UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1069

Reconstructing the Human Past using Ancient and Modern Genomes

PONTUS SKOGLUND

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2013

ISSN 1651-6214
ISBN 978-91-554-8744-7
urn:nbn:se:uu:diva-206787

Dissertation presented at Uppsala University to be publicly examined in Zootissalen, Evolutionary Biology Centre, Norbyvägen 18C, Uppsala, Friday, October 18, 2013 at 10:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Skoglund, P. 2013. Reconstructing the Human Past using Ancient and Modern Genomes. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1069. 68 pp. Uppsala. ISBN 978-91-554-8744-7.

The study of DNA variation is one of the most promising avenues for learning about the evolutionary and historical past of humans and other species. However, the difficulty associated with obtaining DNA directly from ancient remains have for long kept genomic studies of population history trapped in time; confined to interpreting patterns of modern-day variation without direct historical observations. In this thesis, I outline new approaches for the retrieval, analysis and interpretation of large-scale genomic data from ancient populations, including solutions to overcome problems associated with limited genome coverage, modern-day contamination, temporal differences between samples, and post-mortem DNA damage. I integrate large-scale genomic data sets from ancient remains with modern-day variation to trace the human past; from traits targeted by natural selection in the early ancestors of anatomically modern humans, to their descendants' interbreeding with archaic populations in Eurasia and the spread of agriculture in Europe and Africa. By first reconstructing the earliest population diversification events of early modern humans using a novel large-scale genomic data set from Khoe-San populations in southern Africa, I devise a new approach to search for genomic patterns of selective sweeps in ancestral populations and report evidence for skeletal development as a major target of selection during the emergence of early modern humans. Comparing publicly available genomes from archaic humans, I further find that the distribution of archaic human ancestry in Eurasia is more complex than previously thought. In the first direct genomic study of population structure in prehistoric populations, I demonstrate that individuals associated with farming- and hunter-gatherer complexes in Neolithic Scandinavia were strongly genetically differentiated, and direct comparisons with modern-day populations as well as other prehistoric individuals from Southern Europe suggest that this structure originated from Northward expansion of Neolithic farming populations. Finally, I develop a bioinformatic approach for removing modern-day contamination from large-scale ancient DNA sequencing data, and use this method to reconstruct the complete mitochondrial genome sequence of a Siberian Neandertal that is affected by substantial modern-day contamination.

Keywords: population genetics, paleogenomics, human evolution

Pontus Skoglund, Uppsala University, Department of Ecology and Genetics, Evolutionary Biology, Norbyväg 18 D, SE-752 36 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-8744-7
urn:nbn:se:uu:diva-206787 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-206787)

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Schlebusch, C. M.*, **Skoglund, P.***, Sjödin, P., Gattepaille, L. M., Li, S., Hernandez, D., Jay, F., Singleton, A., Blum, M. G. B., Soodyall, H., Jakobsson, M. (2012) Genomic variation of seven Khoe-San groups reveals adaptation and complex African history. *Science*, 338(6105): 374–379

II **Skoglund, P.**, Jakobsson, M. (2011) Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences of the United States of America*, 108(45):18301–18306

III **Skoglund, P.**, Malmström, H., Rhagavan, M., Storå, J., Willerslev, E., Gilbert, M. T. P., Götherström, A.*, Jakobsson, M.* (2012) Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe, *Science,* 336(6080):466–469

IV **Skoglund, P.***, Malmström, H.*, Omrak, A., Rhagavan, M., Willerslev, E., Storå, J., Götherström, A., Jakobsson, M. Ancient genomes mirror mode of subsistence rather than geography in prehistoric Europe, *manuscript*

V **Skoglund, P.**, Northoff, B. H., Shunkov, M. V., Derevianko, A., Pääbo, S., Krause, J., Jakobsson, M. Separating endogenous ancient DNA from modern contamination: application to a Siberian Neandertal, *submitted manuscript*

VI **Skoglund, P.**, Sjödin, P., Skoglund, T., Lascoux, M., Jakobsson, M. Investigating population history using temporal genetic differentiation, *manuscript*

*These authors contributed equally to the study.

I am also a co-author of the following articles that were published during my graduate studies.

Skoglund, P., Storå, J., Götherström, A., Jakobsson, M. (2013) Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science*, 40(12):4477–4482

Schlebusch, C.M.*, Sjödin, P.*, **Skoglund, P.***, Jakobsson, M. (2013) Stronger signal of recent positive selection for lactase persistance in Maasai than Europeans. *European Journal of Human Genetics*, 21(5):550–553

Nyström, V., Humphrey, J., **Skoglund, P.**, McKeown, N.J., Vartanyan, S., Shaw, P.W., Lidén, K., Jakobsson, M., Barnes, I., Angerbjörn, A., Lister, A., Dalén, L. (2012) Microsatellite genotyping reveals end-Pleistocene decline in mammoth autosomal genetic variation. *Molecular Ecology*, 21:3391–3402

Malmström, H., Vretemark, M., Tillmar, A., Brandström-Durling, M., **Skoglund, P.**, Gilbert, M.T.P., Willerslev, E., Holmlund, G., Götherström, A. (2012) Finding the founder of Stockholm – a kinship study based on Y-chromosomal, autosomal and mitochondrial DNA, *Annals of Anatomy*, 194:138–145

Skoglund, P., Götherström, A., Jakobsson, M. (2011) Estimation of population divergence times using non-overlapping genomic sequences: examples from dogs and wolves, *Molecular Biology and Evolution*, 28:1505–1517

Castroviejo-Fisher, S.*, **Skoglund, P.***, Valadez, R., Vilá, C., Leonard, J.A. (2011) Vanishing Native American dog lineages, *BMC Evolutionary Biology*, 11:73

Skoglund, P., Höglund, J. (2010) Sequence polymorphism in candidate loci for phenotypic differences in winter plumage between Scottish and Scandinavian populations of Willow grouse (*Lagopus lagopus*), *PLoS ONE* 5(4): e10334

*These authors contributed equally to the study.

