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 ($\sim$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).