the South Asian Pathan population) were pooled together in different relative abundance, and variants were called across these reads, treated as a single sample, using GATK HaplotypeCaller + GenotypeGVCFs. The horizontal axis shows the copy number of the second chromosome relative to the first chromosome. The vertical axis shows the fraction of heterozygote genotypes that were not called, relative to what was called with a balanced read count from the two chromosomes. The experiment was performed at three different levels of background genome-wide coverage; 25x, 30x and 35x. Twenty replicate down-sampling experiments were performed for the case with a balanced read count, the results of which are displayed with a boxplot.*

4.5 Variant call properties and filtering strategies

There is a myriad of strategies for the calling and filtering of genotypes from whole-genome sequencing data, and strategies also depend on the objective of the given study. I performed various analyses on a preliminary callset of the HGDP samples, produced at the Wellcome Trust Sanger Institute using the GATK HaplotypeCaller software (McKenna, et al. 2010), to try to understand the properties of the variant calls, discover technical issues that need to be addressed and identify an appropriate filtering strategy for this data.

As the sequencing data for this project were produced by two different institutes using a combination of PCR-based and PCR-free libraries, as well as different sequencing platforms, some technical batch effects might be expected, and these might affect downstream ancestry analyses. PCA analyses reveal that there are indeed batch effects, with the strongest one being between WTSI and SGDP datasets, but also some between PCR-based and PCR-free libraries (Figure 4.5). This is also seen in direct *D*-statistic tests of allele frequency correlations between single individuals, for which there is a very strong expectation of a value of zero; for example: $D(\text{Chimp},\text{Biaka}_{\text{SGDP.PCRfree}};\text{Papuan}_{\text{WTSI.PCRfree}},\text{Papuan}_{\text{SGDP.PCRfree}}) = 0.0194$, $Z = 3.771$. The extent of the batch effect revealed by PCA is reduced by applying increasingly stringent filter thresholds as determined by the GATK VQSR (Variant Quality Score Recalibration) filtering engine, as well as by restricting to regions of the genome with good mapping properties for short reads (Figure 4.5). With the 1000 Genomes “strict mask”, which covers ∼78% of the reference genome, the batch effect is not discernible. The same is true in D-statistic tests, e.g. the above test statistic reduces to $D = 0.0064$, $Z = 1.334$ when applying the strict mask (though it cannot be ruled out that there is still a subtle but non-significant effect remaining). These results imply that the batch effect is not primarily driven by genotype differences at genuine and non-problematic sites, but rather by sites in difficult parts of the genome.

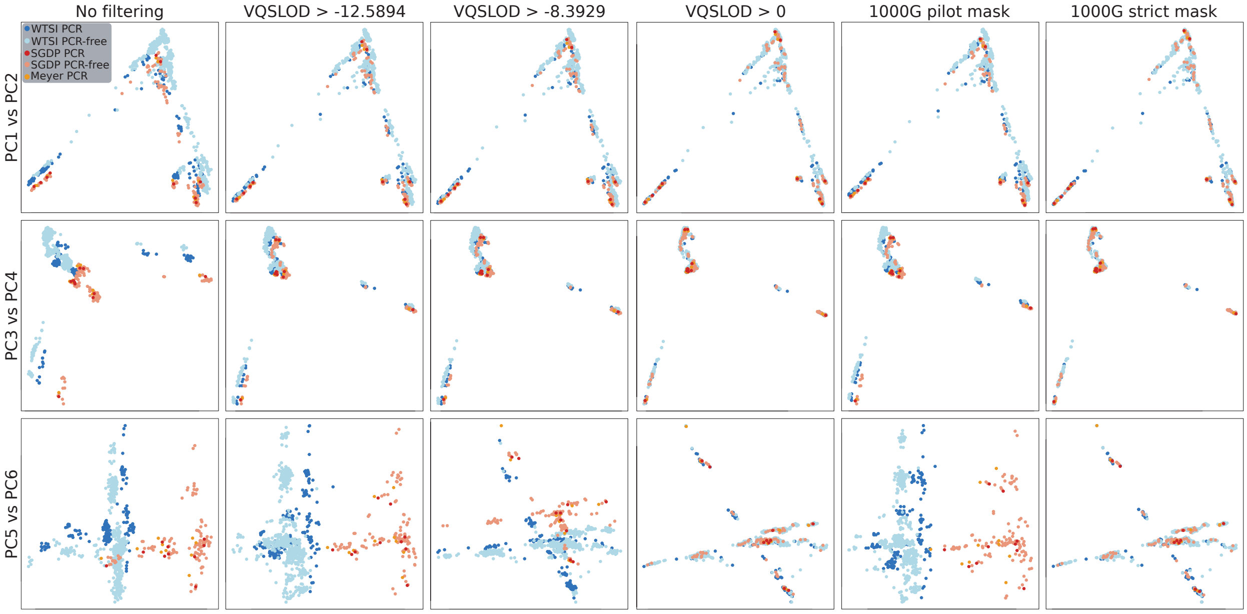

Figure 4.5: Technical batch effects in HGDP genotype calls and the effects of different filters. *Principal component analyses were performed on whole-genome SNP genotype calls, restricted to approximately 1 million variants with a minor allele frequency of at least 5% and pruned for LD. Without filtering (the left-most panel), a clear batch effect is visible between the sets of libraries sequenced with different technologies. This effect is reduced by increasingly stringent VQSR filtering thresholds, as well as by restricting to parts of the genome covered by mappability masks from the 1000 Genomes Project. With more stringent filters (i.e. VSQLOD > 0 and 1000G strict mask), in terms of ancestry, PC1 separates Africans from non-Africans, PC2 separates western non-Africans from eastern non-Africans, PC3 separates out Oceanians, PC4 separates out Native Americans, PC5 separates South Asians from Europeans and Middle Easterners and PC6 separates different African populations. With less stringent filtering, PC5 corresponds to the WTSI vs SGDP batch effect while PC6 separates South Asians from Europeans and Middle Easterners.*

Properties of the subset of SNPs that strongly differentiate between the sample sets might provide some insights into the causes of the batch effect. SNPs that are associated with higher alternative allele frequencies in the WTSI libraries also tend to have higher coverage and lower rates of missing genotypes across individuals in these compared to SGDP libraries (Figure 4.6). Likewise, SNPs that are associated with higher alternative allele frequencies in SGDP libraries tend to have higher coverage in these compared to WTSI libraries (though no clear difference in rates of missing genotypes). The latter SNPs have slightly lower GC content in the 100bp windows surrounding them, but this is not a large difference. Overall, these trends suggest that there are sets of sites in the genome which are differentially accessible to reads from these two sources, in each case leading to reduced coverage and ability to call alternative alleles. One underlying reason for this might be that the WTSI reads are 150 bp long, while the SGDP reads are only 100 bp, likely leading to some difference in mappability in repetitive parts of the genome.

The patterns of missing genotypes across individuals are non-random. There are some differences in the overall rate of missing genotypes across the called variants, with slightly lower rates in the

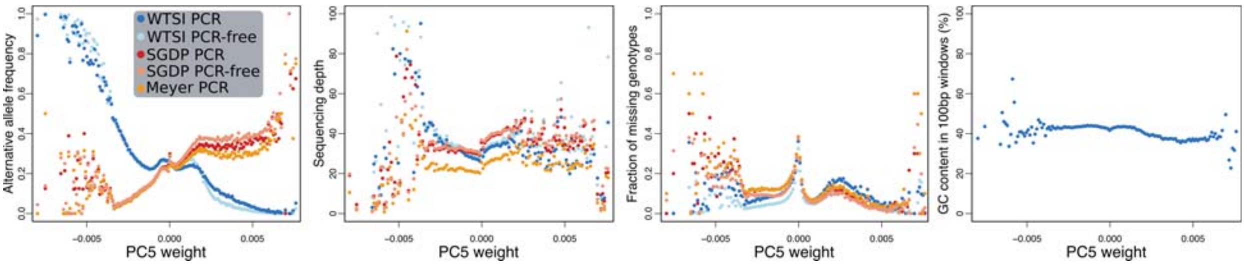

Figure 4.6: ***Properties of SNPs that separate between WTSI and SGDP sequenced libraries.*** *In PCA analyses of unfiltered SNP calls, PC5 separates WTSI and SGDP sequenced libraries (see Figure 4.5). The weights of each SNP for this component were extracted and plotted against various features of these SNPs.*

WTSI PCR-free than in other libraries (Figure 4.7A). Principal component analyses of the patterns of missing genotypes, rather than the genotypes themselves, also reveal a very strong clustering by technology rather than by ancestry, for both SNPs (Figure 4.7B) and indels (Figure 4.7C). These systematic differences in the ability to call genotypes might underlie some of the ancestry batch effects. Further to this, patterns of depth of coverage are also non-random (Figure 4.6), and the variant caller will output genotypes even at sites where there are just a few reads or an excessively high number of reads, resulting in low-confidence genotype calls that might be more susceptible to technical artefacts. Applying a site-level filter on the depth of sequencing coverage, setting any genotype to missing if the number of reads covering the site is lower than one third or higher than double the genome-wide average for the particular library (hereafter referred to as "DP3rd2x"), leads to some reduction in the strength of the batch effect in PCA (not shown). Similarly, the effect is partly reduced in *D*-statistic tests, i.e. the above test statistic reduces to *D* = 0.0141, Z = 2.271.