Contents

Abbreviations

bp	base pairs
DNA	Deoxyribonucleic acid
kya	Thousand years ago
mtDNA	mitochondrial DNA
LGM	Last Glacial Maximum
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
SE	Standard error
SNP	Single Nucleotide Polymorphism

Introduction

As late as 3 million years ago, it is believed that all ancestors of living humans were found in the African continent. Starting ca ~2.3 million years ago, hominin migrations out of Africa resulted in the appearance of several population lineages in all major continents except the New World and Antarctica, succeeding in adapting to a vast diversity of environments from the frigid Siberian tundra to the lush rainforests of Southeast Asia. Over the course of 2 million years, they evolved into biologically diverse groups such as the ~1 meter tall *Homo floresiensis* (Brown et al. 2004) to the remarkably robust Neandertals (Hublin 2009), and gracile anatomically modern humans (Trinkaus 2005), surviving several glacial periods and other climatic turnover events, and giving rise to the cultural diversity and complex societies of the past 10,000 years. However, there is today scientific consensus that most of these ancient human populations did not give rise to the human populations living outside of Africa today. Instead, at least 90% of the ancestry of all modern humans today can be traced to anatomically modern human populations living in Africa about 100,000 years ago (Meyer et al. 2012). Modern population genetics is becoming increasingly important in the scientific endeavor of reconstructing how this process unfolded, and the documentation of the history of past and living human populations.

The appearance of anatomically modern humans in Africa

The African continent hosts a complex fossil record documenting the genus *Homo*, from its appearance around 2 million years ago to the large diversity in present day Africa (Barham& Mitchell 2008). The history and possible species distinction of various hominin populations documented by the fossil record is highly contentious (Stringer 2002), but frequently cited scenarios are roughly as follows; around 2 million years ago, populations attributed to the taxon *Homo erectus* left Africa to colonize Eurasia. Later, between 800 and 300 thousand years, a second migration out of Africa is likely to have given rise to populations attributed to both non-African *Homo heidelbergensis*, Neandertals, and the group known only from DNA information that is designated as Denisovans (Hublin 2009; Meyer et al. 2012; Reich et al.

2010). However, western representatives of these populations were probably connected to populations living in Africa, and the African fossil record has been interpreted to include taxons such as *Homo ergaster*, *Homo habilis* and *Homo heidelbergensis* (Barham& Mitchell 2008). Likewise, *Homo erectus* populations in Eurasia overlapped in time with Neandertals and other descendants of the putative second wave, and may have contributed to some of the ancestry of Denisovans (Krause et al. 2010b; Reich et al. 2010). Furthermore, possibly as recently as 10,000 years ago another hominin population was present on the island of Flores in Southeast Asia (Brown et al. 2004), and it has been suggested that this population could trace its ancestry either to *Homo erectus*-like populations, or perhaps an even earlier, previously unaccounted for, population related to Australopithecines (Jungers et al. 2009; Tocheri et al. 2007).

Between 300 kya and 150 kya, the first evidence of what is referred to as anatomical modernity began to appear in Africa (Barham& Mitchell 2008; Stringer 2002; Tattersall 2009; Trinkaus 2005; White *et al.* 2003) (*Figure 1*). Anatomical modernity refers to fossil morphology that is largely (but allowing for possible exceptions) within the range of what is observed in living human groups (Stringer 2002; Tattersall 2009). These skeletal remains are found both in eastern and southern Africa, and their absence from other parts of Africa could be attributed to less intense excavations or poor preservation conditions. Thus the geographical point of origin of these anatomical features in Africa is not known, but they markedly predate any such evidence from outside of Africa. Instead, clearly anatomically modern features begin to appear outside of Africa around 40,000-60,000 years ago (40-60 kya), in Europe, Oceania and Asia (Mellars 2006; Trinkaus 2005). Interestingly, such features appear in the Levant about 100 kya (for example in Skhul and Qafzeh), but do not seem to expand further into Eurasia, and are later replaced by populations with clear Neandertal morphology (for example in the Tabun Cave) (Trinkaus 2005). Also the earliest evidence of what is called behavioral modernity is represented by what to present-day eyes may look like a rather unremarkable criss-cross pattern on a rock from Blombos cave in South Africa (Henshilwood *et al.* 2002). However, at 70,000 years ago, this is the oldest securely dated evidence of symbolic culture. The definition of behavioral modernity is contentious, but usually includes representative symbolism and other more complex cultural expressions.

