APPLICATIONS OF NEXT-GENERATION SEQUENCING

Learning about human population history from ancient and modern genomes

Mark Stoneking and Johannes Krause‡*

Abstract | Genome-wide data, both from SNP arrays and from complete genome sequencing, are becoming increasingly abundant and are now even available from extinct hominins. These data are providing new insights into population history; in particular, when combined with model-based analytical approaches, genome-wide data allow direct testing of hypotheses about population history. For example, genome-wide data from both contemporary populations and extinct hominins strongly support a single dispersal of modern humans from Africa, followed by two archaic admixture events: one with Neanderthals somewhere outside Africa and a second with Denisovans that (so far) has only been detected in New Guinea. These new developments promise to reveal new stories about human population history, without having to resort to storytelling.

HapMap
An international project with the goal of identifying genetic similarities and differences among human populations. The project has made large amounts of data publicly available.

Admixture
Gene flow between two (or more) groups that have been separated for a long enough period of time to be genetically distinct.

**Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D04103 Leipzig, Germany.*
‡Institute for Archaeological Sciences, University of Tübingen, Germany.
Correspondence to M.S.
e-mail: stoneking@eva.mpg.de
doi:10.1038/nrg3029

The year 2010 saw the publication of the first three ancient hominid nuclear genome sequences[1–3], the first results from the 1000 Genomes Project[4], and several other human genome and exome sequences[5,6]. Moreover, genome-wide SNP data are becoming increasingly available from not just the HapMap 'big three' populations (Europeans, Han Chinese and Yoruba) but also many other populations of anthropological interest[7–13]. With the already existing data, and the promise of much more with the increasing ease and plummeting costs of generating such data, these are exciting times for studies of human population history.

The goal of such studies is finding out what happened in our past, in terms of population origins, migrations, relationships, admixture and changes in population size — that is, the demographic history of populations. Prior to the availability of genome-wide data, human population history studies relied largely on single genetic loci, such as mitochondrial DNA (mtDNA) or the non-recombining regions of the Y chromosome (NRY)[14,15]. Although mtDNA and NRY studies have provided a wealth of information, they have limited power to actually infer the basic parameters of demographic history (FIG. 1a) because a single locus provides only a single window into the past. By contrast, genome-wide data provide many independent windows and hence a much more accurate overall view of the past, enabling more detailed demographic inferences (FIG. 1b). Genome-wide SNP data have been available for a few years now and are being increasingly used for demographic studies. However, there are issues related to how SNPs are chosen for inclusion on SNP arrays (see below); these are circumvented by genomic sequence data.

The study of fossil DNA, or ancient DNA, has been revolutionized by technological developments in high-throughput sequencing, making it feasible to move from focusing on single genetic loci, such as mtDNA, to (almost) complete genome sequences of extinct species and populations[16]. These new technologies have been used to generate genome sequences from two of our closest extinct relatives: the Neanderthals[1] at a coverage of 1.3-fold; and a recently discovered extinct hominin group from Siberia, Denisovans, at a coverage of 1.9-fold[3]. In addition, 20-fold coverage of the genome was achieved from a 4,000-year-old hair sample from a native Greenlander (Saqqaq)[2].

Both the ancient genome sequences and the modern genome-wide data have recently provided answers to long-standing questions about the number of dispersals of modern humans from Africa, as well as several other important insights. Here, we review important features of the recent methodological advances, discuss examples that illustrate how genome-wide data have contributed to our understanding of human population history, and consider what more we might expect to uncover in the next few years through using these methods.

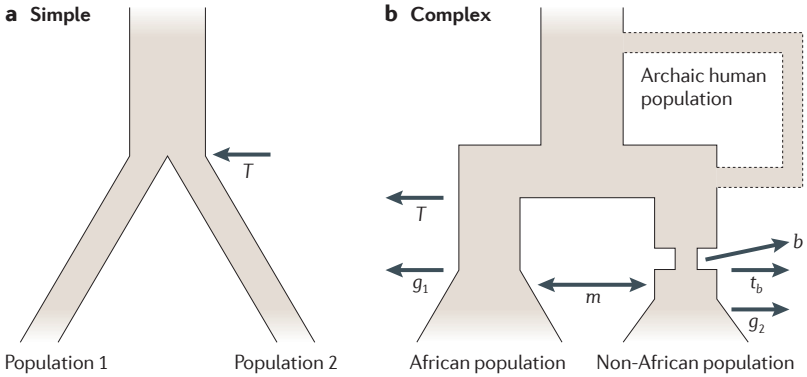


Figure 1 | **Model-based analyses of demographic history.** Demographic analyses based on single or only a few loci can, at best, hope to analyse only very simple models, such as the model shown in **a** of two populations diverging with no subsequent migration or population size changes. With genome-wide data and more sophisticated analyses, much more complex (and hence realistic) models can be analysed (**b**) and many more parameters of interest can be estimated, including divergence time (*T*), migration (*m*), strength of bottlenecks (*b*), time of bottlenecks (t_b), population expansions (*g*) and even admixture from extinct hominins. This figure is illustrative of the complexity of demographic models but does not represent a specific model. Figure is modified, with permission, from REF. 101 © (2009) Oxford University Press.

Demographic history
The history of events that influence the genetic structure of a population, including population size changes, divergence from other populations and migration (gene flow).

SNP arrays
Microarrays that are used to simultaneously genotype several thousand to several hundred thousand SNPs for a single sample.

Hominin
Modern humans, their fossil ancestors, and extinct relatives thereof, up to (but not including) chimpanzees.

Denisovans
Archaic hominins represented by fossil remains from Denisova Cave in southern Siberia; genome sequence data indicate that Denisovans are a sister group to Neanderthals.

Saqqaq
The Saqqaq culture is the archaeological designation of the earliest culture of West and South East Greenland. A 4,000-year-old native Greenlander from the Saqqaq culture, whose hair sample was preserved in permafrost, was used to obtain the first genome sequence from an ancient modern human.

Methods for obtaining genome-scale data

Given the rate at which DNA sequencing costs are dropping[17], in the near future we will have genome sequences from a large number of individuals and populations. However, to date most of the genome-wide population genetic data available for modern humans have come from portions of the genome that have been sequenced by targeted approaches. Below we discuss various methods to obtain ancient as well as modern genome-wide data.

High-throughput sequencing of ancient genomes. Even though the DNA of a deceased organism usually degrades rapidly, some part of it can survive for more than 100,000 years under favourable conditions, such as cold and stable temperatures and a dry environment[18]. Methods such as PCR and bacterial cloning have been used since the early 1980s to amplify and sequence such surviving DNA[19,20]. Such analyses offer great potential to gain insights into the genetic composition of extinct organisms and populations, and can be used to infer phylogenetic relationships, divergence times, population structure, population hybridization and phylogeographic patterns, as well as functional changes and adaptations influenced by the varying environments of extinct and extant organisms. However, ancient DNA is a challenging source of genetic material owing to several factors.

First, it presents a complex mixture of DNA originating from multiple sources, such as endogenous DNA from the target organism itself, microbial and environmental DNA introduced to the fossil during deposition, and DNA contamination introduced after sample collection. Second, ancient DNA is characterized by a short average fragment length, usually below 70 bp[21–23], as well as extremely low quantities of DNA[18]. Third, chemical modifications cause miscoding lesions to the ancient DNA molecules that result in nucleotide misincorporations during amplification and DNA sequencing. Even in the first years of ancient DNA research, criteria were established to ensure the authenticity of results: for example, independent replications[24,25], laboratory practices to avoid contamination of experiments and samples (such as sterile clean rooms, and the use of bleach and ultraviolet light to degrade potential contaminants) and a strict physical separation of modern and ancient DNA work[18]. However, owing to the limited amount of fossil material, traditional PCR approaches — even when performed in multiplex — allow only a maximum of several thousand base pairs of DNA sequence to be obtained from extinct organisms[26]. High-throughput DNA sequencing offers a number of considerable advantages, as well as some important limitations, for sequencing ancient DNA.

The first advantage of a high-throughput approach is that no targeted amplification steps (such as direct PCR) are necessary, as extracted DNA is turned directly into a DNA sequencing library by adding artificial adaptor sequences to both ends of each DNA fragment, thus allowing rapid sequencing-template production (FIG. 2, centre). Sample-specific adaptors can be used to detect potential contamination from other libraries[27]. The artificial adaptors allow all DNA fragments in the library to be amplified in a PCR reaction using adaptor priming sites. Also, aliquots of the amplified library can be reamplified, which provides a renewable source of template for DNA sequencing (FIG. 2). Efficient protocols for preparing sequencing libraries from ancient DNA[28,29] and whole-library amplification therefore allow almost complete ancient genomes to be obtained from less than 50 mg of fossil material[3]. In comparison, almost 1 g of bone was needed to sequence the first few hundred base pairs of Neanderthal mtDNA using previous methods[30].