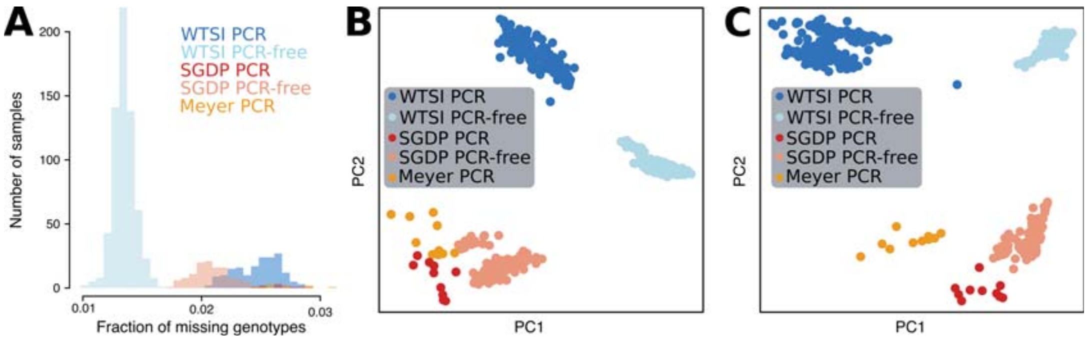

Figure 4.7: Missing genotypes in the HGDP callset*. A) Overall rates of missing genotypes at SNPS in the unfiltered callset for the different sets of libraries. B) A PCA on the patterns of missing genotypes at SNPs reveals very strong clustering by technology. C) A PCA on the patterns of missing genotypes at indels reveals very strong clustering by technology.*

The duplicate sequencing of a number of individuals with more than one technology allows for concordance assessments. Comparing WTSI PCR-free and SGDP PCR-free, representing the two dominant technology sets, there are an average of 263,779 discordant SNP genotypes between duplicate samples if no filtering is applied. This reduces to 133,225 at a VQSR LQSLOD threshold of >-12.5894, 103,325 at VQSLOD >-8.3929, 13,239 at VQSLOD >0, 159,837 with the DP3rd2x filter, 104,903 with the 1000 Genomes pilot mask, 6,669 with the 1000 Genomes strict mask and 6,280 with the combination of the DP3rd2x filter and the latter mask. If restricting to the ~650,000 sites that are present on the Illumina 650K array and therefore very likely to be real, and easily accessible, variants, there are an average of 117 discordant genotypes, which compares favourably to concordance observed between sequencing and array (which generally have very high quality) genotypes. Concordance at indel genotypes is substantially lower, considering that the total number of indel calls is approximately an order of magnitude lower than the number of SNPs calls, with e.g. an average of 197,929 discordant genotypes without filtering and 7,448 with the 1000 Genomes strict mask. Comparisons between duplicate samples in which at least one experiment was done with PCR-based libraries reveal substantially higher indel discordance (e.g. 28,855 discordant genotypes in the strict mask between a WTSI and a SGDP PCR duplicate sample), consistent with the notion that the PCR reaction introduces some level of indel error into the libraries. Overall, the observation that restricting to the strict mask leads to substantial reductions in duplicate discordance is consistent with the notion that most genotype errors are driven by mapping, alignment, coverage or other issues at problematic sites, rather than random sequencing errors or binomial sampling errors at well-behaved sites. Intuitively, the latter should be very infrequent for sequencing experiments with on the order of 30x coverage.

In summary, even with high-coverage sequencing, technology differences lead to considerable batch effects which impact ancestry analyses. However, the strength of these effects can be reduced by filtering. Application of strict VQSR filtering removes most of the batch effects, however this comes at a price of excluding a fairly large number of variants (e.g. at VQSLOD > 0, less than 80% of variants are retained, though many of the excluded ones will not be real variants). There is also a conceptual issue related to VQSR and similar filtering strategies, which is that they only apply to polymorphic sites. Certain population-genetic methods, however, need genotype information also at monomorphic sites, especially methods that in one way or another perform inference on the process of mutation, e.g. the PSMC (Li and Durbin 2011), MSMC (Schiffels and Durbin 2014) and SMC++ (Terhorst, et al. 2017) methods, as well as site-frequency spectrum modelling methods such as fastsimcoal (Excoffier and Foll 2011) (or more generally, any divergence calculation, e.g. the basic divergence between two genomes or the heterozygosity of a single genome). Making use of filters that only apply to polymorphic sites would introduce a bias against such sites, potentially confounding these methods. The above analyses, however, also show that by restricting to the regions of the genome that are easily accessible to short read mapping and non-repetitive, most if not all of the batch effects disappear. This kind of site-level, rather than variant-level filtering, would not introduce a bias against polymorphic sites and the resulting data would therefore be suitable for any kind of population-genetic analysis. It does, however, come at the cost of excluding large parts of the genome – while 78% of the reference genome (which is what the 1000 Genomes strict mask covers) is probably enough for most population-history and demographic analyses, an ideal genomics resource should provide data on as large a fraction of the genome as possible, to also maximally enable e.g. functional, medical and selection studies. There is thus, as always, a trade-off involved in these filtering decisions, and perhaps a sensible approach is to produce data for the whole genome and then restrict to suitable subsets depending on the particular type of analysis to be carried out.

4.6 Rare variant sharing patterns

One of the areas where sequencing of a large number of individuals per population has clear benefits, compared to array genotyping or sequencing of small numbers of individuals per population, is in the analysis of rare variants. There is an increasing interest in using rare variants to learn about population history (1000 Genomes Project Consortium 2012; Mathieson and McVean 2014; 1000 Genomes Project Consortium 2015; Field, et al. 2016; Schiffels, et al. 2016). As they typically result from recent mutations, their distribution across individuals might give insight into more recent population history than do common variants. An analysis of the sharing of doubletons, meaning variants that are observed exactly twice in this particular dataset, reveals an abundance of structure among the HGDP samples (Figure 4.8). This holds promise for the application of e.g. rare variant sharing asymmetry tests conceptually similar to the *D*-statistic, or more sophisticated, model-based approaches (Schiffels, et al. 2016).

On a general level, the dependence of the non-normalized doubleton counts on the background level of genetic diversity is evident in these results. Most strikingly, the number of variants shared between any African populations is greater than between many populations within non-African continental regions, even if the latter will typically be more closely related, reflecting the greater genetic diversity and therefore larger number of rare variant sharing opportunities in Africa. Most non-African populations share more doubletons with the San population than with other Africans, despite likely being less closely related to them, probably similarly reflecting the great genetic diversity of the San (Kim, et al. 2014). This demonstrates the need for appropriate normalization when using rare variant sharing counts to infer shared ancestry.

On a more detailed level, several known features of the history of particular populations are visible in the patterns of doubleton sharing. The South Asian Hazara population displays elevated sharing with East Asians, reflecting East Asian admixture described in this group. Specific South Asian and Middle Eastern individuals display elevated sharing with Africans, consistent with recent sub-Saharan admixture in these. The Uygurs from western China display elevated sharing with South Asians and Europeans, relative to other Chinese populations, reflecting known west Eurasian admixture in this group. The Melanesian population from Bougainville Island in Papua New Guinea display elevated sharing with East Asians (particularly the south-eastern groups of Cambodian and Dai) relative to the mainland Papuans, reflecting the $\sim$20% of their ancestry deriving from Southeast Asian admixture. The substructure within the mainland Papua New Guinean samples, described in Chapter 3, is clearly visible.