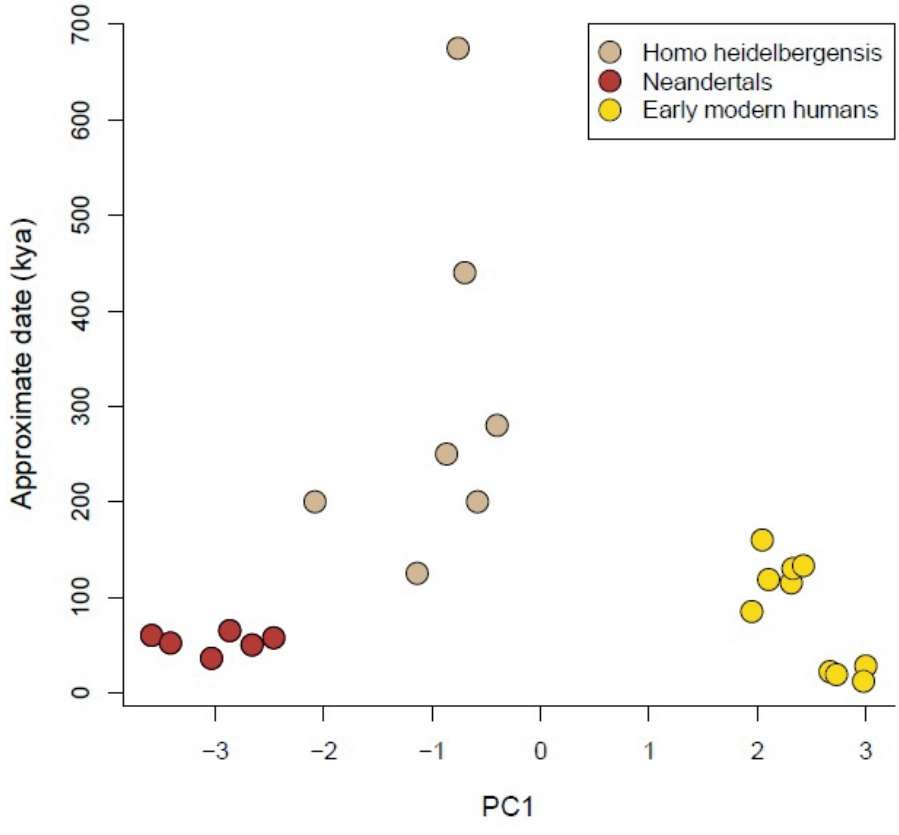

Figure 1. Principal component analysis was performed on 44 skull morphology traits from 23 ancient human remains collected by a previous study (Mounier *et al.* 2011). The first principal component (x-axis) is contrasted with approximate age of the remains (y-axis). The divergence of early modern humans from the putative ancestral population becomes evident from approximately 300,000 years ago.

One of the most heavily debated topics in paleoanthropology was for long the population history behind the appearance of anatomical modernity in Africa and Eurasia (Stringer 2002). The extreme alternative scenarios are sometimes referred to as the out-of-Africa model (or more precisely stated the Recent African Origin model), and the model of multiregional evolution. The out-of-Africa model posits that present-day human populations across the world trace their ancestry to Africa within the past ~100 kya, and thus that the populations with archaic morphology (such as *Homo erectus* and Neandertals) would have been replaced by these newcomers without contributing significantly to their ancestry (Stringer 2002; Stringer& Andrews 1988). The multiregional model however, suggests that present-day human ancestry outside of Africa has its roots in the archaic Eurasian populations, and that anatomical modernity spread as an effect of populations being con-

nected by gene flow, with natural selection acting to raise the frequency of traits associated with anatomical modernity (Wolpoff *et al.* 2000).

Tangled roots: genetic insights into human origins

The first genetic evidence that was taken to support the out-of-Africa hypothesis was from sequencing of mitochondrial DNA variation from worldwide populations (Cann et al. 1987). When a genealogical tree was reconstructed from human mtDNA variation, non-African sequences represent a subset of the variation found in Africa (Cann et al. 1987; Ingman et al. 2000). While this was consistent with the out-of-Africa hypothesis, it did not exclude more complex hypotheses. As has been made apparent by advances in population genetic theory such as Kingman's coalescent (Kingman 1982a; Kingman 1982b; Kingman 1982c), there is great variance in the time at which a sample of genetic lineages find their common ancestor, and the human mitochondrial genealogy only represents a single realized outcome of this highly stochastic process (Nielsen& Beaumont 2009; Rosenberg& Nordborg 2002). However, the first mitochondrial DNA data from Neandertal individuals revealed that they carried lineages that were outside the variation of present-day populations (Green et al. 2008; Krings et al. 2000; Krings et al. 1997; Ovchinnikov et al. 2000) as well as early modern humans (Fu et al. 2013a; Fu et al. 2013b; Krause et al. 2010a; Serre et al. 2004b), which was taken by many as supporting the view that they contributed little to present-day ancestry, while others pointed out that even relatively large contributions could not be excluded based on data from a single locus from a single individual (Nordborg 1998).

The two first genomic studies of Neandertals arrived at different conclusions regarding their genetic relationship between modern humans (Green et al. 2010; Green et al. 2006; Noonan et al. 2006; Wall& Kim 2007), but this was later revealed to be due to contamination from present-day humans affecting one of the two data sets (Green et al. 2009; Wall& Kim 2007). Thus the conclusion from the data that seemed unaffected by contamination was that Neandertals were a distinct population from modern-humans, but that a minor contribution to present-day ancestry could not be excluded (Noonan et al. 2006). Likewise, a remarkably strong correlation between geographic distance from Africa and measures of genetic diversity (Conrad et al. 2006; Jakobsson et al. 2008; Li et al. 2008a; Prugnolle et al. 2005; Ramachandran et al. 2005) seemed to be most parsimoniously explained by a "serial founder model" where populations expanding out of Africa experienced successive founder events and genetic drift (DeGiorgio et al. 2009; Ramachandran et al. 2005). However, decreasing genetic diversity with distance from Africa

could not be taken as rejecting all models with contributions from archaic populations, since it is possible that this pattern could be produced by isolation-by-distance with successively lower effective population size with distance from Africa (Relethford& Harpending 1995).