Second, the vast majority of ancient DNA fragments are too short to be efficiently retrieved by a PCR amplification approach but are readily amenable to high-throughput sequencing. In fact, whereas modern DNA has to be artificially fragmented to produce DNA fragments that are optimal for high-throughput sequencing, the ancient DNA is naturally degraded, so this step can be omitted. Moreover, the short fragment length means that each ancient DNA fragment can be sequenced completely from both ends, which reduces sequencing errors[1]. As any contaminating modern DNA is less likely to consist of short fragments, it is less likely to contribute to high-throughput sequencing results than PCR-based results from the same sample[31].

The direct library sequencing approach also provides further information about potential contamination, as whenever independent fragments (that is, fragments with different start and/or end bases) that overlap a particular DNA position are sequenced, they provide independent internal replication (FIG. 3a). The sequences that can be used in this approach are those that are private to the studied individual or population and in low frequency or absent in potential contaminants[22,32]. For

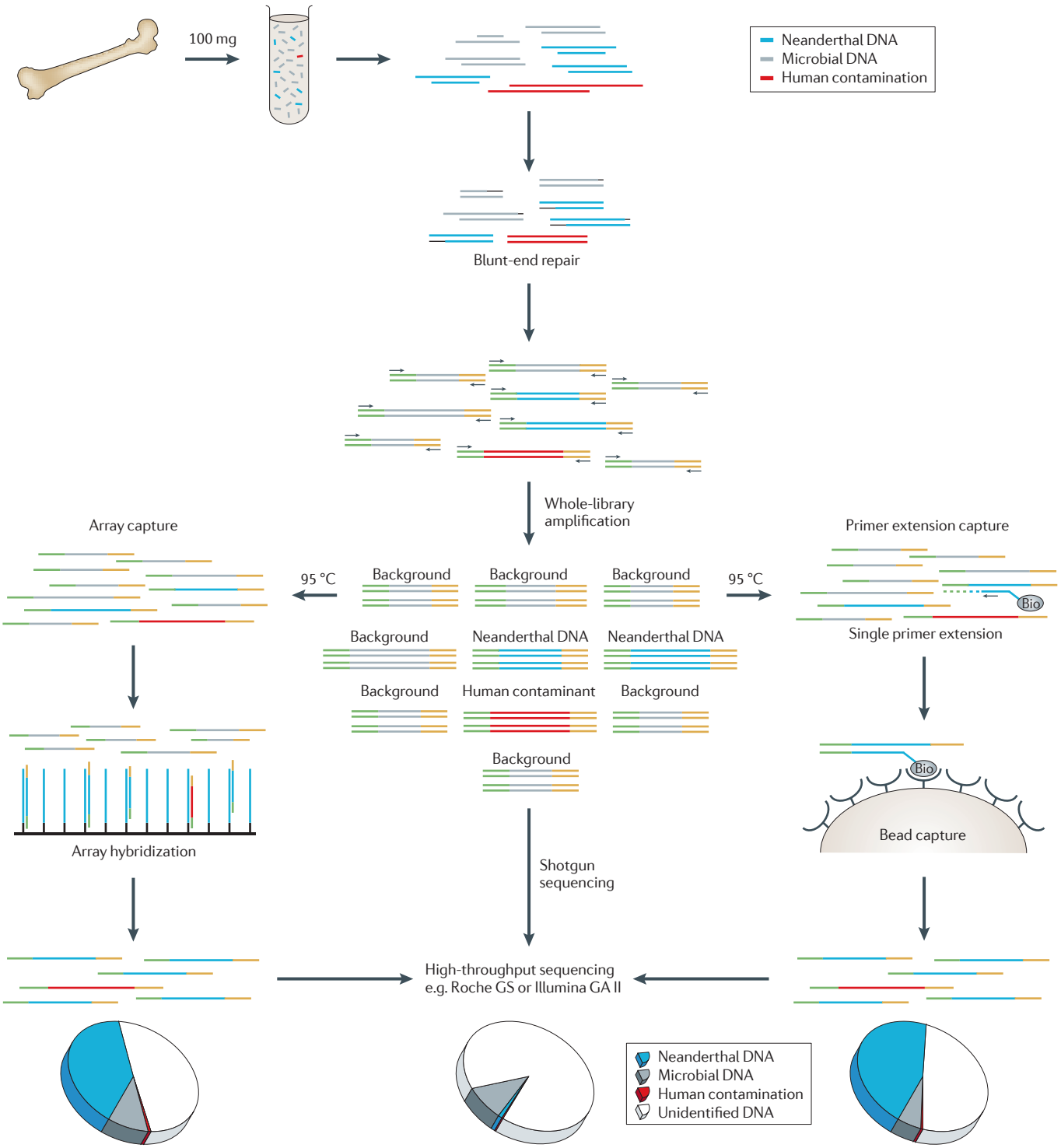


Figure 2 | **High-throughput sequencing of ancient DNA.** A schematic representation of high-throughput sequencing of DNA from fossil remains, here depicted as a Neanderthal bone. The ancient DNA is first blunt-end repaired, and then DNA adaptors are added to each end. The final product, called the sequencing library, serves as the input for various high-throughput sequencing strategies and technologies. All ancient DNA molecules in the library will be first amplified using the adaptors as priming sites in PCR. Aliquots that contain copies of all original ancient DNA molecules can be directly sequenced on a high-throughput sequencer (centre panel) or used in targeted enrichment via array (left panel) or primer extension capture (right panel) methods. The pie charts illustrate the percentage of Neanderthal DNA obtained by each of these approaches, based on data from REF. 45 (array enrichment), REF. 1 (direct shotgun sequencing) and REF. 29 (bead-based enrichment). Figure is modified, with permission, from REF. 102 © (2010) Gesellschaft für Urgeschichte und Förderverein des Urgeschichtlichen Museums Blaubeuren e.V. Bio, biotin; GA, Genome Analyzer; GS, Genome Sequencer.