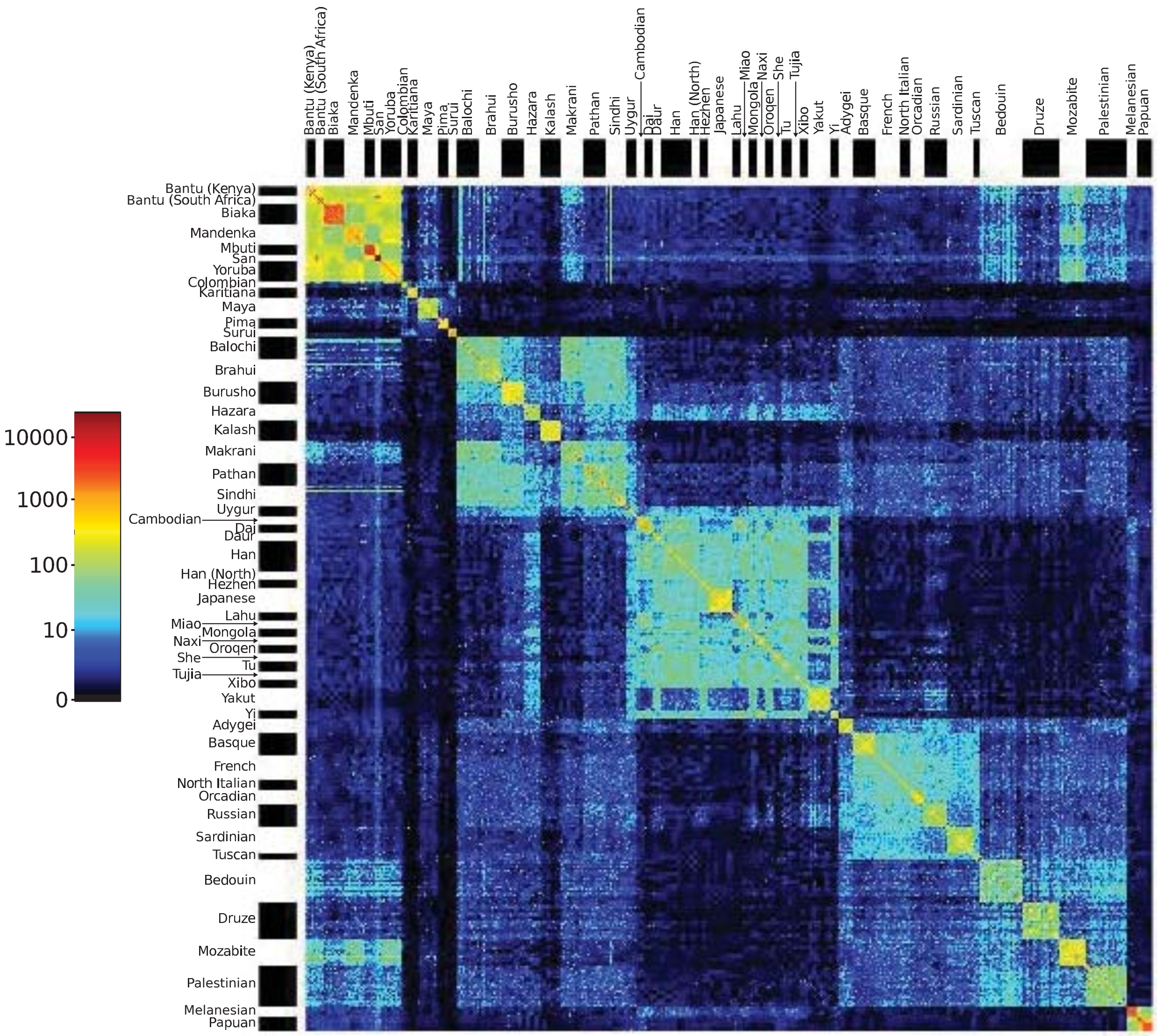

*Figure 4.8: **Doubleton variant sharing in the HGDP whole-genome sequencing data.** The non-normalized counts of shared variants observed exactly twice across the dataset, for each pair of individuals, displayed on a logarithmic colour scale. The boundaries between populations are indicated with alternating black boxes and empty space on each side of the plot.*

4.7 Conclusions

The high-coverage, whole-genome sequencing data generated from the HGDP-CEPH panel of worldwide populations is likely to be of great use in the study of human population history, and human genetics more generally. Compared to the 1000 Genomes Project (1000 Genomes Project Consortium 2015), which is the most commonly used panel representing worldwide human diversity, this resource will have several advantages. The main one is the more diverse populations represented in the HGDP panel, including many of particular historical interest and utility in analyses. Another one is the high sequencing coverage, enabling unproblematic application of methods that rely on high-quality genotype calls. Lower-coverage sequencing projects have typically relied on genotype imputation to achieve high-quality genotypes; however, there is concern about imputation in population history contexts as the ancestry composition of the panel used for imputation might introduce biases. A disadvantage of the HGDP collection is the relatively small number of individuals in some populations, e.g. 6-10 for the smallest ones, relative to ∼100 individuals for all populations in 1000 Genomes.

High-coverage sequencing currently provides the highest accuracy genotypes, but this project demonstrates that there are still technical issues that need consideration. The DNA for this sequencing project was extracted from cell lines and in some cases had chromosomal abnormalities, leading to a decision to exclude a small number of individuals. The chromosomal copy number abnormalities have affected array genotypes from the same panel, used for a decade across hundreds of studies; however, it appears that the negative effects are less pronounced for high-coverage sequencing data. Sequencing of individuals across different technologies, i.e. in different centres on different types of Illumina machines, and with PCR-based or PCR-free libraries, results in batch effects in the genotype calls with effects on ancestry analyses. These can be reduced through filtering; however, overly aggressive filtering by necessity means less of the genome is available for use in downstream analyses.

4.8 Materials and methods

Genotype array data from the HGDP-CEPH samples generated on the Illumina 650K array were obtained from (Li, et al. 2008) and chromosomal coordinates were lifted over from GRCh36 to GRCh38 using the NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap), as well as using dbSNP rs IDs. Genotype data on these samples generated on the Human Origins array were obtained from (Patterson, et al. 2012) and lifted over from GRCh37 to GRCh38 using the NCBI Genome Remapping Service. Genotypes for chimpanzee were extracted from the UCSC hg38-panTro4 axtNet alignments. Sequencing coverage along the reference genome was calculated for each sample using the "depth" command from the samtools software (Li, et al. 2009), applying the "-aa" argument. Linkage disequilibrium pruning and principal component analyses were performed using AKT (Arthur, et al. 2017). In order to perform PCA on the patterns of missing genotypes rather than the genotypes themselves, missing genotypes were recoded as 0 and non-missing as 1. The "pilot" and "strict" accessibility masks for GRCh38 were obtained from the 1000 Genomes Project. The rare variant sharing heatmap was made using the heatmap.2 function from the gplots R package.

Chapter 5: Future directions

5.1 Future directions for the population history of Sahul

The broad outlines of the genetic history of Sahul are starting to become clear, and that history appears to be relatively simple at the coarsest level of resolution, with genetic isolation from the rest of the world from the initial settlement $\sim$50 kya until the period of Southeast Asian contact and European colonization. However, much is still unknown about population history within Sahul, especially Australia, and a simple relationship to worldwide populations does not mean that the internal history was also simple. Indeed, using the well-studied history of Europe as an example, much of the complex Holocene population changes and admixture there occurred between populations that were all different branches of the broad western Eurasian ancestry. It is entirely possible that similarly dramatic processes have occurred within Sahul, involving differentiated branches of the broad Sahul ancestry, while not altering the relationships to worldwide outgroups. The first in-depth look at population structure in PNG, presented in chapter 3, hints at such complex processes, at least here. Sahul represents an opportunity to study a part of the world in isolation, providing independent insight into patterns of human evolution and the behaviour of populations, especially as many of the major cultural changes in Eurasia and Africa appear to be connected to some extent.

One of the major events that we do currently have knowledge of is the seemingly deep split time between Aboriginal Australians and Papuans. This establishes Sahul as a continent with a present-day population structure almost as old as that between western and eastern Eurasia. A better understanding of this split is therefore of great interest. Firstly, more confidence is needed in the actual split time estimates, with e.g. current MSMC estimates being associated with technical uncertainty surrounding haplotype phasing quality. Experimental phasing, e.g. with 10x Genomics as described in Chapter 3 here, might be a promising approach to resolve this. Secondly, more comprehensive sampling of populations in northern Australia, as well the Torres Strait Islands between Australia and New Guinea, is needed to better characterize the seemingly very sharp genetic divide between the two. Current data, as described in Chapter 2, already gives some hints of a closer relationship to New Guineans in some northern Australian individuals. Further work is needed to determine if this just reflects admixture in relatively recent times, or if there have been more long-standing contacts. The separation between Aboriginal Australian and New Guinean populations, long before rising sea levels, has no obvious geographical, climatic or cultural explanation. It clearly predates the development of agriculture in PNG, excluding this as a cause of cultural differentiation. In terms of geography and climate, northern Australia with its rain forests might in some respects be more similar to New Guinea than to the rest of Australia. Further understanding of the causes for this early split might need to come from fields other than population genetics.

Another question in the early history of Sahul relates to the peopling of the Melanesian islands, and the relationship of the populations here to those in the rest of continent. The more or less implicit assumption in the literature so far, seemingly quite natural on geographical grounds, is that Melanesian islanders are a sister group to Papua New Guineans, to the exclusion of Aboriginal Australians. The analyses here of Bismarck Archipelago groups (New Ireland and New Britain) are consistent with this; however, another study reported an increased Aboriginal Australian affinity in certain groups further east in the Pacific (Skoglund, et al. 2016). In some explorative analyses presented here, the Bismarck groups also displayed difficult-to-interpret behaviour in this direction. Further analyses, and more comprehensive sampling across near and far Melanesia, are needed to better understand these relationships. Another question is the time-scale of separation between Melanesian islanders and New Guinea mainlanders. The array genotypes indicate strong allele frequency differentiation, but whole-genome sequencing data will likely be needed to get at the question of actual split time. Such analyses will also need to overcome the confounding effects of the Southeast Asian admixture that is present in all island populations.

A question related to this, and one that will increase in relevance with an increasingly detailed understanding of Sahul and Melanesian population structure, is that of the origin of the Sahul-related ancestry of Polynesians. Seemingly all Polynesian groups, as far as Hawaii and Rapa Nui, carry on the order of 20% Sahul-related ancestry. Little is known about where they picked this up: whether it was from the mainland of New Guinea, the large Bismarck islands, smaller islands further out, or some combination of these.

While the big picture of Sahul history is that of isolation from the rest of the world, it has not been completely isolated. In New Guinea there has of course been Southeast Asian admixture in the Holocene. While the dominant view is that of a relatively rapid expansion in the last few thousand years originating in Taiwan, there are also suggestions of earlier migrations (Soares, et al. 2016). There is also still uncertainty about the timing of arrival from Southeast Asia of certain domesticated crops, as well as domesticated pigs. More genome-wide data from many different Southeast Asian groups might be needed to better characterize the history of migration into New Guinea. In Australia, the dingo is evidence of external contacts, and while no Southeast Asian or other pre-European colonization admixture has been found in Aboriginal Australian genomes, more geographically comprehensive studies are needed before it can be ruled out completely. Further, in particular genome-wide, studies of dingos, as well as the related New Guinean Singing dogs, and Southeast Asian and perhaps Polynesian dogs might also provide answers as to who brought these animals to Sahul.