The most powerful investigation into the question of the relationship between present-day non-Africans and archaic Eurasian populations so far was the analyses of ~1-fold coverage draft genomes of Neandertals and an archaic human individual from the Denisova cave (Green *et al.* 2010; Reich *et al.* 2010). Analyses of the Neandertal genome revealed that while present-day individuals on the whole are substantially more related to each other than Neandertals, the patterns of relatedness to Neandertals were not geographically homogenous. Instead, all non-Africans were significantly closer to Neandertals than Africans, to a degree corresponding to 1-4% Neandertal-related ancestry in all present-day non-African human populations (Green *et al.* 2010). Subsequent analysis of the individual from Denisova cave (Krause *et al.* 2010b) revealed that it was from a population lineage that shared origin with Neandertals (about as distantly related from Neandertals as the most distantly related human groups today), but that Oceanian populations appeared to harbor as much as 5% of their ancestry from Denisovans or a Denisova-related population (Reich *et al.* 2010). Subsequent analyses of higher-quality data from these individuals have refined estimates of total archaic ancestry of ~1.0% in Western Eurasians, ~1.7% in Eastern Eurasians, and up to ~4.7% in Oceanians (Meyer *et al.* 2012), and analyses are ongoing. Thus, while the absolute majority of the ancestry of present-day people in Eurasia, Oceania and the Americas appears to be derived from African populations in the last ~100 kya, a small proportion of genetic material traces its ancestry to the archaic populations that came to Eurasia much earlier. A recent study (Sankararaman *et al.* 2012), used the length of Neandertal-related haplotypes observed in modern-day populations to estimate that the influx of genetic material from Neandertals occurred between 37,000 and 82,000 years ago, possibly coinciding with the potential window of overlap between early modern humans and Neandertals in the Levant.

These ancient DNA studies have also contributed to our view of archaic human populations themselves. The genetic diversity of both Neandertals (Dalén *et al.* 2012) and Denisovans (Meyer *et al.* 2012) appear to be significantly lower than that of any continental human population alive today, suggesting that they were subject to a high degree of genetic drift starting from around half a million years ago (Meyer *et al.* 2012). The substantial genetic divergence of Neandertals and Denisovans also suggest East-West substructure in Eurasia, but the appearance of a highly diverged mtDNA lineage in Denisovans might point to contributions from a more basal archaic population to their ancestry (Krause *et al.* 2010b; Reich *et al.* 2010). Investigations

into the time of population divergence of Neandertals and Denisovan ancestors from present-day African ancestors are complicated by the wide uncertainties surrounding the human autosomal mutation rate, which are necessary to convert estimates of genetic divergence to chronological time. Recent estimates on the mutation rate based on direct assessment of the number of mutations observed between parents and offspring are about half as large as those obtained when calibrating against the human fossil record (Altshuler *et al.* 2010; Kong *et al.* 2012; Scally& Durbin 2012; Scally *et al.* 2012a; Sun *et al.* 2012). These two high- and low estimates of the human mutation rate together with the assumption of instant divergence between Neandertals and present-day West Africans result in population divergence estimates of 170-440 kya, and 410-700 kya, respectively (Meyer *et al.* 2012). The rate of point mutations in the human genome is also central to inferring the history of present-day populations, such as the mode and timing of migration waves out of Africa, to the Americas, and to Oceania (Hawks 2012; Scally& Durbin 2012).

Human genetic variation

At the individual base pair level, a modern human differs from a chimpanzee at approximately 1.1-1.2% of all sites (Mikkelsen et al. 2005; Prado-Martinez et al. 2013) and from a gorilla at about 1.4% of all positions (Scally et al. 2012b). In comparison, a modern human differs from the Denisova or Neandertal genomes at about ~0.15% of all sites (Green et al. 2010; Meyer et al. 2012; Reich et al. 2010), and two modern humans from different populations differ from each other at about 0.08% to 0.12% of all sites (Figure 2). Indeed, the variation in an individual modern human is on the same order of magnitude, around 0.1%, which in this case represents the number of differences between DNA segments inherited from the individual's mother and father. Thus, the variation within most humans is on the same order of magnitude as the differences between two individuals from different populations, but as we shall see this does not prevent statistically powerful inferences about ancestry and population structure (Edwards 2003).