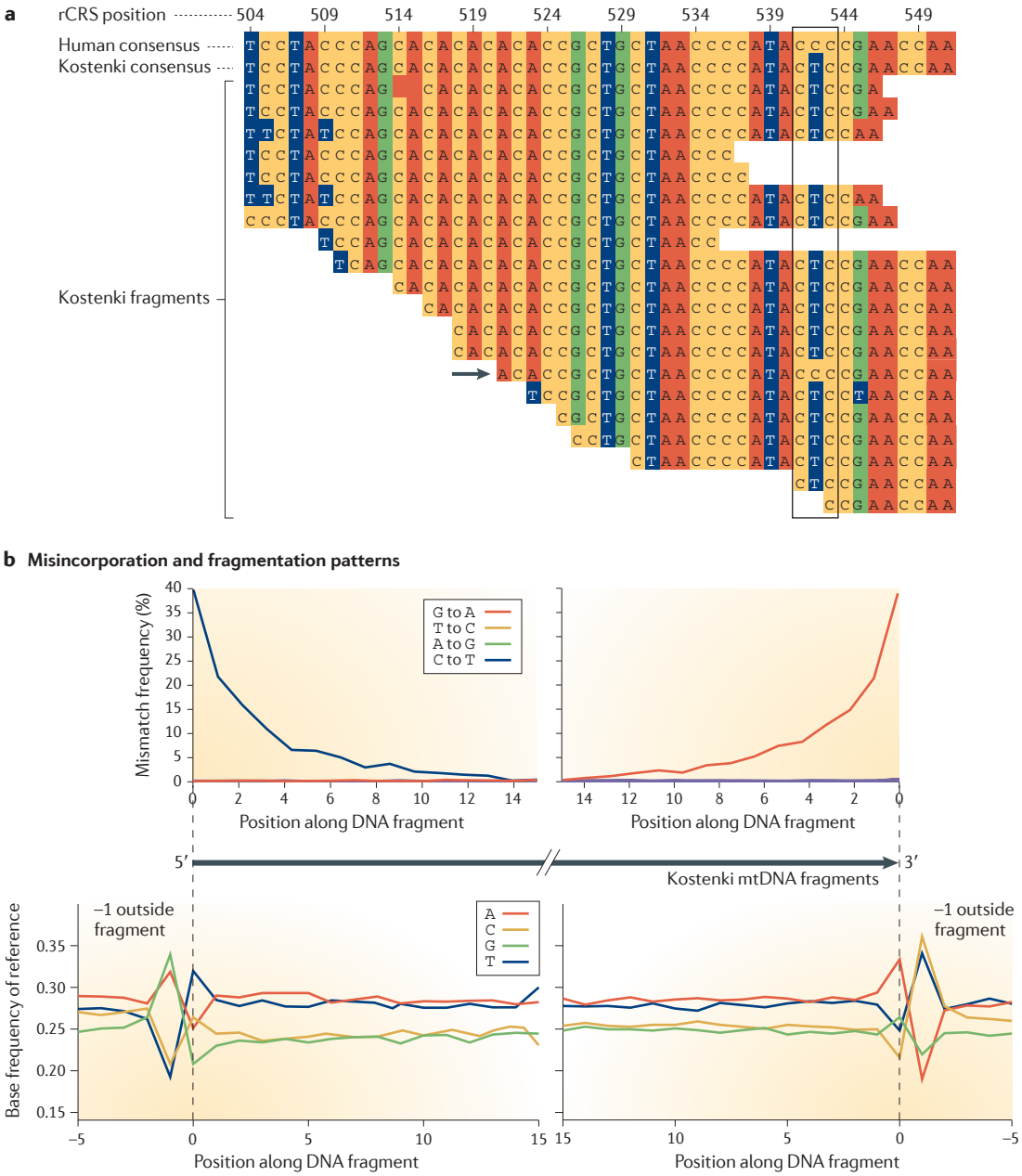


Figure 3 | **Distinguishing ancient from modern DNA. a** | Estimating contamination with modern DNA. Shown is a section of an alignment of the complete mitochondrial DNA (mtDNA) genome (total 16,570 positions) of an early modern human from the Kostenki site, Russia[31]. The positions are based on the revised Cambridge reference sequence (rCRS)[103]. The first line of the alignment shows the consensus sequence obtained from 311 worldwide modern human mtDNAs. The second line shows the consensus sequence for 10,664 mtDNA fragments retrieved from the 30,000-year-old Kostenki early modern human bone. To get an estimate of contamination with modern human DNA, positions were identified where more than 99% of 311 modern human mtDNAs are different from the Kostenki consensus sequence. All fragments that overlap such a position (boxed) and are different from the Kostenki consensus base are likely to be modern human contamination. Only one fragment (indicated by an arrow) is inconsistent, suggesting a very low level of contemporary modern human contamination (1 out of 16 fragments that overlap this position and 1 out of 77 for the complete Kostenki mtDNA data set). **b** | The spatial distribution of DNA degradation patterns that are typical for ancient DNA[27], shown here for the mtDNA fragments from the Kostenki early modern human[31]. The upper panel shows DNA mismatches to a reference sequence for all ancient mtDNA fragments: more than 40% of Cs are seen as Ts at the 5′ end of the mtDNA fragments (left) and more than 40% of Gs are seen as As at the 3′ end (right). The lower panel shows the base frequency of the reference sequence: left, purines (A and G) are in high frequency one base pair upstream of the 5′ end of the start of the mtDNA sequence; right, pyrimidines (C and T) are in high frequency one base pair downstream of the 3′ end of the mtDNA sequence. The presence of such patterns can be used to test the authenticity of ancient modern human DNA[31]. Figure is modified, with permission, from REF. 102 © (2010) Gesellschaft für Urgeschichte und Förderverein des Urgeschichtlichen Museums Blaubeuren e.V.

Endogenous DNA
In the ancient DNA field, endogenous DNA usually refers to the original DNA from the actual organism that was sampled. In some publications, endogenous DNA includes the microbial DNA that is common to most ancient samples versus exogenous DNA that is brought onto or into the sample after excavation.

Nucleotide misincorporations
Erroneous incorporations of nucleotides during the synthesis of the complementary DNA strand by a polymerase (for example, during PCR) that are caused by chemical modifications of the template nucleotides. For example, deamination of cytosine leads to uracil, which is read by the DNA polymerase as thymine and as a consequence instead of a guanine an adenine is incorporated into the complementary strand.

Sequencing library
This consists of DNA samples that have been prepared for high-throughput DNA sequencing by adding artificial oligonucleotides to both ends of the template molecules. The adaptors can be used to bind the DNA to a surface and clonally amplify each molecule before or during high-throughput DNA sequencing.

example, the mtDNA of the Neanderthal differs by at least 133 positions from all modern human mtDNAs; when the Neanderthal genome was sequenced[1], about 27,000 independent mtDNA fragments were determined that overlap at least one of those positions. Less than 0.2% were found to resemble modern human mtDNA, providing a precise estimate of human mtDNA contamination. Similarly, Y chromosomal DNA fragments found in sequences from female samples can be used to quantify the amount of modern male contamination[21]. Although not all regions of the genome can be assayed for contamination, it is possible to extrapolate to the entire genome because contamination would be a mixture of DNA from all genetic loci.

However, it is important to have a large number of informative fragments; the first study of Neanderthal genomic sequence data underestimated the amount of human contamination because this study used only a small number of informative fragments (approximately ten)[21,22,33]. An estimate of the amount of contamination using direct sequencing therefore crucially depends on a sufficient number of individual fragments overlapping informative differences. All three ancient human genomes sequenced to date have contamination estimates of less than 1% derived from large numbers of informative fragments at several genetic loci; therefore it is unlikely that downstream analysis of the data is affected by contamination. In summary, as long as humans handle old remains and do the laboratory work, there will always be a risk of contamination, but the next-generation sequencing methods do provide a greatly improved means of assessing the degree of human contamination.

A limitation of direct sequencing of DNA from ancient remains is that there is enormous variation in the percentage of endogenous DNA that actually stems from the target organism rather than from microbes or some other source. Even within a single excavation site the percentage of endogenous DNA can span more than two orders of magnitude, from close to 100% to less than 0.1%[3,16]. Therefore the amount of sequence information that can be retrieved in a sequencing run varies greatly between samples; for example, the draft woolly mammoth genome was achieved with 30 million reads from remains with >90% endogenous mammoth DNA[16], whereas it took 1.5 billion reads to obtain the draft Neanderthal genome from samples with <5% endogenous DNA[1]. The high-quality results from permafrost remains, such as from woolly mammoths[16] or the hair of the Saqqaq native Greenlander[2], suggest that cold preservation enhances the retrieval of authentic ancient genomic DNA (whether there is any specific advantage in extracting DNA from hair rather than other tissues from permafrost remains needs further investigation).

Furthermore, ancient DNA tends to be affected by post-mortem chemical damage that changes the structure of the DNA molecule and induces nucleotide misincorporations during library preparation and sequencing. The most substantial such damage appears to be cytosine deamination, in which cytosine is converted into uracil and subsequently 'interpreted' as thymine when sequenced[34]. Next-generation sequencing of ancient DNA revealed a specific deamination pattern in which 5′ ends show high rates of C to T changes and 3′ ends show high rates of G to A changes (FIG. 3b). The G to A changes can be attributed to the blunt-ending reaction during the library preparation[27]. In this step, uracils on overhanging 3′ ends will be degraded, whereas uracils on overhanging 5′ ends will serve as a template during the 3′ fill-in step and cause G to A nucleotide misincorporations. The 3′ end therefore presents the reverse complementary pattern to the 5′ end but is caused by the same process of cytosine deamination. The high amount of deamination at the ends of the ancient DNA molecules — where up to 40% of cytosines exhibit the signature of deamination — is in contrast to the finding that only 2% of internal cytosines show such misincorporations (FIG. 3b); this is presumably because the ends are likely to be single-stranded[27,35].

Unless incorporated into the analysis models, such nucleotide misincorporations result in massive amounts of false changes in the DNA sequence of the ancient organism in evolutionary comparisons[1]. One approach to avoid misleading results because of cytosine deamination is to use polymerases that cannot replicate uracil, thereby massively reducing nucleotide misincorporations[2]. However, this approach will also exclude the majority of ancient DNA fragments that contain uracil, which is particularly problematic for samples with very low amounts of endogenous DNA. An alternative approach that avoids loss of precious ancient DNA templates is to repair ancient DNA lesions using uracil DNA glycosylase. This enzyme removes uracil, leaving an abasic site that is subsequently repaired with endonuclease VIII, resulting in a truncated fragment that can be used for library preparation[36].

For some studies, however, DNA misincorporation patterns can be useful for differentiating ancient DNA from modern DNA contamination. For example, modern human DNA contamination in Neanderthal remains, sequenced with high-throughput technologies, showed a more than eightfold reduction in cytosine deamination at the ends of DNA fragments compared to the endogenous Neanderthal DNA[31]. This feature provides a potential means to test the authenticity of DNA sequences derived from the ancient remains of modern humans, which is otherwise extremely problematic as it is nearly impossible to distinguish contaminating human DNA sequences from authentic human DNA sequences from such remains. DNA sequences that stem from a single source (that is, sequences that are specific to this individual or rare in the population) and exhibit typical misincorporation patterns are likely to derive from endogenous ancient human DNA rather than contaminating modern human DNA. However, further studies are needed to investigate the rate at which these nucleotide misincorporation patterns build up over time in order to exclude the misidentification of old contamination — for example, in samples collected in the nineteenth century — as authentic endogenous DNA[37].

Post-mortem chemical damage
Chemical modifications to DNA that happen after the death of the organism: for example, hydrolytic deamination of cytosine.

Cytosine deamination
In the context of ancient DNA, a post-mortem hydrolytic chemical reaction that changes cytosine to uracil, releasing ammonia in the process.

Assembly of ancient genomes. For the two extinct hominin genomes sequenced so far, less than twofold coverage was obtained, preventing *de novo* genome assembly. Therefore, both the Neanderthal and Denisovan genomes were analysed by mapping them to human and chimpanzee genomes and to the *in silico*-reconstructed genome of the last common ancestor of chimpanzees and humans[38].