Major questions remain on the population history within New Guinea and within Australia, histories we are now only beginning to scratch the surface of. In Australia, very little is known about population structure, and studies are complicated by the widespread and extensive European admixture as well as the fact that many Aboriginal Australians are unaware of their deeper geographical origins because of forced movements during European colonial rule. Efforts to make use of well-documented museum collections of hair samples, many probably also predating much of the European admixture, might be a promising way to overcome this issue (Tobler, et al. 2017). From current data, population structure at least between western and eastern groups appears to be fairly old, at least pre-Holocene. A study of mitochondrial genomes suggested that separation between present-day Aboriginal Australian groups even dates back all the way to the initial peopling $\sim$50 kya (Tobler, et al. 2017). However, the limited and stochastic information provided by uniparental lineages, especially in the presence of potentially strong drift in small sub-populations, arguably makes such a conclusion difficult to draw, and furthermore it is not compatible with current estimates of the split time from Papua New Guineans. In any case, a major challenge is to explain how potentially fairly deep genetic divergences within Australia relate to the fact that a single language family, the Pama-Nyungan family, is spoken across 90% of the continent. The shared languages imply contact on the timescale of the last few, perhaps at most 10, thousand years, and lessons from the rest of the world suggest that languages typically spread with genes. There was no agricultural or similar transition in Australia that could have driven this, but there was a change in stone tool technology, and population expansions among hunter-gatherers cannot be ruled out. Indeed, such expansions have been demonstrated from ancient DNA in Pleistocene Europe (Fu, et al. 2016). Understanding population dynamics within Australia, including studying the non-Pama-Nyungan speaking groups in the northern parts of the continent, should be a key focus of future studies. Lastly, the genetic history of Tasmania remains virtually completely unknown, except for mitochondrial studies which at least have linked them to mainland Aboriginal Australians (Presser, et al. 2002; McAllister, et al. 2013). No unadmixed Aboriginal Tasmanians remain today, meaning that any insights would need to come from analyses of admixed genomes or ancient DNA.

In New Guinea, further work is needed to understand the population genetic consequences of the agricultural transition, and the extent to which current population structure was shaped by this as opposed to earlier, or later, processes. Whole-genome sequencing data from larger number of lowland groups (only one, from the Sepik region, was available for the analyses presented here) is needed to obtain a better understanding of the timing of separation between highlanders and lowlanders, and the demographic histories of the latter. This might illuminate whether or not the separation definitely occurred prior to the development of agriculture in the highlands, and thereby how agriculture and Trans-New Guinea languages reached the lowlands. Additionally, data from the Indonesian half of New Guinea island, so far very much underrepresented in genetic studies relative to the PNG half, is needed to complete the picture of New Guinean population structure.

Very little is known about positive selection in the history of the populations of Sahul. An analysis in (Malaspinas, et al. 2016) identified a number of candidate loci under selection in Aboriginal Australians, with possible links to adaptation to desert cold and dehydration; however, these results were tentative. Positive selection studies in other populations have uncovered several interesting instances of adaptation to local environments and lifestyles, and given the long, independent history of Sahul there might be good opportunities to make more here. Such discoveries could provide general insights into human biology and physiology, as well as insights specifically into the population and lifestyle histories of humans in Sahul. A question of particular interest is if any genetic variants were exposed to positive selection during the agricultural transition in New Guinea, and if so, how that process compares to agricultural adaptations in other parts of the world.

Ultimately, ancient DNA will be necessary to gain a truly in-depth understanding of Sahul population history. Data from ancient individuals would illuminate every aspect of this history, including the timing and nature of the Aboriginal Australian and Papuan split, positive selection, population structure and dynamics within Australia and the agricultural transition in New Guinea. While e.g. the first of these questions might require DNA from samples that are very old, perhaps on the order of 30 ky or more, the latter two would be greatly informed by samples that are 10 ky old or less. While successful DNA extraction has traditionally been mostly limited to samples from colder environments, recent technological improvements have enabled analyses of samples from e.g. eastern Africa (Gallego Llorente, et al. 2015), the Near East (Lazaridis, et al. 2016; Haber, et al. 2017) and Pacific islands (Skoglund, et al. 2016). There are thus real prospects for ancient DNA from Sahul, especially from the drier environments of Australia. The humid climate of PNG will likely prove more challenging, but on the other hand, the highlands are relatively cool and even feature snow-covered mountaintops. Ancient DNA is currently transforming our understanding of human evolutionary history in many parts of the world, and will be essential in Sahul as well.

5.2 Future directions for human population history generally

While the genetic history of western Eurasian and Native American populations is now becoming understood in some detail, much is still unclear about the history of eastern Eurasians. This includes the early events in the diversification between the ancestors of present-day East Asians, the populations of Sahul and those of Island Southeast Asia, as well as the admixture with Denisovans.

A number of so-called Negrito populations in Southeast Asia appear to share some parts of their early history with Aboriginal Australian and Papuans, after the separation from the ancestors of East Asians. This includes certain populations in the Philippines and Malaysia, as well as the Andaman islanders. However, the shape of the relationships of these various groups to each other is not understood with high confidence. Many of these Southeast Asian groups display a higher affinity to East Asians than Aboriginal Australian and Papuans do, likely due to East Asian-related admixture. This admixture could be relatively recent, i.e. occurring in the last 10 ky following the expansions of agriculturalists across Southeast Asia, and/or potentially more ancient, occurring during the early diversification of these lineages $\sim$50 kya. The observation that certain Philippine groups harbour more Denisovan admixture than expected given their levels of East Asian versus Sahul related ancestry suggests a more complex history than one of simple East Asian dilution (Reich, et al. 2011). Some models of the population split and admixture topologies relating these various lineages have been put forth (Reich, et al. 2011; Lipson and Reich 2017), but further work is needed. Furthermore, it has arguably not been conclusively determined if the small amounts (on the order of 0.1%) of Denisovan ancestry that is present in South and East Asian populations derives from the same admixture event as the material that is present in large amounts (∼4%) in Aboriginal Australians and Papuans. Analyses of predicted Denisovan haplotypes across these groups, including the Philippine Negritos, might provide further insights into the details of Denisovan gene flow into modern humans, as well as into the relationships between the modern human groups themselves.

The relationships between these populations also need to be placed onto a temporal axis – little is known about e.g. the genetic split times between the various Negrito groups, Sahul populations and East Asians. Current split time estimates between Sahul populations and East Asians are associated with technical uncertainty and need to be made more confident, e.g. through application of MSMC to experimentally phased genomes as well as independent dating methods. The determination of split times between Southeast Asian groups will likely be complicated by complex histories of admixture, such that simple pairwise split estimates might not be very meaningful. Decomposition of such estimates given knowledge of overall ancestry proportions (Pagani, et al. 2016), or more explicit temporal model fitting based on e.g. the site frequency spectrum (Excoffier and Foll 2011) or rare variant sharing (Schiffels, et al. 2016), might be fruitful.

There are observations that hint at the possibility that populations with a Negrito-like ancestry were once more widespread, not just across Island Southeast Asia, but perhaps even mainland Asia. First, the genetic split time between Andamanese islanders, in the past sometimes informally hypothesized to have been isolated in the Indian Ocean since shortly after the migration of humans out of Africa, and other Eurasian populations appears to have occurred closer to ∼20 kya (Mondal, et al. 2016). Second, a slightly higher affinity towards Negrito and Sahul populations among some South American groups relative to other Native Americans (Skoglund, et al. 2015) suggests structure in the Siberian source populations ∼20 kya that was correlated to Negrito affinity. Ancient DNA from East Asia might reveal populations with this kind of ancestry, that perhaps have been replaced or absorbed by expanding groups harbouring the ancestry that today dominates across mainland East Asia.

In mainland East Asia, very little is known about the processes underlying current population structure. One of the key aspects that needs an explanation is the genetic homogeneity of the region, which overall is comparable to Europe, and strikingly different from Papua New Guinea. For example, F_{ST} between Japanese and Vietnamese is approximately 1%. It is tempting to speculate that, similarly to Europe, this is the result of one or more recent expansions after the development of agriculture. Studying the factors behind the genetic homogenization of East Asia, and comparing them to what did or did not happen in Europe, Papua New Guinea and elsewhere, will provide further fuel to this "comparative population genetics" approach to understanding the driving forces in human history. Further work is needed in East Asia, and, as elsewhere, ancient DNA will be required to achieve an in-depth understanding.

In a global context, the biggest gaps in our understanding of human population history are arguably in Africa. A few basic features seem to be firmly established, including the early divergence of the Khoe-San people of southern Africa, the divergence of central African rainforest hunter-gatherers after that, the diversification of what is largely the present-day western and eastern African ancestries after that, large-scale migration and admixture following the Bantu expansion in the last few thousand years, and gene flow from non-African sources in the Middle East also in the last few thousand years. Some of the earlier population structure has likely been obscured by the recent expansions and admixture. There is still great uncertainty about the shape and timing of the early diversification events within Africa, which likely occurred somewhere between 100-200 kya, but potentially even earlier. Understanding the time depth of human population structure is of key interest not least because it puts a lower bound on the emergence of anatomically modern human traits.

An increased understanding of African population history might also aid the understanding of the dispersal out of Africa. In principle, one kind of observation that would provide evidence that different non-African populations derive ancestry from different migrations out of Africa would be differences in their relationships to African population structure. For example, if Aboriginal Australians and Papuan showed higher affinity to western Africans while Eurasian showed higher affinity to eastern Africans, this would constitute such evidence. A major obstacle, however, is the extensive backflow that has occurred into Africa from sources more closely related to Middle Easterners and Europeans than to other worldwide populations, affecting especially eastern, but also southern, Africa. Further to this, a key question is where in the African population history topology the group that became non-Africans fits in, and the related question of which present-day African populations are most closely related to non-Africans. Because of population dynamics within Africa in the last 50 ky, it is, however, very unlikely that the direct sister group to non-Africans will still exist in un-admixed form. Modelling approaches that incorporate past admixture events will be required to elucidate the complex population history within Africa.