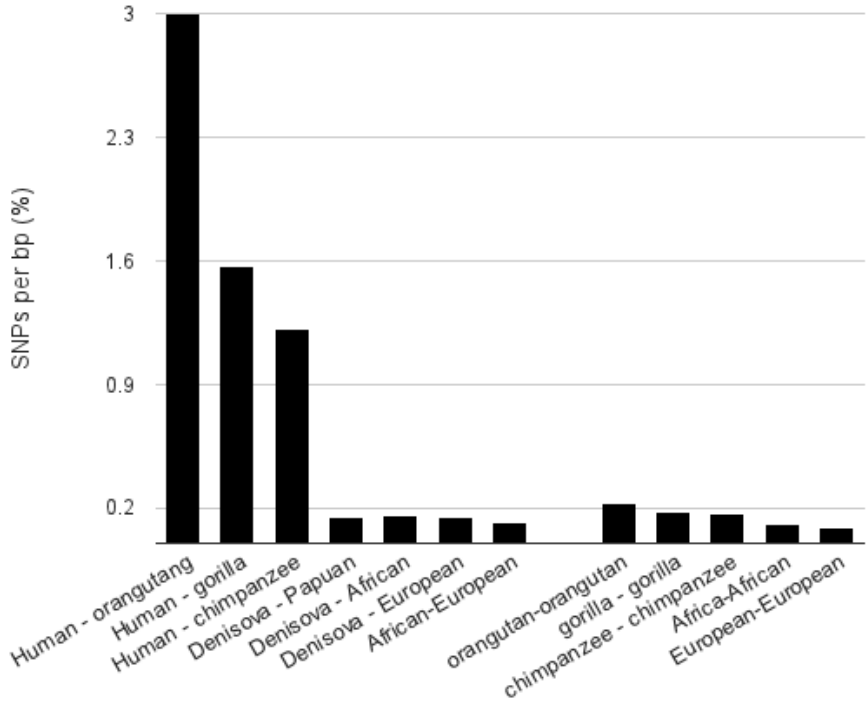

Figure 2. Number of differences observed when comparing sequences from different hominids (Meyer *et al.* 2012; Prado-Martinez *et al.* 2013). On the left side of the histogram are comparisons of sequences from different populations, illustrating the genetic proximity of archaic Denisovans to modern-day humans. On the right are estimates of variation within single individuals from different populations, illustrating that humans are among the least genetically diverse of all hominids (Great apes), and that variation within and between humans of the same population is about as great as between populations.

It is important to distinguish these values of quantitative differences between DNA sequences with commonly cited estimates of shared ancestry between populations. For example, a commonly cited figure is the 1-4% of the ancestry of non-Africans that can be traced to Neandertals (Green *et al.* 2010). This can be thought of as a model where, as we trace the ancestry of different genetic loci in the genome of a non-African individual backwards in time towards the past, 1-4% of these lineages will at some point find themselves in a Neandertal individual. The remaining loci may also coalesce with the Neandertal genome at different time points, but it will always be prior to the divergence of the two populations some 300-800 kya, in the ancestral population of both anatomically modern humans and Neandertals. This value thus conceptually corresponds to *e.g.* the average ~18% of European ancestry found in African Americans (Bryc *et al.* 2010; Parra *et al.* 2001; Tishkoff *et al.* 2009). However, in contrast to recently admixed populations such as African Americans which show great variation in European ancestry (25th–

75th percentiles in one study were 12–28%) (Bryc *et al.* 2010), there is currently no evidence for within-population variation in Neandertal-related ancestry (Green *et al.* 2010), owing to the long time since admixture, which has acted to break up and distribute Neandertal haplotypes evenly in the modern population.

When the variation between humans is more closely investigated, approximately 80-95% of all genetic variation is found between individuals, with the remaining variation being between different subpopulations (Barbujani& Di Benedetto 2001; Lewontin 1972; Li *et al.* 2008b). This variance is represented mainly by small differences in allele frequencies, with some variants being present in certain populations but not others, but so far not a single genetic variant has been found to separate all members of a particular population from any other population. When studying population structure and population history of modern humans, it is thus the small allele frequency differences between populations that make up these 5-20% of the total genetic variance that allows us to make inferences, but it turns out that even just a handful of randomly chosen loci of the large number of polymorphisms in the human genome (estimated to 10-100 million) can provide a great deal of statistical power to infer population structure in human populations (Edwards 2003; Rosenberg *et al.* 2002). Recent studies of large numbers of human individuals are now demonstrating that the total number of human genetic variants has previously been greatly underestimated (Gravel *et al.* 2011; Nelson *et al.* 2012; Tennessen *et al.* 2012). This is due to the recent super-exponential growth of many human populations (Keinan& Clark 2012), which results in many mutations having arisen just within the past 5,000-10,000 years (Fu *et al.* 2013c). Projections suggest that perhaps almost all non-lethal positions in the human genome have a variant present in at least one individual on earth (Nelson *et al.* 2012).

The Neolithic transition in Europe

Managed cultivation of plants and animals arose about 10,000 years ago in the Levant, when hunter-gatherer communities increasingly shifted their lifestyle to a sedentary basis and used the favorable growing conditions in the Fertile Crescent to initiate a transformation of human activities in Western Eurasia from a foraging to food-producing economy (Childe 1925). This historical process is known as the Neolithic transition ('Neolithic' is derived from 'New Stone Age' in Greek). Meanwhile in Europe, the warming climate was also bringing in more suitable foraging conditions for human populations, ushering in a transition from the Palaeolithic ('Old Stone Age') to the Mesolithic ('Middle Stone Age'). These populations could likely trace

their ancestry to the first anatomically modern humans in Europe 45,000 years ago, and had survived the frigid winters of the Last Glacial Maximum (LGM) between 25,000 and 19,000 years ago by retreating to ice-free refugia in Iberia, the Balkans and the Ukrainian steppes, eventually recolonizing the continent at the onset of the Holocene approximately 11,000 years ago (Pinhasi et al. 2012). In addition to being one of the most widely studied periods in archaeology, the central debate of human population genetics in Europe since the inception of the field has circled around the relative impact of the LGM and the Neolithic transition on European genetic variation (Ammerman& Cavalli-Sforza 1984). Specifically, a central question during the last near-century has been two contrasting models for the spread of the Neolithic, via means of "demic diffusion"—where migration of people is the main agent of propagating farming economy and associated material culture expressions across the continent (Ammerman& Cavalli-Sforza 1984; Cavalli-Sforza et al. 1994; Childe 1925; Menozzi et al. 1978) —and "cultural diffusion", which posits a model where non-Neolithic populations take up the new lifestyle and associated practices from neighbors or trade networks (Barker 1985; Champion et al. 2009). However, these two models should not be viewed as mutually exclusive, but rather as two different modes of spread that could have co-occurred. In genetic terms however, they provide testable hypotheses about ancestry and possibly population continuity through time.