However, there are important limitations to current approaches to ancient genome assembly owing to the short length of ancient DNA fragments and the repetitive nature of large parts of mammalian genomes (which creates ambiguities in sequence read mapping). For example, short fragments can cause mapping bias, as highly divergent short fragments cannot be accurately mapped to a reference genome. Fragments may also map to different locations in different reference genomes depending on the completeness and accuracy of the reference genomes. For example, to calculate divergence times between an ancient hominin genome sequence, modern humans and chimpanzees, it is important to first verify that the ancient DNA sequences map to orthologous positions in both the human and chimpanzee genomes[1]. These issues mean that even at 20-fold coverage (which was the coverage obtained for the Saqqaq genome) not more than 85% of the genome could be reconstructed[2]; full genome sequences from fossil samples can probably never be achieved with current methods.

Targeted approaches using SNP arrays. As noted above, despite the decreasing cost of sequencing, by far the most common approach for population genetic studies is to use SNP genotyping arrays. These arrays currently allow more than 2 million SNPs to be assayed simultaneously at a cost that makes it feasible to apply this technology to population samples. The main disadvantage of SNP arrays is that only previously described genetic variation is targeted. This has two important consequences: novel variation in a new population of interest will not be detected; and diversity in populations that are closely related to those used to identify the SNPs included on the genotyping array will be overestimated relative to more distantly related populations[39]. This ascertainment bias can be easily seen in estimates of heterozygosity derived from SNP arrays, in which heterozygosity in European populations is overestimated relative to non-European populations[9,10]. If not properly accounted for, ascertainment bias in the SNPs can severely influence the estimation of demographic parameters of interest from the genetic data[39–41]. However, as discussed in more detail below, there are ways to circumvent the ascertainment bias issue, and SNP arrays are currently the most important source of genome-wide data from contemporary populations for studies of human population history[7–13,42,43].

Targeted DNA hybridization capture. Another example of a targeted approach is hybridization capture, which uses synthesized or PCR-amplified biotinylated DNA as a capture device ('bait') that is either bound to a surface or in solution (FIG. 2, right panels). Complementary sequences in the sample DNA bind to the bait, the unbound DNA is washed away, and the enriched target DNA is then eluted and sequenced. This approach has been used successfully to capture and sequence complete exomes or complete mtDNA genomes from a large number of individuals in a parallel manner[28,44]. Moreover, hybridization capture can also be applied to fossil DNA to overcome the limitation of small percentages of endogenous DNA. For example, using microarray capture and synthetic probes, the protein-coding regions that exhibit amino acid differences between humans and chimpanzees were targeted in the Neanderthal genome[45], and complete mtDNA genomes were sequenced from a number of extinct Pleistocene hominins[3,29,46]. In principle, it should be possible to enrich for a complete genome sequence from an ancient hominin fossil, even if the DNA extracted is <1% hominin. Here, the close evolutionary relationship of humans to our nearest living relative, chimpanzees (and, by extension, to all extinct hominins), is advantageous in that using the human and/or chimpanzee genome sequences to design the bait for such enrichment approaches is expected to work for any extinct hominin.

Making inferences about population history

Unsupervised approaches. In addition to the technical advances in producing genome-wide data, increases in computational power have enabled more sophisticated use of such data. There have been two important advances for human population studies. The first involves 'unsupervised analyses', in which the individual, and not the population, is the unit of analysis. Most population genetic analyses are based on populations, which means that the investigator must first classify individuals into populations. Classifying individuals wrongly can thus lead to inaccurate results. Unsupervised analyses avoid the use of pre-defined group labels in the analysis; examples include principle components analysis (PCA)[47] or ancestry components (for example, STRUCTURE or frappe)[48,49]. Such analyses enable detailed description of the genetic structure and admixture history of human populations (BOX 1), but they also lend themselves to 'storytelling', as it is tempting to relate particular ancestry components or PCA results to particular migrations. It is important to remember that PCA or STRUCTURE approaches do not reveal the actual history that produced the observed patterns, and in fact many potential histories would be compatible with the results of such descriptive analyses.

Model-based approaches. To discern underlying population history (FIG. 1), a more promising approach is the use of model-based analyses to directly compare how well different models of population history fit genome-wide data, and to estimate demographic parameters of interest (effective population size, population divergence time, migration rate, and so on) for the best-fitting model[50–52]. The basic idea is to compute various summary statistics from the observed genome-wide data that capture important aspects of the data (such as the site frequency spectrum and genetic distance values between

Ascertainment bias
Sampling bias that arises from how SNPs are chosen for inclusion on SNP arrays; SNPs that are known to be polymorphic in a particular population will overestimate genetic variation in that population relative to other populations.

Hybridization capture
A method that allows selective capture of genomic regions of interest from a complex DNA sample before DNA sequencing. It is based on hybridization between DNA fragments in the sample and chosen 'bait' sequences.

Pleistocene
Geological epoch that spans the time period from about 2.5 million years ago to 12,000 years ago.

Unsupervised analyses
Analyses that are done at the individual instead of the population level and do not require that population labels are applied to individuals.

Ancestry components
A pre-defined number of subgroups with distinctive allele frequencies, inferred from genome-wide data, which are then used to assign the ancestry of each individual without specifying the population to which the individual belongs.

Model-based analyses
Analyses that specify demographic models, investigate which demographic model best fits the genetic data and infer parameters of interest (such as population size changes, divergence times and migration events) for the best-fitting model.

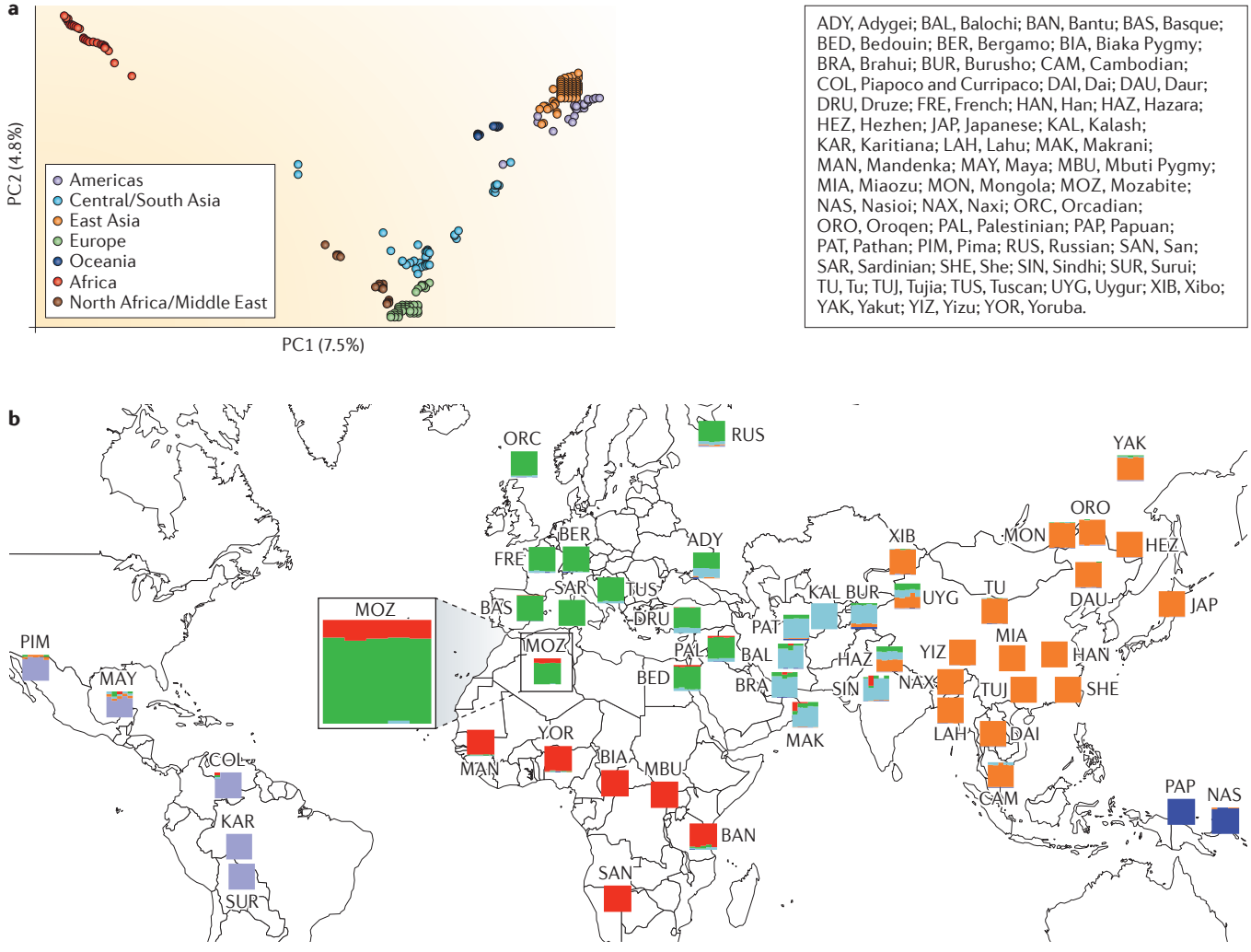
Summary statistics
Statistics that summarize various aspects of genetic data, such as heterozygosity (for within population variation) or F_{ST} values (for between population variation). Summary statistics are conventionally used to investigate the fit of demographic models to the actual genetic data.