In recent years there have been a number of archaeological findings of human presence outside of Africa at times earlier than predicted by current population genetic models, e.g. in China 100 kya (Liu, et al. 2015), North America 120 kya (Holen, et al. 2017), Australia 65 kya (Clarkson, et al. 2017), as discussed in Chapter 2, and Sumatra (Westaway, et al. 2017) 63-73 kya, and a similar suggestion from analyses of Neanderthal genomes (Kuhlwilm, et al. 2016). An important question is if these are traces of people that are the ancestors of present-day non-Africans, and that we are currently underestimating the time depth of human dispersal, or if they reflect earlier migrating groups that died out and did not contribute detectable ancestry to present-day people. There are observations that arguably provide reasonably firm upper bounds on the timing of certain events, e.g. the timing of Neanderthal admixture (Fu, et al. 2014) or the divergence of non-African uniparental chromosomes from African ones (Fu, Mittnik, et al. 2013; Poznik, et al. 2016), and which therefore are incompatible with at least some of the earliest archaeological findings being associated with ancestors of present-day populations. However, recent findings from ancient DNA are showing that humans have often been highly mobile in the past, and individuals from ancient populations that do not seem to have contributed to present-day people have been discovered (Fu, et al. 2014; Fu, et al. 2015), so the idea of small groups of humans moving large distances and living in a new area for a while before dying out might now appear more plausible.

Another question that is still unresolved is that of whether or not there was admixture into modern humans from any archaic human groups within Africa. In the absence of ancient DNA from any such groups, studies analysing particularly divergent haplotypes have suggested evidence for such "ghost admixture" in certain present-day African populations (Hammer, et al. 2011; Hsieh, Woerner, et al. 2016). Such analyses are necessarily highly statistical and indirect, but perhaps larger amounts of data from diverse African populations will allow these analyses to be more conclusive. However, while conditions for ancient DNA preservation are poor in most of Africa, ideally direct tests will be made possible at some point, with DNA isolated from human remains or possibly even recovered from the environment (Willerslev, et al. 2003; Slon, et al. 2017).

5.3 Concluding remarks

It is truly an exciting time for the field of human population genetics. Many important findings have been made even just while the work described in this thesis was carried out, and more will undoubtedly come in the near future. It is becoming increasingly apparent that ancient DNA is extremely valuable for understanding population histories. While inference from modern genomes often requires highly sophisticated analyses to infer properties of past populations, the power of ancient DNA is that it allows us to observe genomes from such populations directly. However, there is still arguably value in generating data from modern genomes. The higher quality and larger amounts of data typically obtained serve as useful reference points for ancient data, and also more readily enable certain analyses that dependent on high data quality. For example, ancient DNA studies have not provided much insight into changes in effective population sizes over time, while analyses of modern genomes have. There might also be practical limits on how much ancient DNA will be generated in certain parts of the world and from certain time periods. In particular, to address the deep history of humanity in Africa on the order of $\sim$200 kya, or maybe even earlier, we might still have to rely on indirect inferences from combinations of modern and ancient genomes, rather than just observations of ancient DNA directly from those periods. Nonetheless, ancient DNA will likely drive most progress in the field from now onwards.

Over the next few years and onwards, very large numbers of human genomes will be sequenced, mainly for biomedical purposes. There are potentially exciting prospects for gaining population-genetic insights from such data. While simply having a very large number of genomes from a given population is unlikely to provide any dramatic gains to the understanding of the history of that population and its relationships to others, such datasets might greatly inform the more functional and mechanistic aspects of human genetics. This could include questions related to the mechanisms and consequences of mutation and recombination, gene essentiality and disease penetrance, positive and purifying selection and classical genetics concepts such as epistasis, dominance and heritability. With a greater functional characterization of human biology, including an ever-expanding list of variants that are associated with trait variation, will also come challenges and opportunities to understand this biology in a population history framework.

Lastly, efforts need to be made to ensure that not only people in the more economically privileged parts of the world benefit from the progress in human genomics. This argument is often made in the context of medical genetics, where results from studies carried out in populations of European ancestry might not translate well to other populations, but a version of the argument could also be applied to the study of population history. Currently, the part of the world with by far the best understood genetic history is Europe. There are practical and logistical reasons for this, including good conditions for ancient DNA preservation in many parts of Europe and comprehensive previous archaeological work, but greater efforts need nonetheless be made to study other parts of the world in greater detail as well. Such studies will be needed to gain a truly comprehensive picture of human evolutionary history and diversity.

Bibliography

1000 Genomes Project Consortium. 2015. *A global reference for human genetic variation*. Nature 526:68-74.

1000 Genomes Project Consortium. 2012. *An integrated map of genetic variation from 1,092 human genomes.* Nature 491:56-65.

1000 Genomes Project Consortium. 2010. *A map of human genome variation from population-scale sequencing.* Nature 467:1061-1073.

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

Allentoft ME, Sikora M, Sjogren KG, Rasmussen S, Rasmussen M, Stenderup J, Damgaard PB, Schroeder H, Ahlstrom T, Vinner L, et al. 2015. *Population genomics of Bronze Age Eurasia.* Nature 522:167-172.

Ardalan A, Oskarsson M, Natanaelsson C, Wilton AN, Ahmadian A, Savolainen P. 2012. *Narrow genetic basis for the Australian dingo confirmed through analysis of paternal ancestry*. Genetica 140:65-73.

Argue D, Groves CP, Lee MSY, Jungers WL. 2017. *The affinities of Homo floresiensis based on phylogenetic analyses of cranial, dental, and postcranial characters.* Journal of Human Evolution 107:107-133.

Arthur R, Schulz-Trieglaff O, Cox AJ, O'Connell J. 2017. *AKT: ancestry and kinship toolkit*. Bioinformatics 33:142-144.

Athanasiadis G, Cheng JY, Vilhjalmsson BJ, Jorgensen FG, Als TD, Le Hellard S, Espeseth T, Sullivan PF, Hultman CM, Kjaergaard PC, et al. 2016. *Nationwide Genomic Study in Denmark Reveals Remarkable Population Homogeneity.* Genetics 204:711-722.

Berg JJ, Coop G. 2014. *A population genetic signal of polygenic adaptation*. PLoS Genet 10:e1004412.

Bergström A, Nagle N, Chen Y, McCarthy S, Pollard MO, Ayub Q, Wilcox S, Wilcox L, van Oorschot RA, McAllister P, et al. 2016. *Deep Roots for Aboriginal Australian Y Chromosomes.* Curr Biol 26:809-813.

Bergström A, Oppenheimer SJ, Mentzer AJ, Auckland K, Robson K, Attenborough R, Alpers MP, Koki G, Pomat W, Siba P, et al. 2017. *A Neolithic expansion, but strong genetic structure, in the independent history of New Guinea.* Science 357:1160-1163.

Bigham AW. 2016. *Genetics of human origin and evolution: high-altitude adaptations.* Curr Opin Genet Dev 41:8-13.

Bourke RM, Harwood T, Ed. 2009. *Food and agriculture in Papua New Guinea*. The Australian National University: ANU E Press.

Bowern C. 2010. *Historical linguistics in Australia: trees, networks and their implications.* Philos Trans R Soc Lond B Biol Sci 365:3845-3854.

Brisbin A, Bryc K, Byrnes J, Zakharia F, Omberg L, Degenhardt J, Reynolds A, Ostrer H, Mezey JG, Bustamante CD. 2012. *PCAdmix: principal components-based assignment of ancestry along* *each chromosome in individuals with admixed ancestry from two or more populations.* Hum Biol 84:343-364.

Brown P. 2013. *Palaeoanthropology: Of humans, dogs and tiny tools.* Nature 494:316-317.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, et al. 2002. *A human genome diversity cell line panel.* Science 296:261-262.

Cann RL, Stoneking M, Wilson AC. 1987. *Mitochondrial DNA and human evolution.* Nature 325:31-36.

Clarkson C, Jacobs Z, Marwick B, Fullagar R, Wallis L, Smith M, Roberts RG, Hayes E, Lowe K, Carah X, et al. 2017. *Human occupation of northern Australia by 65,000 years ago.* Nature 547:306-310.

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

Darwin C. 1859. *On the Origin of Species by means of Natural Selection, or the preservation of favoured races in the struggle for life.* London: John Murray.

Delaneau O, Marchini J, Zagury JF. 2011. *A linear complexity phasing method for thousands of genomes.* Nat Methods 9:179-181.

Denham T, Barton H. 2006. *The emergence of agriculture in New Guinea: a model of continuity from pre-existing foraging practices.* In: Kennett DJ, Winterhalder B, editors. Behavioral Ecology and the Transition to Agriculture:237–264

Dillehay TD, Ramirez C, Pino M, Collins MB, Rossen J, Pino-Navarro JD. 2008. *Monte Verde: seaweed, food, medicine, and the peopling of South America.* Science 320:784-786.

Duggan AT, Stoneking M. 2014. *Recent developments in the genetic history of East Asia and Oceania.* Curr Opin Genet Dev 29:9-14.

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

Excoffier L, Foll M. 2011. *fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios.* Bioinformatics 27:1332-1334.

Fan S, Hansen ME, Lo Y, Tishkoff SA. 2016. *Going global by adapting local: A review of recent human adaptation.* Science 354:54-59.

Feil DK. 1987. *The Evolution of Highland Papua New Guinea Societies.* Cambridge University Press.

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

Field LL, Tobias R, Robinson WP, Paisey R, Bain S. 1998. *Maternal uniparental disomy of chromosome 1 with no apparent phenotypic effects.* Am J Hum Genet 63:1216-1220.

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, et al. 2016. *Detection of human adaptation during the past 2000 years.* Science 354:760-764.

Forsberg LA. 2017. *Loss of chromosome Y (LOY) in blood cells is associated with increased risk for disease and mortality in aging men.* Hum Genet 136:657-663.