The earliest evidence of Neolithic communities in Europe is in Cyprus and in Thessaly in mainland Greece by 8500 BP. The manner of this spread seems to indicate a reliance on maritime transportation, and as the Neolithic culture spread along the Mediterranean route into Western Europe and Iberia, a coastal route also appears to have been the main mode of movement. In contrast, the Neolithic cultures that spread into Central Europe seems to have followed the river valleys of the Danubian and Dnieper. These two paths describe the two main routes of expansion and are associated with the Impressed Ware and Cardial culture in the southwest and the Lineal Pottery complex in Central Europe, and the manner by which their spread was facilitated in terms of population history may have differed considerably (Deguilloux *et al.* 2012; Lacan *et al.* 2012; Lacan *et al.* 2011). After the establishment of Neolithic communities in central Europe, Scandinavia was reached approximately 6,000 years ago. As in the region of origin of the Neolithic in the Near East, the archaeological record documents highly increased population numbers associated with the new food-producing economies in Europe, as well as the presence of ceramics, fortified settlements, grains such as barley and lentils (originating from the Near East) and domestic animals such as cattle, pigs, goats and sheep (Barker 1985; Scarre 2009).

In Scandinavia, the oldest Neolithic material culture expressions are assigned to the Funnel Beaker Culture (TRB; after the German term

Trichtenbecher kultur) (Fischer 2002; Hallgren 2008), which shows clear connection with contemporaneous Neolithic cultures in Central and Western Europe (Malmer 2002; Midgley 2008). However, the foraging lifestyle seems to either persist or reappear in the region, represented by the ceramic-using Pitted Ware Culture (PWC) between 5,300 – 4,000 BP (Malmer 2002). The PWC culture is thus the last hunter-gatherer culture in and around Scandinavia (Malmer 2002), a region which had been inhabited by Mesolithic groups since the ice sheets retreated from Fennoscandia approximately 9,000 years ago. However, further to the North foraging remained an important part of subsistence for considerably longer. Osteological and genetic studies have suggested that individuals associated with the Pitted Ware Culture showed morphological evidence of cold-adaptation that is absent in individuals associated with the TRB (Ahlström 1997a; Ahlström 1997b), nearly lacked the lactase persistence associated allele near the *LCT* gene (Malmström *et al.* 2010) that is also absent in Central European Neolithic farmers associated with the *Linearbandkeramik* culture (LBK) (Burger *et al.* 2007), and showed a high frequency of haplogroup U (Malmström *et al.* 2009)—potentially consistent with continuity from the Mesolithic (Pinhasi *et al.* 2012). Several other DNA studies on genetic sex, contamination and DNA preservation have also been performed on the PWC remains (Götherström *et al.* 2002; Götherström *et al.* 1997; Linderholm *et al.* 2008; Malmström *et al.* 2005; Skoglund *et al.* 2013), and isotopic evidence suggests that individuals associated with TRB and PWC kept distinct dietary patters (Lidén 1996; Lidén *et al.* 2004; Lindqvist& Possnert 1997), focusing on terrestrial and marine resources respectively, even when present side by side on the island of Öland in the Baltic Sea (Eriksson *et al.* 2008), despite evidence of cultural contacts.

Whereas the pioneering analyses of autosomal classical genetic markers by Cavalli-Sforza and colleagues (Ammerman& Cavalli-Sforza 1984; Cavalli-Sforza *et al.* 1994; Menozzi *et al.* 1978) emphasized similarities between the orientation of present-day European genetic differentiation, later studies have demonstrated that these patterns can be caused by a variety of factors (Arenas *et al.* 2013; DeGiorgio& Rosenberg 2013; François *et al.* 2010; Novembre& Stephens 2008). Direct genetic evidence from ancient specimens is almost exclusively from mitochondrial DNA, and indicates that pre-agricultural populations in Europe carried lineages exclusively, or almost exclusively, of a haplogroup (U) (Bramanti *et al.* 2009; Der Sarkissian *et al.* 2013; Fu *et al.* 2013b; Hervella *et al.* 2012; Krause *et al.* 2010a; Malmström *et al.* 2009; Sánchez-Quinto *et al.* 2012) that is present at much lower frequency (<10%) in most European populations today but at higher frequencies in the Baltic region and Northern Scandinavia (20-40%). In sharp contrast to this picture is the mtDNA of early Neolithic farmers in both Central and Northern Europe (Bramanti *et al.* 2009; Deguilloux *et al.* 2011; Haak *et*

al. 2005; Linderholm 2008) and Iberia (Gamba *et al.* 2012; Hervella *et al.* 2012), in which mtDNA haplogroup U lineage is at much lower frequency or absent. In the later farmers and hunter-gatherer populations of the Neolithic period, the difference in mtDNA haplogroup frequencies appears to attenuate, possibly due to admixture (Pinhasi *et al.* 2012).