Box 1 | **Unsupervised analyses**

Unsupervised analyses are done at the individual and not the population level, and hence avoid any preconceived grouping of individuals into populations that may bias the results. After the analyses are done, group labels can be added to assess how the genetic structure at the individual level corresponds to the groupings.

An example of two such analyses is shown in part **a** of the figure for data on approximately 1 million SNPs genotyped in five individuals from each of the Human Polymorphism Study Centre (CEPH) Human Genetic Diversity Panel populations[10]. The principle components analysis (PCA) identifies independent dimensions that represent most of the information in the data[47], with principal component 1 (PC1) explaining the most variation in the data and subsequent PCs explaining less of the variation. The plot of PC1 versus PC2, in which each label represents an individual, shows one axis extending from Africa to North Africa, Europe and the Middle East and a second axis extending from these latter populations to East Asia, Oceania and the New World. Although this commonly observed pattern is often interpreted as supporting an African origin, note that there is no inherent information about population origins in the plot — one could just as easily suppose that humans originated in East Asia and spread from there, or in Europe or the Middle East and spread bidirectionally.

The second analysis (see part **b** of the figure) is of ancestry components. The analysis assumes that the multilocus genotype of each individual can be explained by one or more ancestry components, where each ancestry component has a characteristic set of allele frequencies at each locus. Carrying out such an analysis involves specifying the number of ancestry components, inferring the allele frequencies for each ancestry component and then assigning the genotypes of individuals probabilistically to one or more components[48,49]. The analysis is again done on individuals, without specifying any group labels; group labels can then be added afterward. Part **b** of the figure shows the apportionment of six ancestry components for the same data[10] as in the PC plot, with each column indicating a single individual's ancestry and the differing colours representing different ancestry components. Note that the presence of multiple ancestry components in a population is often taken as an indication of admixture in the history of that population, but there is nothing in the analysis itself that leads to that conclusion. For example, the North African Mozabites (inset) exhibit both red ('African') and green ('European') ancestry components, which could mean any of the following: Mozabites are of European origin and have experienced African admixture; Mozabites are of African origin and have experienced European admixture; or the ancestral population of Mozabites had both ancestry components and there was never any admixture in their history. There is nothing in the analysis itself that allows one to favour one explanation over another, although that unfortunately has not stopped investigators from fitting the results of such analyses into their favourite stories. The real utility of these unsupervised analyses is in providing descriptive insights into the genetic structure and relationships of populations, for which competing explanations can then be subject to direct testing by model-based approaches (see BOX 2 for an example). Part **b** of the figure is modified from REF. 10.

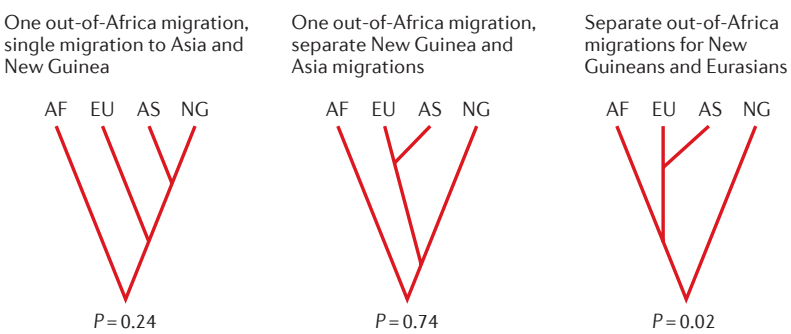


Box 2 | **Testing the 'early southern route' hypothesis**

Based largely on archaeological and fossil evidence, it was proposed that there were multiple dispersals of modern humans from Africa[71]. In particular, the earliest migration of modern humans was hypothesized to have occurred by a southern route along the coast of India that reached Sahul (the combined Australia–New Guinea landmass) around 40,000–50,000 years ago. Rising sea levels and subsequent migrations would then have erased most of the evidence for this early southern route dispersal, except for the preservation of a genetic record in certain populations, including Andamanese Islanders, so-called 'Negrito' groups of South East Asia, and Aboriginal Australians and New Guineans[67]. Indeed, mitochondrial DNA and Y-chromosome evidence has been interpreted as supporting the early southern route dispersal[94–98], although other interpretations are possible[99]. Moreover, genome-wide SNP data have been argued to support a single dispersal of modern humans out of Africa[57], as well as a single wave of migration to East Asia, rather than multiple dispersals[100]. However, this latter study was based on a limited number of SNPs (~50,000) and did not include Australians or New Guineans; also, although some Negrito groups from Malaysia and the Philippines were analysed, these groups have probably admixed heavily with neighbouring non-Negrito groups[100], thereby obscuring their origins.

A recent study[13] of approximately 1 million SNPs in populations from Borneo, New Guinea, Fiji and Polynesia used a novel approach to account for the ascertainment bias associated with the SNPs included on the genotyping platform and then tested several models of human dispersal to determine which model best fits the observed data. The authors compared the summary statistics based on SNPs included on the genotyping platform with the same statistics for full sequence data from ENCODE regions for Yoruba, Chinese and European-Americans from the International HapMap Project. They then used the difference between these statistics to estimate the most likely composition of the discovery sample for ascertaining the SNPs on the genotyping platform. This was then used as a prior in an approximate Bayesian computation (ABC) approach to select the best-fitting model and estimate demographic parameters.

To test the early southern route dispersal hypothesis, three scenarios were evaluated, as shown in the figure (AF, African; EU, European; AS, Asian; NG, New Guinean). The first assumes a single dispersal of modern humans from Africa, followed by a single migration to Asia and New Guinea, and receives moderate support from the data ($P = 0.24$). The second assumes a single dispersal from Africa followed by separate migrations from this ancestral non-African source population, and this scenario receives the strongest support from the data ($P = 0.74$). The third scenario, which corresponds to the original hypothesis of multiple dispersals[71], assumes separate migrations from Africa in the ancestry of New Guineans versus Eurasians. This scenario receives very little support from the data ($P = 0.02$). Thus, dense genome-wide SNP data most strongly support a modified version of the early southern route dispersal hypothesis, in which there was a single migration of modern humans from Africa, followed by separate dispersals from an ancestral non-African population to New Guinea and to East Asia. Encouragingly, this modified southern route dispersal hypothesis is also supported by the signals of gene flow from extinct hominins in modern human genomes[1,3]. First, all studied non-African modern humans have a signal of Neanderthal gene flow, supporting a single dispersal of modern humans from Africa; and second, New Guineans also have a signal of Denisovan gene flow that is not present in other East Asian populations surveyed to date, suggesting a separate migration of the ancestors of New Guineans. Still, the fact that the scenario of a single migration to Asia and New Guinea receives some support from genome-wide SNP data may indicate that this is the correct scenario. Alternatively, it may reflect a subsequent migration (or migrations) from East Asia to New Guinea, although this explanation needs to be tested explicitly. Figure is modified, with permission, from REF. 13 © (2010) Elsevier.



populations) that are expected to be influenced by the demographic history of the population. One then simulates appropriate data under various histories, computes the summary statistics from the simulated data and then identifies the simulated demographic history with summary statistics that are the closest fit to the observed summary statistics. The conclusion is then that the simulated history with the closest fit is the best explanation for the observed data.

Ascertainment bias in SNP array data is an important issue in the selection of summary statistics because it is difficult to adequately account for this bias when simulating genome-wide data. However, approaches for choosing the best-fitting model and inferring demographic parameters have been developed that appear to adequately handle ascertainment bias in SNP array data[11,13], and an example is provided in BOX 2. Model-based approaches have not yet been applied to whole-genome sequence data, but such data should be amenable to this sort of analysis, and would moreover have the advantage of avoiding the ascertainment bias of SNP array data. In principle, even a genome sequence from a single individual could be used to infer the demographic history of that individual's population, as each independent locus in the genome sequence is an independent realization of the population history.

A limitation of current model-based methods is that they rely on summary statistics, which do not make full use of the information in genome-wide data. Some alternative approaches[53,54] make more use of the information in the data; however, these currently require too much computational time to apply to genome-wide data, and therefore further advances in computational power are required before such methods can be applied.

Admixture. Another important aspect of human population history is admixture: that is, genetic contributions of one population to another. Analyses of genome-wide data have made a significant contribution by increasing the recognition of the important role of admixture in human population history (see below). Several approaches now exist for quantifying admixture at either the individual or the population level[11,13,48,49,55–58]. Importantly, these methods are either not influenced by ascertainment bias or can be adjusted to account for ascertainment bias, and thus can be applied to SNP array data. Estimating the time of admixture events is a more difficult problem, especially for old admixture events or for admixture involving closely related populations, but some recent progress has been made[59,60].

What have we learned so far?