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

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo JH, Koki G, Hodgson JA, et al. 2008. *The genetic structure of Pacific Islanders.* PLoS Genet 4:e19.

Fu Q, Hajdinjak M, Moldovan OT, Constantin S, Mallick S, Skoglund P, Patterson N, Rohland N, Lazaridis I, Nickel B, et al. 2015. *An early modern human from Romania with a recent Neanderthal ancestor.* Nature 524:216-219.

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

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Paabo S. 2013. *DNA analysis of an early modern human from Tianyuan Cave, China.* Proc Natl Acad Sci U S A 110:2223-2227.

Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, et al. 2013. *A revised timescale for human evolution based on ancient mitochondrial genomes.* Curr Biol 23:553-559.

Fu Q, Posth C, Hajdinjak M, Petr M, Mallick S, Fernandes D, Furtwangler A, Haak W, Meyer M, Mittnik A, et al. 2016. *The genetic history of Ice Age Europe.* Nature 534:200-205.

Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jorgensen ME, Korneliussen TS, Gerbault P, Skotte L, Linneberg A, et al. 2015. *Greenlandic Inuit show genetic signatures of diet and climate adaptation.* Science 349:1343-1347.

Gallego Llorente M, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, et al. 2015. *Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent.* Science 350:820-822.

Garrison E, Marth G. 2012. *Haplotype-based variant detection from short-read sequencing.* arXiv:1207.3907.

Genome of the Netherlands C. 2014. *Whole-genome sequence variation, population structure and demographic history of the Dutch population.* Nat Genet 46:818-825.

Gittelman RM, Schraiber JG, Vernot B, Mikacenic C, Wurfel MM, Akey JM. 2016. *Archaic Hominin Admixture Facilitated Adaptation to Out-of-Africa Environments.* Curr Biol 26:3375-3382.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Genomes P, Bustamante CD. 2011. *Demographic history and rare allele sharing among human populations.* Proc Natl Acad Sci U S A 108:11983-11988.

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

Gunther T, Jakobsson M. 2016. *Genes mirror migrations and cultures in prehistoric Europe-a population genomic perspective.* Curr Opin Genet Dev 41:115-123.

Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. *The African Genome Variation Project shapes medical genetics in Africa.* Nature 517:327-332.

Haak W, Lazaridis I, Patterson N, Rohland N, Mallick S, Llamas B, Brandt G, Nordenfelt S, Harney E, Stewardson K, et al. 2015. *Massive migration from the steppe was a source for Indo-European languages in Europe.* Nature 522:207-211.

Haber M, Doumet-Serhal C, Scheib C, Xue Y, Danecek P, Mezzavilla M, Youhanna S, Martiniano R, Prado-Martinez J, Szpak M, et al. 2017. *Continuity and Admixture in the Last Five Millennia of Levantine History from Ancient Canaanite and Present-Day Lebanese Genome Sequences.* Am J Hum Genet 101:274-282.

Haber M, Mezzavilla M, Xue Y, Tyler-Smith C. 2016. *Ancient DNA and the rewriting of human history: be sparing with Occam's razor.* Genome Biol 17:1.

Hammer MF, Woerner AE, Mendez FL, Watkins JC, Wall JD. 2011. *Genetic evidence for archaic admixture in Africa.* Proc Natl Acad Sci U S A 108:15123-15128.

Harris K, Nielsen R. 2016. *The Genetic Cost of Neanderthal Introgression.* Genetics 203:881-891.

Hartl DL, Clark AG. 2007. *Principles of population genetics.* New York: W. H. Freeman ; Basingstoke : Palgrave.

Helgason A, Einarsson AW, Gudmundsdóttir VB, Sigurdsson Á, Gunnarsdóttir ED, Jagadeesan A, Ebenesersdóttir SS, Kong A, Stefánsson K. 2015. *The Y-chromosome point mutation rate in humans.* Nat Genet 47:453-457.

Henn BM, Botigue LR, Gravel S, Wang W, Brisbin A, Byrnes JK, Fadhlaoui-Zid K, Zalloua PA, Moreno-Estrada A, Bertranpetit J, et al. 2012. *Genomic ancestry of North Africans supports back-to-Africa migrations.* PLoS Genet 8:e1002397.

Higham T, Compton T, Stringer C, Jacobi R, Shapiro B, Trinkaus E, Chandler B, Groning F, Collins C, Hillson S, et al. 2011. *The earliest evidence for anatomically modern humans in northwestern Europe.* Nature 479:521-524.

Holen SR, Demere TA, Fisher DC, Fullagar R, Paces JB, Jefferson GT, Beeton JM, Cerutti RA, Rountrey AN, Vescera L, et al. 2017. *A 130,000-year-old archaeological site in southern California, USA.* Nature 544:479-483.

Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD, et al. 2015. *Genomic Insights into the Ancestry and Demographic History of South America.* PLoS Genet 11:e1005602.

Howie BN, Donnelly P, Marchini J. 2009. *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet 5:e1000529.

Hsieh P, Veeramah KR, Lachance J, Tishkoff SA, Wall JD, Hammer MF, Gutenkunst RN. 2016. *Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection.* Genome Res 26:279-290.

Hsieh P, Woerner AE, Wall JD, Lachance J, Tishkoff SA, Gutenkunst RN, Hammer MF. 2016. *Model-based analyses of whole-genome data reveal a complex evolutionary history involving archaic introgression in Central African Pygmies.* Genome Res 26:291-300.

Hubisz MJ, Falush D, Stephens M, Pritchard JK. 2009. *Inferring weak population structure with the assistance of sample group information.* Mol Ecol Resour 9:1322-1332.

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

Huffman TN. 1982. *Archaeology and Ethnohistory of the African Iron-Age*. Annual Review of Anthropology 11:133-150.

Hugo Pan-Asian SNP Consortium, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen CH, Chen J, et al. 2009. *Mapping human genetic diversity in Asia.* Science 326:1541-1545.

International HapMap Consortium. 2005. *A haplotype map of the human genome.* Nature 437:1299-1320.

Jablonski NG, Chaplin G. 2010. *Human skin pigmentation as an adaptation to UV radiation.* Proc Natl Acad Sci U S A 107 Suppl 2:8962-8968.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, et al. 2008. *Genotype, haplotype and copy-number variation in worldwide human populations.* Nature 451:998-1003.

Jobling MA, Hurles M, Tyler-Smith C. 2004. *Human evolutionary genetics: origins, peoples & disease.* New York ; Abingdon: Garland Science.

Juric I, Aeschbacher S, Coop G. 2016. *The Strength of Selection against Neanderthal Introgression.* PLoS Genet 12:e1006340.

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. *Natural selection and infectious disease in human populations.* Nat Rev Genet 15:379-393.

Karmin M, Saag L, Vicente M, Wilson Sayres MA, Jarve M, Talas UG, Rootsi S, Ilumae AM, Magi R, Mitt M, et al. 2015. *A recent bottleneck of Y chromosome diversity coincides with a global change in culture.* Genome Res 25:459-466.

Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill P, Shen P, Oefner P, Tommaseo-Ponzetta M, Stoneking M. 2003. *Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea.* Am J Hum Genet 72:281-302.

Kayser M, Lao O, Saar K, Brauer S, Wang X, Nurnberg P, Trent RJ, Stoneking M. 2008. *Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians.* Am J Hum Genet 82:194-198.

Kim BY, Lohmueller KE. 2015. *Selection and reduced population size cannot explain higher amounts of Neandertal ancestry in East Asian than in European human populations*. Am J Hum Genet 96:454-461.

Kim HL, Ratan A, Perry GH, Montenegro A, Miller W, Schuster SC. 2014. *Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history.* Nat Commun 5:5692.

King DA, Fitzgerald TW, Miller R, Canham N, Clayton-Smith J, Johnson D, Mansour S, Stewart F, Vasudevan P, Hurles ME, et al. 2014. *A novel method for detecting uniparental disomy from trio genotypes identifies a significant excess in children with developmental disorders.* Genome Res 24:673-687.

Kuhlwilm M, Gronau I, Hubisz MJ, de Filippo C, Prado-Martinez J, Kircher M, Fu Q, Burbano HA, Lalueza-Fox C, de la Rasilla M, et al. 2016. *Ancient gene flow from early modern humans into Eastern Neanderthals.* Nature 530:429-433.

Lahr MM, Foley R. 1994. *Multiple dispersals and modern human origins.* Evolutionary Anthropology 3:48–60.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. *Initial sequencing and analysis of the human genome*. Nature 409:860-921.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, et al. 2008. *Correlation between genetic and geographic structure in Europe*. Curr Biol 18:1241-1248.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. *Inference of population structure using dense haplotype data*. PLoS Genet 8:e1002453.

Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, Fernandes D, Novak M, Gamarra B, Sirak K, et al. 2016. *Genomic insights into the origin of farming in the ancient Near East.* Nature 536:419-424.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. *Ancient human genomes suggest three ancestral populations for present-day Europeans.* Nature 513:409-413.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. *Analysis of protein-coding genetic variation in 60,706 humans.* Nature 536:285-291.

Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, Hutnik K, Royrvik EC, Cunliffe B, Wellcome Trust Case Control C, et al. 2015. *The fine-scale genetic structure of the British population.* Nature 519:309-314.

Lewis MP, Simons GF, Fennig CD. 2016. *Ethnologue: Languages of the World*, Nineteenth edition.

Lewis SE, Sloss CR, Murray-Wallace CV, Woodroffe CD, Smithers SG. 2013. *Post-glacial sea-level changes around the Australian margin: a review.* Quaternary Science Reviews 74:115-138.