Holocene history of southern Africa

The warming climate of the Holocene marked a change in the trajectory of human societies not only in Western Eurasia but across the world (Scarre 2009). In southern Africa, the other case of interactions between human populations with different modes of subsistence that I consider here, the onset of the Holocene coincides with the more marked appearance of both cultural expressions and skeletal morphology that are within the range of the modern-day indigenous populations of the region (Stynder et al. 2007a; Stynder et al. 2007b). These populations are traditionally speakers of languages of the Khoisan family, which incorporates distinctive click consonants, but many populations speak languages that are mutually unintelligible. Despite their diverse cultures and identities, anthropologists many times refer to these groups collectively as the Khoe-San (which is the preferred term by the San Council) (Schlebusch 2010). San is usually used to refer to the traditionally hunter-gatherer populations that formerly inhabited most of the region but are now most numerous in Namibia and Botswana, and Khoe refers to the speakers of the KhoeKhoe branch of the Khoe-Kwadi languages who relied on pastoralist economies at the time of colonization. How long the ancestors of the Khoe-San have been living in the region is unknown, but evidence of material culture highly similar to that used by San hunter-gatherers today and in historical times (*e.g.* digging sticks and ostrich shell beads) has been dated to 44,000 years ago (d'Errico et al. 2012). The historical distribution of the ancestors of these populations is unclear (Scheinfeldt et al. 2010). It has been suggested both on linguistic and genetic grounds that they share a more recent genetic relationship with East African populations, possibly related to the introduction of pastoralism (Güldemann 2008; Henn et al. 2008; Pickrell et al. 2013), and it is possible that this connection extends further back in prehistory to the present-day Hadza and Sandawe hunter-gatherer populations (Pickrell et al. 2012; Scheinfeldt et al. 2010).

Early evidence of pastoralism in southern Africa involving sheep and goats date to about 2,000 years ago (Pleurdeau *et al.* 2012; Sadr 1998; Smith 1992) and has been suggested to result from contact with Eastern Africa that may have coincided with the appearance of the Khoe-Kwadi language family (Güldemann 2008). The first adoption of pastoralist practices is thus likely to

predate the arrival of the ancestors of present-day Bantu-speaking populations both in the western and eastern parts of southern Africa. The latter colonization also introduced iron-working societies, and had a tremendous impact on the demographics on the region, with Khoe-San populations now being absent from the eastern parts of south Africa, which both historical sources and evidence such as rock art indicate as being once populated by Khoe-San peoples. The final major event that has been documented for southern African population history is the arrival of European colonists in the 16th century, which also marked the arrival of diseases that indigenous populations had not been previously exposed to.

Early genetic studies identified the Khoe-San as carrying the deepest mitochondrial DNA lineages among all modern human populations (Barbieri *et al.* 2013; Behar *et al.* 2008; Schlebusch *et al.* 2013). Also analyses of microsatellites and small indels (Tishkoff *et al.* 2009) reached similar conclusions of the Khoe-San being relatively distantly related to other African and non-African populations, as did analyses of genome sequence data from half a dozen individuals (Green *et al.* 2010; Gronau *et al.* 2011; Schuster *et al.* 2010). Closer analyses have found support for a model where the Khoe-San diverged from Western Africans at least 100,000 years ago (Green *et al.* 2010; Veeramah *et al.* 2012). However, the genome-wide studies have so far been limited to a small set of populations, and the population history of Khoe-San populations as well as their relationship to other populations remains largely unknown. Central unresolved questions include whether all Khoe-San groups share a common origin, the timing of the earliest diversifications between the Khoe-San populations, their relationship with Khoisan-speakers of East Africa, the genetic impact of the introduction of pastoralism, and the arrival of Bantu-speakers in southern Africa.

Methods

Using DNA information to learn about population history

Population genomic markers

The analysis of population genetic structure is made possible by the presence of polymorphic loci in the genome, which arise by the process of genetic mutation. A polymorphic locus that is used for genetic analysis is often referred to as a genetic marker. Different frequencies of alleles at polymorphic loci provide information about population relationships, diversity, and history. If there was no mutation and hence no genetic markers, there could very well be population structure, but we would have no ability to infer it using genomic data. As population geneticists, we also benefit from the process of recombination, which creates statistical independence between markers. If two loci are statistically independent, they can be seen as independent stochastic outcomes of the same ancestral process, and it is for this reason that our resolution of population genetic structure tends to be greater the more genetic markers we analyze.

Single Nucleotide Polymorphisms

Single Nucleotide Polymorphisms (SNPs) are variants at a single base pair position, and the only indivisible genetic marker. SNPs have been the workhorse of human population structure studies since large-scale genotyping arrays were made available based on the data generated in the HapMap project (Novembre& Ramachandran 2011). Studies using a set of known SNPs which are genotyped in additional individuals provide a powerful yet economical way to obtain large numbers of reasonably independent markers from across the genome, but are unable to discover previously undocumented variants. Since mutations occur rarely in most organisms, it is usually reasonable to assume that there are only two alleles at a single SNP locus (*i.e.* in most cases, only two of the bases A, T, G or C exist at a certain position in the genome).

One of the main ailments of studying known SNPs is what is known as ascertainment bias (Clark *et al.* 2005). This phenomenon arises from the fact

that the allele frequencies of the SNPs discovered in a panel of individuals will tend to be of intermediate frequency in the population(s) to which the individuals belong. Thus, there are two main general effects: we will tend to miss SNPs of low minor allele frequency (low or high derived allele frequency); and the SNPs obtained will be biased toward populations related to the discovery panel (Albrechtsen *et al.* 2010). This leads to distorted allele-frequency spectra, in particular for populations genetically close to the discovery panel, which in turn can cause genetic diversity to be overestimated in these populations. Many approaches have been suggested to correct for ascertainment bias (Albrechtsen *et al.* 2010; Clark *et al.* 2005; Keinan *et al.* 2007; Ramírez-Soriano& Nielsen 2009; Wollstein *et al.* 2010), but this usually requires detailed knowledge of the process that was used to discover the SNPs, a criterion that is not fulfilled for most commercially available SNP-arrays. One clear advantage however, can be gained by analyzing SNPs that have been ascertained (discovered) in a population that is symmetrically related to all study populations*, i.e.* an historical outgroup, in which case no bias is expected with regards to the allele frequency divergence between populations (Patterson *et al.* 2012; Wang& Nielsen 2012).