Hominin relationships. Genome sequences provide direct estimates of the relationships and divergence times of ancient and modern hominins. In particular, genome sequences indicate that the Neanderthal and Denisovan genomes diverged from modern human genomes on average about 820 thousand years ago (kya), whereas the Neanderthal and Denisovan genomes diverged from each other about 680 kya[1,3]. Thus, Neanderthals and Denisovans are sister groups

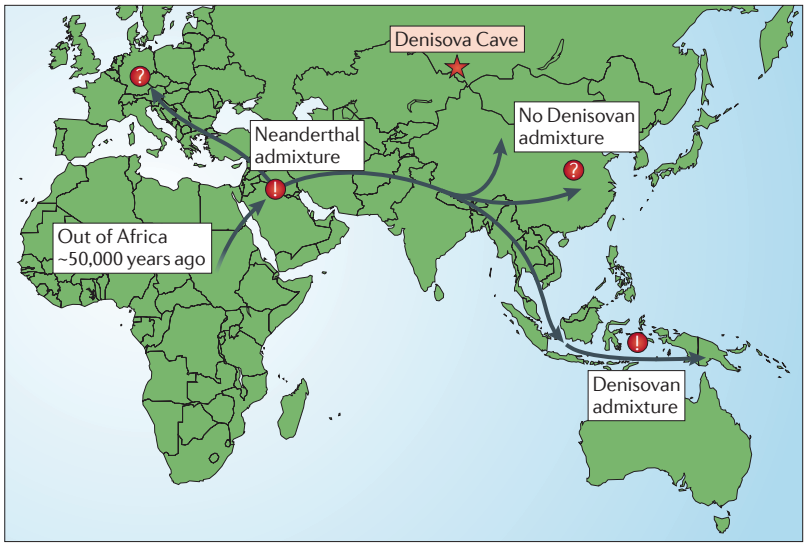


Figure 4 | **Dispersal of modern humans from Africa.** A map illustrating the dispersal of modern humans from Africa about 50,000 years ago, followed by admixture with Neanderthals in the ancestry of all non-Africans, followed by admixture with Denisovans in the ancestry of New Guineans. Arrows indicate general directionality and not specific migration routes — in general we only know for sure the end points of migrations, not the routes. The red star indicates the location of Denisova Cave. The exclamation marks indicate admixture, but there is extreme uncertainty as to where the Neanderthal and Denisovan admixture occurred. Question marks indicate regions where no additional admixture was detected even though archaeological findings suggest that Neanderthals and Denisovans overlapped with modern human populations in those regions.

with a genomic divergence similar to, or slightly larger than, the deepest divergence known among present-day humans, that between San and other groups[3,6]. After Neanderthals and Denisovans diverged they experienced independent histories, which are reflected in their genetic contributions to different present-day human groups. Such information was not apparent from the morphology of the meagre fossil remains attributed to Denisovans. Identifying the source of a particular fossil in the absence of any informative morphology, and even identifying previously unknown hominin groups, as in the case of Denisovans, is likely to be a powerful application of ancient DNA in the future.

It should be stressed that genome sequence divergence times are older than population divergence times because of genetic polymorphism in the ancestral populations[61]. That is, if at the time of population divergence there is polymorphism at a nucleotide site within the ancestral population, then the genetic divergence at that nucleotide site must be older than the population divergence time. However, with some assumptions about population history, genome sequence data can be used to estimate population divergence times, and the resulting estimate for the divergence of the ancestors of Neanderthals and Denisovans from the ancestors of modern humans is about 350 kya[1,3]. This presumably reflects the time when a hominin population left Africa and evolved into Neanderthals and Denisovans, while other hominins in Africa evolved into modern humans.

African origin of modern humans. Single-locus studies of mtDNA and NRY variation in modern human populations have strongly supported a recent African origin of our species, in terms of Africa being the source of the deepest lineages and harbouring the greatest diversity[14,62–65]. Genome-wide SNP data is consistent with this view[7–9], and genome sequences from several modern humans indicate that the deepest population divergences within modern humans are between San individuals from southern Africa and other groups, approximately 115 kya[1,6]. Genetic data indicate that modern humans first dispersed from Africa about 50 kya with divergences among non-African populations dating to 35–50 kya[13,66–68]. One of the most convincing indications of a strong signal of a recent African origin throughout our genome was the demonstration of an astonishingly close correlation between the amount of genetic diversity in a population and the geographic distance of that population from East Africa[69]. This 'serial bottleneck' model strongly implies an African origin of modern humans; in summary, the genetic evidence for an African origin of modern humans is overwhelming.

Dispersal from Africa: replacement or assimilation? Given that modern humans arose recently in Africa and given that other hominins (such as Neanderthals and Denisovans) already existed outside Africa, what happened when modern humans dispersed from Africa and encountered these other hominins? Was there interbreeding, thereby leading to genetic contributions to modern humans from these non-African hominins, or were the non-African hominins replaced without any interbreeding? The extent to which non-African hominins might have contributed to the genomes of modern humans has been one of the long-standing controversies in human evolution[70]. In our opinion this can now be laid to rest, thanks to the Neanderthal and Denisovan genome sequences[1,3]: all non-Africans (and no sub-Saharan Africans) examined to date show about the same amount of gene flow from Neanderthals, with an estimated 2–4% of the genomes of non-Africans coming from Neanderthals. It is also possible to explain this signal of Neanderthal gene flow by a more complicated scenario involving deep population structure within Africa[1]. However, the finding of a signal of gene flow into some modern humans from the Denisova hominin renders this alternative explanation less likely[3], and in our opinion the model that best explains human origins is a recent African origin followed by a small amount of admixture (or assimilation) with non-African hominins (FIG. 4).

Dispersal from Africa: how many times, and which way did they go? The Neanderthal and Denisovan genome sequences give us new insights into human migrations, as the presence (or absence) of the signal of a genetic contribution from a particular extinct hominin can be used as a marker of population relationships. Whether there was a single dispersal or multiple dispersals of modern humans from Africa has been a long-standing question[67,71]. The finding that all modern non-Africans

examined to date exhibit about the same amount of gene flow from Neanderthals[1,3] argues strongly for a single dispersal. Presumably the ancestral non-African population admixed with Neanderthals; where and when this occurred is still uncertain. The current best guess would be somewhere in the Middle East (FIG. 4), where Neanderthals and early modern humans coexisted[72], and some time between 50 kya (when modern humans are estimated to have left Africa) and 35 kya (by which time non-African human populations are estimated to have diverged from one another). A single dispersal of modern humans from Africa is also supported by model-based analyses of genome-wide SNP data[13] (BOX 2).

Another long-standing question about human migrations is whether there was a separate dispersal from Africa of modern humans along a southern route that reached as far as Sahul, followed by a later migration that colonized East Asia, or whether Sahul and East Asia were colonized as part of the same migration[73]. As discussed in BOX 2, the finding that individuals in New Guinea and Bougainville show a signal of gene flow from Denisovans[3] has provided new insights into this issue. However, more sampling of populations in South East Asia and Oceania is needed to fully evaluate the extent of Denisovan gene flow in contemporary human populations.

Other migrations of modern humans. An enduring feature of modern humans is that they have migrated around the world, more so than any other primate species, and genome-wide SNP data are providing insights into the direction, timing and other features of such migrations. For example, the colonization of the Pacific has long been of interest, given the long open-ocean voyages required to reach the far-flung islands of Polynesia. Linguistic and archaeological data have consistently pointed towards the large impact of a recent expansion of a population of Austronesian-speakers from East Asia (probably Taiwan); it is thought that this population may have spread south through the Philippines and Indonesia, and then eastward along the coast of New Guinea and nearby islands in the Bismarck Archipelago, eventually reaching Fiji and then the farthest reaches of the Pacific[74,75]. However, other explanations have been proposed for the spread of Austronesian languages and the origins of Polynesians[76,77]. mtDNA and NRY data have pointed to an admixed origin of Polynesians, with mostly East Asian maternal ancestry and Near Oceanian paternal ancestry[78,79], but this evidence has also been disputed[80]. A recent study of genome-wide SNP data has confirmed the admixed origin of Polynesians, showing about 80% Asian ancestry and 20% Near Oceanian ancestry[80], in line with previous estimates from microsatellite data[81,82]. Moreover, two different approaches have dated the time of Asian–Near Oceanian admixture in the ancestry of Polynesians to about 3 kya[13,60], which is in excellent agreement with the linguistic and archaeological evidence.

As more genome-wide SNP data become available from more populations, the importance of admixture in human population history becomes increasingly evident. This should perhaps not be too surprising given that the two things humans are especially fond of are migration and mating, which then leads to admixture between populations. In addition to the Asian–Near Oceanian admixture in Polynesia, genome-wide SNP data have provided clear evidence of past admixture events in Indian populations[11,83] and in populations of hunter-gatherers from Southern Africa[7]. In fact, as methods are now improving to the point where they can distinguish even subtle signatures of admixture, it is quite likely that in the future it will be found that every human population has experienced some admixture with other groups[57].

Not all migrations were successful in terms of leaving descendants among contemporary populations, and ancient genome sequences can provide evidence of such migrations. One potential example comes from the Saqqaq genome sequence, which appears to indicate a migration to the New World that is distinct from the migrations that led to the origin of all contemporary native North American and Greenland groups[2]. Although confirmation from additional ancient remains of this proposed separate migration to the New World would be desirable, this example nonetheless illustrates the potential of ancient genome analyses for expanding our knowledge of human migrations beyond what can be gleaned from analysing contemporary populations.