Li H. 2011. *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics 27:2987-2993.

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

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

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

Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. 2008. *Worldwide human relationships inferred from genome-wide patterns of variation.* Science 319:1100-1104.

Li S, Schlebusch C, Jakobsson M. 2014. *Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples*. Proc Biol Sci 281.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. *Low-coverage sequencing: implications for design of complex trait association studies.* Genome Res 21:940-951.

Lipson M, Loh PR, Patterson N, Moorjani P, Ko YC, Stoneking M, Berger B, Reich D. 2014. *Reconstructing Austronesian population history in Island Southeast Asia.* Nat Commun 5:4689.

Lipson M, Reich D. 2017. *A Working Model of the Deep Relationships of Diverse Modern Human Genetic Lineages Outside of Africa.* Mol Biol Evol 34:889-902.

Liu W, Martinon-Torres M, Cai YJ, Xing S, Tong HW, Pei SW, Sier MJ, Wu XH, Edwards RL, Cheng H, et al. 2015. *The earliest unequivocally modern humans in southern China.* Nature 526:696-699.

Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, Schoenherr S, Forer L, McCarthy S, Abecasis GR, et al. 2016. *Reference-based phasing using the Haplotype Reference Consortium panel.* Nat Genet 48:1443-1448.

Luca F, Perry GH, Di Rienzo A. 2010. *Evolutionary adaptations to dietary changes.* Annu Rev Nutr 30:291-314.

Lynch M. 2007. *The origins of genome architecture.* Sunderland, Mass.: Sinauer Associates.

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, et al. 2005. *Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes.* Science 308:1034-1036.

Macknight CC. 1986. *Macassans and the Aboriginal Past.* Archaeology in Oceania 21:69-75.

Malaspinas AS, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE, et al. 2016. *A genomic history of Aboriginal Australia.* Nature 538:207-214.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A, et al. 2016. *The Simons Genome Diversity Project: 300 genomes from 142 diverse populations.* Nature 538:201-206.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. *RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference.* Am J Hum Genet 93:278-288.

Mathieson I, McVean G. 2014. *Demography and the age of rare variants.* PLoS Genet 10:e1004528.

McAllister P, Nagle N, Mitchell RJ. 2013. *The Australian Barrineans and their relationship to Southeast Asian negritos: an investigation using mitochondrial genomics.* Hum Biol 85:485-494.

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. 2016. *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet 48:1279-1283.

Mcconvell P. 1996. *Backtracking to Babel: The Chronology of Pama-Nyungan Expansion in Australia.* Archaeology in Oceania 31:125-144.

McDermott F, Grun R, Stringer CB, Hawkesworth CJ. 1993. *Mass-spectrometric U-series dates for Israeli Neanderthal/early modern hominid sites.* Nature 363:252-255.

McEvoy BP, Lind JM, Wang ET, Moyzis RK, Visscher PM, van Holst Pellekaan SM, Wilton AN. 2010. *Whole-genome genetic diversity in a sample of Australians with deep Aboriginal ancestry.* Am J Hum Genet 87:297-305.

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

McVean G. 2009. *A genealogical interpretation of principal components analysis*. PLoS Genet 5:e1000686.

Mellars P. 2006. *A new radiocarbon revolution and the dispersal of modern humans in Eurasia.* Nature 439:931-935.

Metzker ML. 2010. *Sequencing technologies - the next generation.* Nat Rev Genet 11:31-46.

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, et al. 2012. *A high-coverage genome sequence from an archaic Denisovan individual*. Science 338:222-226.

Migliano AB, Romero IG, Metspalu M, Leavesley M, Pagani L, Antao T, Huang DW, Sherman BT, Siddle K, Scholes C, et al. 2013. *Evolution of the pygmy phenotype: evidence of positive selection fro genome-wide scans in African, Asian, and Melanesian pygmies.* Hum Biol 85:251-284.

Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, Comas D, Laayouni H, Li Q, Majumder PP, et al. 2016. *Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation*. Nat Genet 48:1066-1070.

Moorjani P, Gao Z, Przeworski M. 2016. *Human Germline Mutation and the Erratic Evolutionary Clock.* PLoS Biol 14:e2000744.

Moorjani P, Sankararaman S, Fu Q, Przeworski M, Patterson N, Reich D. 2016. *A genetic method for dating ancient genomes provides a direct estimate of human generation interval in the last 45,000 years.* Proc Natl Acad Sci U S A 113:5652-5657.

Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh PR, Govindaraj P, Berger B, Reich D, Singh L. 2013. *Genetic evidence for recent population mixture in India.* Am J Hum Genet 93:422-438.

Moreno-Estrada A, Gignoux CR, Fernandez-Lopez JC, Zakharia F, Sikora M, Contreras AV, Acuna-Alonzo V, Sandoval K, Eng C, Romero-Hidalgo S, et al. 2014. *The genetics of Mexico recapitulates Native American substructure and affects biomedical traits.* Science 344:1280-1285.

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martinez RJ, Hedges DJ, Morris RW, et al. 2013. *Reconstructing the population genetic history of the Caribbean.* PLoS Genet 9:e1003925.

Moreno-Mayar JV, Rasmussen S, Seguin-Orlando A, Rasmussen M, Liang M, Flam ST, Lie BA, Gilfillan GD, Nielsen R, Thorsby E, et al. 2014. *Genome-wide ancestry patterns in Rapanui suggest pre-European admixture with Native Americans*. Curr Biol 24:2518-2525.

Nagle N, Ballantyne KN, van Oven M, Tyler-Smith C, Xue Y, Taylor D, Wilcox S, Wilcox L, Turkalov R, van Oorschot RA, et al. 2016. *Antiquity and diversity of aboriginal Australian Y-chromosomes.* Am J Phys Anthropol 159:367-381.

Nagle N, Ballantyne KN, van Oven M, Tyler-Smith C, Xue Y, Wilcox S, Wilcox L, Turkalov R, van Oorschot RA, van Holst Pellekaan S, et al. 2017. *Mitochondrial DNA diversity of present-day Aboriginal Australians and implications for human evolution in Oceania*. J Hum Genet 62:343-353.

Nakatsuka N, Moorjani P, Rai N, Sarkar B, Tandon A, Patterson N, Bhavani GS, Girisha KM, Mustak MS, Srinivasan S, et al. 2017. *The promise of discovering population-specific disease-associated genes in South Asia*. Nat Genet.

Nelson MR, Bryc K, King KS, Indap A, Boyko AR, Novembre J, Briley LP, Maruyama Y, Waterworth DM, Waeber G, et al. 2008. *The Population Reference Sample, POPRES: a resource for* *population, disease, and pharmacological genetics research.* Am J Hum Genet 83:347-358.

Nordborg M. 1998. *On the probability of Neanderthal ancestry.* Am J Hum Genet 63:1237-1240.

Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. 2008. *Genes mirror geography within Europe*. Nature 456:98-101.

O'Connell JF, Allen J. 2015. *The process, biotic impact, and global implications of the human colonization of Sahul about 47,000 years ago.* Journal of Archaeological Science 56:73-84.

Orlando L, Gilbert MT, Willerslev E. 2015. *Reconstructing ancient genomes and epigenomes.* Nat Rev Genet 16:395-408.

Oskarsson MC, Klutsch CF, Boonyaprakob U, Wilton A, Tanabe Y, Savolainen P. 2012. *Mitochondrial DNA data indicate an introduction through Mainland Southeast Asia for Australian dingoes and Polynesian domestic dogs*. Proc Biol Sci 279:967-974.

Pagani L, Kivisild T, Tarekegn A, Ekong R, Plaster C, Gallego Romero I, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, et al. 2012. *Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool.* Am J Hum Genet 91:83-96.

Pagani L, Lawson DJ, Jagoda E, Morseburg A, Eriksson A, Mitt M, Clemente F, Hudjashov G, DeGiorgio M, Saag L, et al. 2016. *Genomic analyses inform on migration events during the peopling of Eurasia.* Nature 538:238-242.

Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, Xue Y, Haber M, Ekong R, Oljira T, et al. 2015. *Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians.* Am J Hum Genet 96:986-991.

Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, et al. 2017. *Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America.* Science 356:543-546.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. *Ancient admixture in human history.* Genetics 192:1065-1093.

Patterson N, Price AL, Reich D. 2006. *Population structure and eigenanalysis*. PLoS Genet 2:e190.

Pawley A. 2005. *The chequered career of the Trans New Guinea hypothesis: recent research and its implications.* In: Papuan Pasts: cultural, linguistic and biological histories of Papuan-speaking peoples. A. Pawley, R. Attenborough, J. Golson, R. Hide. Eds.:67–107.

Perry GH, Foll M, Grenier JC, Patin E, Nedelec Y, Pacis A, Barakatt M, Gravel S, Zhou X, Nsobya SL, et al. 2014. *Adaptive, convergent origins of the pygmy phenotype in African rainforest hunter-gatherers*. Proc Natl Acad Sci U S A 111:E3596-3603.

Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, Pakendorf B, Reich D. 2014. *Ancient west Eurasian ancestry in southern and eastern Africa.* Proc Natl Acad Sci U S A 111:2632-2637.

Ponting CP. 2017. *Biological function in the twilight zone of sequence conservation.* BMC Biol 15:71.

Posth C, Renaud G, Mittnik A, Drucker DG, Rougier H, Cupillard C, Valentin F, Thevenet C, Furtwangler A, Wissing C, et al. 2016. *Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe.* Curr Biol 26:827-833.

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

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. *Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences.* Nat Genet 48:593-599.

Presser JC, Redd AJ, Stoneking M. 2002. *Tasmanian Aborigines and DNA.* Papers and Proceedings of the Royal Society of Tasmania 136:35-38.