What more can we expect to learn?

Here we highlight what are, in our opinion, the most interesting open questions about human population history that could be addressed with genome-wide data, as well as some possible approaches.

Ancient hominin admixture. What remains to be found in the genomes of modern humans in the way of genetic contributions from other extinct hominins (besides Neanderthals and Denisovans)? Answering this question will rely to some extent on serendipity, in terms of finding appropriate fossils with sufficient uncontaminated DNA for analysis. However, at least from a technical standpoint, our ability to retrieve genome sequences from fossils has progressed rapidly, almost to the point where if there is any surviving DNA, it can be sequenced. Moreover, as we learn more about how much of the DNA variation in modern human genomes results from their recent African origin, it may become feasible to identify genomic sequences in modern humans that are more likely to reflect archaic admixture than descent from this recent African origin. Ancient admixture is an extremely useful marker of population relationships, and both the presence and absence of an ancient admixture signal provides information about population history.

Ancient modern humans. It used to be thought that analysis of ancient DNA from remains of modern humans was rather hopeless, given the pervasive nature of contamination with, and inability to distinguish authentic ancient human DNA from, contemporary human DNA. However, as discussed above, high-throughput sequencing offers novel means of assessing the authenticity of ancient DNA, even from modern humans. This could

Sahul
The combined Australia–New Guinea landmass that existed periodically during cold periods in the Pleistocene, including during the initial colonization of Australia and New Guinea about 50,000 years ago, up until rising sea levels separated Australia from New Guinea about 8,000 years ago.

Bougainville
A large island in the Pacific that politically is part of Papua New Guinea but geographically is part of the main Solomon Islands chain.

Austronesian
The most geographically widespread family of languages, extending from Taiwan through mainland and island southeast Asia, Near Oceania, Remote Oceania and even Madagascar.

Near Oceania
Refers to New Guinea and nearby offshore islands, including the main Solomon Islands chain (excluding Santa Cruz); Near Oceania was first colonized by humans at least 40,000 years ago, whereas Remote Oceania (Santa Cruz and all islands to the east) was only colonized by humans beginning about 3,200 years ago.

provide the opportunity to address numerous interesting questions. For example, the extent of Neolithic versus Palaeolithic origins of modern Europeans remains uncertain; there have been differing interpretations of mtDNA sequences with respect to how much of the mtDNA gene pool of contemporary Europeans was contributed by Neolithic farmers migrating from the Near East[84–87]. Genome sequences from pre-Neolithic and early Neolithic skeletons should help to resolve this issue.

Such genome sequences would also help to resolve a puzzling feature of the signal of Neanderthal admixture in contemporary human populations: namely, if modern humans and Neanderthals interbred shortly after the out-of-Africa dispersal and then coexisted for several thousand additional years in Europe, why is there not an increased signal of Neanderthal admixture in Europeans? One potential explanation is that there was indeed such additional Neanderthal admixture with Europeans, but their descendants were then replaced by Neolithic farmers coming from the Middle East. This explanation would then predict higher levels of Neanderthal admixture in pre-Neolithic Europeans. Ancient genome sequences from modern humans would also offer the opportunity to find signatures of past migration events that may not have left descendants in contemporary human populations, as discussed above with respect to the Saqqaq sequence[2]. Ancient DNA analyses of appropriate remains from North and South America would allow the testing of hypotheses concerning single versus multiple waves of migration to the New World[88–92].

More data from contemporary populations. With the large quantities of genome-wide data (both SNPs and sequences) that are now available, it may seem as if all the data are now in hand to address all of the interesting questions about human population history. Nothing could be further from the truth: even the completion of the current goals of the 1000 Genomes Project will not provide sufficient data, as most of the large SNP or sequencing projects have, for logistical reasons, tended to focus on only a few populations from a limited number of geographic regions. Even the most comprehensive resource of human population diversity currently available, the Human Polymorphism Study Centre (CEPH) Human Genetic Diversity Panel[93] with 1,050 individuals from 52 populations, has significant gaps. Thus, sampling, genotyping and sequencing of additional populations will continue to be important. For example, comprehensive analyses of genome-wide SNP (or sequence) data from contemporary populations would contribute to answering the question noted above of single versus multiple waves of migration to the New World.

New methods for inferring history from genome-wide data. Human population history may be about telling stories, but we need better ways to discern what happened in the past without resorting to storytelling. Substantial advances have been made in applying model-based approaches for testing different models of population history and for estimating demographic parameters corresponding to the best-fitting model[52,53]. Also, although it is still challenging, some progress has been made in methods for estimating the timing of old admixture events and investigating admixture involving closely related populations[59,60]. However, there is much more that needs to be done, particularly in fitting more complex (and hence realistic) models.

The good news (especially for students who may fear that all interesting questions have been answered) is that when it comes to human population history, there are still many stories waiting to be told.

1. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
The first genome sequence from an extinct hominin, Neanderthals, demonstrating a signal of Neanderthal admixture in the genome of all studied non-African modern humans.
2. Rasmussen, M. *et al.* Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757–762 (2010).
The first genome sequence from an ancient human. This study demonstrates the feasibility of obtaining high-quality genome sequences from permafrost-preserved human hair and suggests the occurrence of a migration event that is not evident from contemporary human populations.
3. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
The second genome sequence from an extinct hominin, Denisovans, demonstrating that they were a sister group to Neanderthals and that they admixed with the ancestors of Melanesians.
4. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
5. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nature Genet.* **42**, 969–972 (2010).
6. Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).
7. Henn, B. M. *et al.* Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proc. Natl Acad. Sci. USA* **108**, 5154–5162 (2011).
8. Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003 (2008).
9. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100–1104 (2008).
10. Lopez Herraez, D. *et al.* Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS ONE* **4**, e7888 (2009).
11. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
In addition to using genome-wide SNP data to provide a detailed genetic history of India, this study introduced several important methods for analysing such data. These methods were used in subsequent studies to demonstrate an admixture signal in modern humans with Neanderthals and Denisovans.
12. Xing, J. *et al.* Toward a more uniform sampling of human genetic diversity: a survey of worldwide populations by high-density genotyping. *Genomics* **96**, 199–210 (2010).
13. Wollstein, A. *et al.* Demographic history of Oceania inferred from genome-wide data. *Curr. Biol.* **20**, 1983–1992 (2010).
These authors used genome-wide SNP data and a novel approach for accounting for ascertainment bias to infer multiple dispersals of humans to Asia and Oceania, and to investigate the complicated admixture history of Remote Oceanian populations.
14. Jobling, M. A. & Tyler-Smith, C. The human Y chromosome: an evolutionary marker comes of age. *Nature Rev. Genet.* **4**, 598–612 (2003).
15. Pakendorf, B. & Stoneking, M. Mitochondrial DNA and human evolution. *Annu. Rev. Genomics Hum. Genet.* **6**, 165–183 (2005).
16. Miller, W. *et al.* Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**, 387–390 (2008).
17. Lander, E. S. Initial impact of the sequencing of the human genome. *Nature* **470**, 187–197 (2011).
18. Paabo, S. *et al.* Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38**, 645–679 (2004).
19. Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A. & Wilson, A. C. DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312**, 282–284 (1984).
20. Paabo, S. Molecular cloning of Ancient Egyptian mummy DNA. *Nature* **314**, 644–645 (1985).
21. Green, R. E. *et al.* The Neandertal genome and ancient DNA authenticity. *EMBO J.* **28**, 2494–2502 (2009).
22. Green, R. E. *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416–426 (2008).
23. Poinar, H. N. *et al.* Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311**, 392–394 (2006).
24. Cooper, A. & Poinar, H. N. Ancient DNA: do it right or not at all. *Science* **289**, 1139 (2000).
25. Paabo, S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl Acad. Sci. USA* **86**, 1939–1943 (1989).
26. Krause, J. *et al.* Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae. *Nature* **439**, 724–727 (2006).
27. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA* **104**, 14616–14621 (2007).

28. Maricic, T., Whitten, M. & Paabo, S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5**, e14004 (2010).
29. Briggs, A. W. *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**, 318–321 (2009).
30. Krings, M. *et al.* Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19–30 (1997).
31. Krause, J. *et al.* A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.* **20**, 231–236 (2010).
32. Gilbert, M. T. *et al.* Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. *Science* **320**, 1787–1789 (2008).
33. Wall, J. D. & Kim, S. K. Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genet.* **3**, 1862–1866 (2007).
34. Hofreiter, M., Jaenicke, V., Serre, D., Haeseler Av, A. & Paabo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29**, 4793–4799 (2001).
35. Brotherton, P. *et al.* Novel high-resolution characterization of ancient DNA reveals C > U-type base modification events as the sole cause of post mortem miscoding lesions. *Nucleic Acids Res.* **35**, 5717–5728 (2007).
36. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of *in vivo* methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
37. Adler, C. J., Haak, W., Donlon, D. & Cooper, A. Survival and recovery of DNA from ancient teeth and bones. *J. Arch. Sci.* **38**, 956–964 (2011).
38. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
39. Nielsen, R. Population genetic analysis of ascertained SNP data. *Hum. Genomics* **1**, 218–224 (2004).
40. Albrechtsen, A., Nielsen, F. C. & Nielsen, R. Ascertainment biases in SNP chips affect measures of population divergence. *Mol. Biol. Evol.* **27**, 2534–2547 (2010).
41. Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502 (2005).
42. Lao, O. *et al.* Correlation between genetic and geographic structure in Europe. *Curr. Biol.* **18**, 1241–1248 (2008).
43. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
44. Hodges, E. *et al.* Genome-wide *in situ* exon capture for selective resequencing. *Nature Genet.* **39**, 1522–1527 (2007).
45. Burbano, H. A. *et al.* Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* **328**, 723–725 (2010).
46. Krause, J. *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894–897 (2010).
47. Reich, D., Price, A. L. & Patterson, N. Principal component analysis of genetic data. *Nature Genet.* **40**, 491–492 (2008).
48. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000). **This paper introduced the widely used STRUCTURE program for inferring ancestry and admixture from multi-locus data at the individual level rather than the population level.**
49. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. *Genet. Epidemiol.* **28**, 289–301 (2005).
50. Beaumont, M. A. & Rannala, B. The Bayesian revolution in genetics. *Nature Rev. Genet.* **5**, 251–261 (2004).
51. Hey, J. & Machado, C. A. The study of structured populations — new hope for a difficult and divided science. *Nature Rev. Genet.* **4**, 535–543 (2003).
52. Kuhner, M. K. Coalescent genealogy samplers: windows into population history. *Trends Ecol. Evol.* **24**, 86–93 (2009).
53. Hey, J. Isolation with migration models for more than two populations. *Mol. Biol. Evol.* **27**, 905–920 (2010).
54. Hey, J. & Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl Acad. Sci. USA* **104**, 2785–2790 (2007).
55. Bertorelle, G. & Excoffier, L. Inferring admixture proportions from molecular data. *Mol. Biol. Evol.* **15**, 1298–1311 (1998).
56. Bryc, K. *et al.* Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl Acad. Sci. USA* **107**, 786–791 (2010).
57. Hellenthal, G., Auton, A. & Falush, D. Inferring human colonization history using a copying model. *PLoS Genet.* **4**, e1000078 (2008).
58. Moorjani, P. *et al.* The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genet.* **7**, e1001373 (2011).
59. Price, A. L. *et al.* Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.* **5**, e1000519 (2009).
60. Pugach, I., Matveyev, R., Wollstein, A., Kayser, M. & Stoneking, M. Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol.* **12**, R19 (2011).
61. Arbogast, B., Edwards, S., Wakeley, J., Beerlie, P. & Slowinski, J. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annu. Rev. Ecol. Syst.* **33**, 707–740 (2002).
62. Cann, R. L., Stoneking, M. & Wilson, A. C. Mitochondrial DNA and human evolution. *Nature* **325**, 31–36 (1987).
63. Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713 (2000).
64. Underhill, P. A. *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358–361 (2000).
65. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
66. Lohmueller, K. E., Bustamante, C. D. & Clark, A. G. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* **182**, 217–231 (2009).
67. Mellars, P. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* **313**, 796–800 (2006).
68. Schaffner, S. F. *et al.* Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583 (2005).
69. Prugnolle, F., Manica, A. & Balloux, F. Geography predicts neutral genetic diversity of human populations. *Curr. Biol.* **15**, R159–R160 (2005).
70. Stoneking, M. Human origins. The molecular perspective. *EMBO Rep.* **9** (Suppl. 1), 46–50 (2008).
71. Lahr, M. & Foley, R. Multiple dispersals and modern human origins. *Evol. Anthropol.* **3**, 48–60 (1994).
72. Grun, R. *et al.* U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *J. Hum. Evol.* **49**, 316–334 (2005).
73. Lahr, M. M. & Foley, R. Multiple dispersals and modern human origins. *Evol. Anthropol.* **3**, 48–60 (1994).
74. Gray, R. D., Drummond, A. J. & Greenhill, S. J. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science* **323**, 479–483 (2009).
75. Kirch, P. Peopling of the Pacific: a holistic anthropological perspective. *Annu. Rev. Anthropol.* **39**, 131–148 (2010).
76. Donohue, M. & Denham, T. Farming and language in Island Southeast Asia. *Curr. Anthropol.* **51**, 223–256 (2010).
77. Terrell, J. in *Lapita: Ancestors and Descendants* (eds Sheppard, P., Thomas, T. & Summerhayes, G.) 255–269 (Publishing Press Ltd, Auckland, 2009).
78. Kayser, M. The human genetic history of Oceania: near and remote views of dispersal. *Curr. Biol.* **20**, R194–R201 (2010).
79. Kayser, M. *et al.* Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. *Mol. Biol. Evol.* **23**, 2234–2244 (2006).
80. Soares, P. *et al.* Ancient voyaging and Polynesian origins. *Am. J. Hum. Genet.* **88**, 239–247 (2011).
81. Friedlaender, J. S. *et al.* The genetic structure of Pacific Islanders. *PLoS Genet.* **4**, e19 (2008).
82. Kayser, M. *et al.* Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. *Am. J. Hum. Genet.* **82**, 194–198 (2008).
83. Chaubey, G. *et al.* Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. *Mol. Biol. Evol.* **28**, 1013–1024 (2011).
84. Haak, W. *et al.* Ancient DNA from European early neolithic farmers reveals their near eastern affinities. *PLoS Biol.* **8**, e1000536 (2010).
85. Haak, W. *et al.* Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310**, 1016–1018 (2005).
86. Sampietro, M. L. *et al.* Palaeogenetic evidence supports a dual model of Neolithic spreading into Europe. *Proc. Biol. Sci.* **274**, 2161–2167 (2007).
87. Bramanti, B. *et al.* Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**, 137–140 (2009).
88. Fagundes, N. J., Kanitz, R. & Bonatto, S. L. A reevaluation of the Native American mtDNA genome diversity and its bearing on the models of early colonization of Beringia. *PLoS ONE* **3**, e3157 (2008).
89. Hubbe, M., Neves, W. A. & Harvati, K. Testing evolutionary and dispersion scenarios for the settlement of the new world. *PLoS ONE* **5**, e11105 (2010).
90. Kitchen, A., Miyamoto, M. M. & Mulligan, C. J. A three-stage colonization model for the peopling of the Americas. *PLoS ONE* **3**, e1596 (2008).
91. Mulligan, C. J., Kitchen, A. & Miyamoto, M. M. Updated three-stage model for the peopling of the Americas. *PLoS ONE* **3**, e3199 (2008).
92. Ray, N. *et al.* A statistical evaluation of models for the initial settlement of the american continent emphasizes the importance of gene flow with Asia. *Mol. Biol. Evol.* **27**, 337–345 (2010).
93. Cann, H. M. *et al.* A human genome diversity cell line panel. *Science* **296**, 261–262 (2002).
94. Macaulay, V. *et al.* Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308**, 1034–1036 (2005).
95. Thangaraj, K. *et al.* Reconstructing the origin of Andaman Islanders. *Science* **308**, 996 (2005).
96. Gunnarsdottir, E. D., Li, M., Bauchet, M., Finstermeier, K. & Stoneking, M. High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Res.* **21**, 1–11 (2011).
97. Endicott, P. *et al.* The genetic origins of the Andaman Islanders. *Am. J. Hum. Genet.* **72**, 178–184 (2003).
98. Forster, P. Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Phil. Trans. R. Soc. Lond. B* **359**, 255–264 (2004).
99. Cordaux, R. & Stoneking, M. South Asia, the Andamanese, and the genetic evidence for an "early" human dispersal out of Africa. *Am. J. Hum. Genet.* **72**, 1586–1590; author reply 1590–1593 (2003).
100. The HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in Asia. *Science* **326**, 1541–1545 (2009).
101. Wall, J. D., Lohmueller, K. E. & Plagnol, V. Detecting ancient admixture and estimating demographic parameters in multiple human populations. *Mol. Biol. Evol.* **26**, 1823–1827 (2009).
102. Krause, J. From genes to genomes: what is new in ancient DNA? *MGfU* **19**, 11–33 (2010).
103. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genet.* **23**, 147 (1999).

Acknowledgements
We thank A. Briggs, B. Pakendorf and S. Pääbo for helpful comments. The work of the authors is supported by the Max Planck Society (M.S.) and the University of Tübingen, Germany (J.K.).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Mark Stoneking's homepage: http://www.eva.mpg.de/genetics/files/team_stoneking.html
Johannes Krause's homepage: http://www.geo.uni-tuebingen.de/arbeitsgruppen/urgeschichte-und-naturwissenschaftliche-archaeologie/palaeogenetik/mitarbeiter/krause.html
1000 Genomes Project: http://www.1000genomes.org
CEPH Human Genetic Diversity Panel: http://www.cephb.fr/en/hgdp/diversity.php
***Nature Reviews Genetics* series on Applications of next-generation sequencing:** http://www.nature.com/nrg/series/nextgeneration/index.html
ALL LINKS ARE ACTIVE IN THE ONLINE PDF