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S. 2009. *Sensitive detection of chromosomal segments of distinct ancestry in admixed populations.* PLoS Genet 5:e1000519.

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

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

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

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

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford TW, Jr., Orlando L, Metspalu E, et al. 2014. *Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans*. Nature 505:87-91.

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

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

Redd AJ, Roberts-Thomson J, Karafet T, Bamshad M, Jorde LB, Naidu JM, Walsh B, Hammer MF. 2002. *Gene flow from the Indian subcontinent to Australia: evidence from the Y chromosome.* Curr Biol 12:673-677.

Redd AJ, Stoneking M. 1999. *Peopling of Sahul: mtDNA variation in aboriginal Australian and Papua New Guinean populations.* Am J Hum Genet 65:808-828.

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

Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, Pugach I, Ko AMS, Ko YC, Jinam TA, Phipps ME, et al. 2011. *Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania.* American Journal of Human Genetics 89:516-528.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. *Reconstructing Indian population history.* Nature 461:489-494.

Reyes-Centeno H, Ghirotto S, Detroit F, Grimaud-Herve D, Barbujani G, Harvati K. 2014. *Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia.* Proc Natl Acad Sci U S A 111:7248-7253.

Riley ID. 1983. *Population-Change and Distribution in Papua-New-Guinea - an Epidemiological Approach*. Journal of Human Evolution 12:125-132.

Rosenberg NA. 2006. *Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives.* Ann Hum Genet 70:841-847.

Roullier C, Benoit L, McKey DB, Lebot V. 2013. *Historical collections reveal patterns of diffusion of sweet potato in Oceania obscured by modern plant movements and recombination.* Proc Natl Acad Sci U S A 110:2205-2210.

Sankararaman S, Mallick S, Dannemann M, Prufer K, Kelso J, Paabo S, Patterson N, Reich D. 2014. *The genomic landscape of Neanderthal ancestry in present-day humans.* Nature 507:354-357.

Sankararaman S, Mallick S, Patterson N, Reich D. 2016. *The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans.* Curr Biol 26:1241-1247.

Sankararaman S, Patterson N, Li H, Paabo S, Reich D. 2012. *The date of interbreeding between Neandertals and modern humans.* PLoS Genet 8:e1002947.

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

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

Schiffels S, Haak W, Paajanen P, Llamas B, Popescu E, Loe L, Clarke R, Lyons A, Mortimer R, Sayer D, et al. 2016. *Iron Age and Anglo-Saxon genomes from East England reveal British migration history.* Nat Commun 7:10408.

Shirley MD, Baugher JD, Stevens EL, Tang Z, Gerry N, Beiswanger CM, Berlin DS, Pevsner J. 2012. *Chromosomal variation in lymphoblastoid cell lines.* Hum Mutat 33:1075-1086.

Siddle KJ, Quintana-Murci L. 2014. *The Red Queen's long race: human adaptation to pathogen pressure.* Curr Opin Genet Dev 29:31-38.

Siska V, Jones ER, Jeon S, Bhak Y, Kim HM, Cho YS, Kim H, Lee K, Veselovskaya E, Balueva T, et al. 2017. *Genome-wide data from two early Neolithic East Asian individuals dating to 7700 years ago.* Science Advances 3.

Skoglund P, Jakobsson M. 2011. *Archaic human ancestry in East Asia.* Proc Natl Acad Sci U S A 108:18301-18306.

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hunemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D. 2015. *Genetic evidence for two founding populations of the Americas.* Nature 525:104-108.

Skoglund P, Malmstrom H, Omrak A, Raghavan M, Valdiosera C, Gunther T, Hall P, Tambets K, Parik J, Sjogren KG, et al. 2014. *Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers*. Science 344:747-750.

Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, Clark GR, Reepmeyer C, Petchey F, Fernandes D, et al. 2016. *Genomic insights into the peopling of the Southwest Pacific.* Nature 538:510-513.

Skoglund P, Reich D. 2016. *A genomic view of the peopling of the Americas*. Curr Opin Genet Dev 41:27-35.

Slatkin M, Racimo F. 2016. *Ancient DNA and human history.* Proc Natl Acad Sci U S A 113:6380-6387.

Slon V, Hopfe C, Weiss CL, Mafessoni F, de la Rasilla M, Lalueza-Fox C, Rosas A, Soressi M, Knul MV, Miller R, et al. 2017. *Neandertal and Denisovan DNA from Pleistocene sediments.* Science 356:605-608.

Soares PA, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, Mormina M, Brandao A, Fraser RM, Wang TY, et al. 2016. *Resolving the ancestry of Austronesian-speaking populations.* Hum Genet 135:309-326.

Song S, Sliwerska E, Emery S, Kidd JM. 2017. *Modeling Human Population Separation History Using Physically Phased Genomes*. Genetics 205:385-395.

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

Stoneking M, Jorde LB, Bhatia K, Wilson AC. 1990. *Geographic variation in human mitochondrial DNA from Papua New Guinea.* Genetics 124:717-733.

Summerhayes GR, Leavesley M, Fairbairn A, Mandui H, Field J, Ford A, Fullagar R. 2010. *Human Adaptation and Plant Use in Highland New Guinea 49,000 to 44,000 Years Ago*. Science 330:78-81.

Sutikna T, Tocheri MW, Morwood MJ, Saptomo EW, Jatmiko, Awe RD, Wasisto S, Westaway KE, Aubert M, Li B, et al. 2016. *Revised stratigraphy and chronology for Homo floresiensis at Liang Bua in Indonesia*. Nature 532:366-369.

Swadling P, Wiessner P, Tumu A. 2008. *Prehistoric stone artefacts from Enga and the implication of links between the highlands, lowlands and islands for early agriculture in Papua New Guinea*. Le Journal de la Société des Océanistes:271-292.

Tassi F, Ghirotto S, Mezzavilla M, Vilaca ST, De Santi L, Barbujani G. 2015. *Early modern human dispersal from Africa: genomic evidence for multiple waves of migration.* Investig Genet 6:13.

Taylor D, Nagle N, Ballantyne KN, van Oorschot RA, Wilcox S, Henry J, Turakulov R, Mitchell RJ. 2012. *An investigation of admixture in an Australian Aboriginal Y-chromosome STR database.* Forensic Sci Int Genet 6:532-538.

Terhorst J, Kamm JA, Song YS. 2017. *Robust and scalable inference of population history from hundreds of unphased whole genomes.* Nat Genet 49:303-309.

Timmermann A, Friedrich T. 2016. *Late Pleistocene climate drivers of early human migration*. Nature 538:92-95.

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. 2007. *Convergent adaptation of human lactase persistence in Africa and Europe*. Nat Genet 39:31-40.

Tobler R, Rohrlach A, Soubrier J, Bover P, Llamas B, Tuke J, Bean N, Abdullah-Highfold A, Agius S, O'Donoghue A, et al. 2017. *Aboriginal mitogenomes reveal 50,000 years of regionalism in Australia.* Nature 544:180-184.

van Oven M, Kayser M. 2009. *Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation.* Hum Mutat 30:E386-394.

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

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. *The sequence of the human genome.* Science 291:1304-1351.

Vernot B, Akey JM. 2015. *Complex history of admixture between modern humans and Neandertals*. Am J Hum Genet 96:448-453.

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

Vernot B, Tucci S, Kelso J, Schraiber JG, Wolf AB, Gittelman RM, Dannemann M, Grote S, McCoy RC, Norton H, et al. 2016. *Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals*. Science 352:235-239.

Veth P, Ward I, Manne T, Ulm S, Ditchfield K, Dortch J, Hook F, Petchey F, Hogg A, Questiaux D, et al. 2017. *Early human occupation of a maritime desert, Barrow Island, North-West Australia*. Quaternary Science Reviews 168:19-29.

Wall JD. 2017. *Inferring Human Demographic Histories of Non-African Populations from Patterns of Allele Sharing*. Am J Hum Genet 100:766-772.

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

Wei W, Ayub Q, Xue Y, Tyler-Smith C. 2013. *A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping.* Forensic Sci Int Genet 7:568-572.

Weissensteiner H, Pacher D, Kloss-Brandstatter A, Forer L, Specht G, Bandelt HJ, Kronenberg F, Salas A, Schonherr S. 2016. *HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing.* Nucleic Acids Res 44:W58-63.

Westaway KE, Louys J, Awe RD, Morwood MJ, Price GJ, Zhao JX, Aubert M, Joannes-Boyau R, Smith TM, Skinner MM, et al. 2017. *An early modern human presence in Sumatra 73,000-63,000 years ago.* Nature 548:322-325.

Willerslev E, Hansen AJ, Binladen J, Brand TB, Gilbert MT, Shapiro B, Bunce M, Wiuf C, Gilichinsky DA, Cooper A. 2003. *Diverse plant and animal genetic records from Holocene and Pleistocene sediments.* Science 300:791-795.

Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nurnberg P, Stoneking M, Kayser M. 2010. *Demographic history of Oceania inferred from genome-wide data*. Curr Biol 20:1983-1992.

Wood JW. 1987. *The Genetic Demography of the Gainj of Papua-New-Guinea. 2. Determinants of Effective Population-Size*. American Naturalist 129:165-187.

Woodroffe CD, Kennedy DM, Hopley D, Rasmussen CE, Smithers SG. 2000. *Holocene reef growth in Torres strait*. Marine Geology 170:331-346.

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, et al. 2003. *The genetic legacy of the Mongols*. Am J Hum Genet 72:717-721.

Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, et al. 2016. *Haplotyping germline and cancer genomes with high-throughput linked-read sequencing*. Nat Biotechnol 34:303-311.