Haplotypes

Increased resolution for inferring population structure can many times be gained by utilizing the variation created not just by mutation processes but also by the process of recombination (Conrad et al. 2006; Gattepaille& Jakobsson 2012; Jakobsson et al. 2008; Lawson et al. 2012; Loh et al. 2012; Moorjani et al. 2011; Patterson et al. 2012). This can be achieved by analyzing allelic configurations on different locations on a chromosome and taking their correlation structure into account. A prerequisite for these types of analyses is in many cases, but not always (Loh et al. 2012; Moorjani et al. 2011; Patterson et al. 2012), that there is information on the phase of markers on the two chromosomes. This is usually achieved by leveraging population information in a model-based approach (Browning& Browning 2007; Li& Stephens 2003; Scheet& Stephens 2006) but can also be done more directly, for instance by using genetic information from both parents of an individual.

Complete genome sequences

The ongoing rapid advance in sequencing technology is catapulting genomic studies of human population genetic variation to a stage where the near-complete set of SNPs in an individual genome is possible to genotype. The main problem for current high-throughput sequencing technologies—the difficulty of distinguishing between rare biological variants and sequencing error (Nielsen *et al.* 2011)—is perhaps of less concern for studies of popula-

tion structure since the main source of information is the sharing of alleles between individuals. However, it is likely that rare alleles that nevertheless are shared between populations are particularly powerful for inferring fine-scale population structure due to more recent history (Gravel *et al.* 2011; Nelson *et al.* 2012; The 1000 genomes project consortium 2012). Approaches for utilizing this information will likely receive increased attention in the future. Likewise, future methods for inferring population structure may explicitly incorporate data quality measures, for example by propagating mapping and sequencing errors in downstream analyses. Another rising problem is that subtle differences between sequencing platforms and bioinformatic pipelines may result in biases that while small, are on the same order of magnitude as the fine-scale signatures that are being studied (The 1000 genomes project consortium 2012), such as admixture fractions around 1% (Reich *et al.* 2010). In a PCA analysis of low-coverage sequence data by the 1000 genomes project, sequencing technology was found to contribute to a greater proportion of genetic variation than geographical sampling groups (The 1000 genomes project consortium 2012). One solution to standardizing biases between samples is to sequence a pool of samples that have been mixed at equimolar DNA concentrations (Meyer *et al.* 2012), but this would mean that every new study would have to resequence genomes previously analyzed. While higher-coverage data sets are likely to be less sensitive to these issues, data merging biases may be one of the major obstacles in the first generation of complete genome sequencing studies.

Describing population structure and relatedness

Principal component analysis and model-based clustering

Since the early work on theoretical models of population structure pioneered by Sewall Wright in the early 20th century (Wright 1949), the data revolution in population genetics has caused a shift in focus from that of theory towards developing inferential tools and studying empirical data. One of the earliest approaches to analyzing multilocus genetic data, at the time from so-called classical genetic markers such as allozyme variants and blood groups, was principal component analysis (PCA). In a landmark study (Menozzi et al. 1978), principal component analysis of genetic variation was used to study European population structure, revealing that the major axis of genetic differentiation between European populations was oriented from southeast to northwest. PCA can be thought of as a way to distill the highly multidimensional data represented by individuals and their genetic variants into major components of variation, resulting in the identification of axes of variation that reflect relatedness between individuals (Engelhardt& Stephens 2010; McVean 2009; Novembre& Stephens 2008).

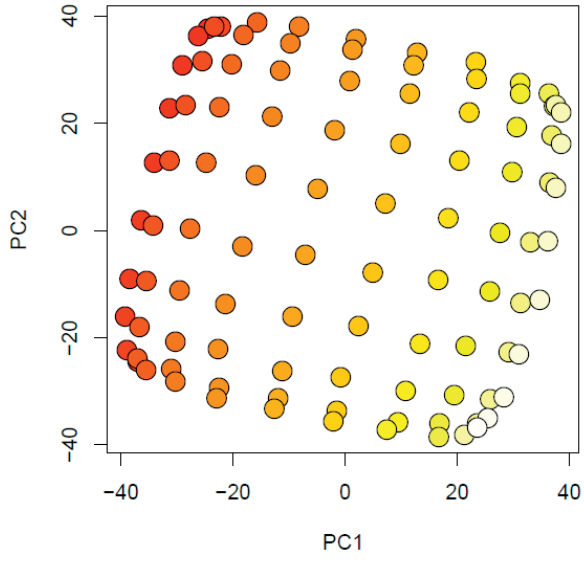

Figure 3. PCA of a two dimensional stepping-stone model of migration where 100 subpopulations in a 10x10 habitat are connected to adjacent populations by migration with parameter $4N_em$=1 (where N_e is the effective population size and m is the number of migrants per generation). We simulated 100,000 independent SNPs using Hudson's *ms* (Hudson 2002) and sampled a single haploid individual from each population.

PCA is attractive due to many features; it is used as a statistical tool in many fields outside population genetics, it can be used to formally test for the presence of population structure using the Tracy-Widom distribution (Patterson *et al.* 2006), and it has been shown that expectations of PCA loadings can be viewed as corresponding to relative coalescent rates between individuals (McVean 2009), much like Wright's commonly used parameter of population structure F_{ST} (Slatkin 1991). If there are two continuous dimensions of population genetic differentiation, such as the two spatial dimensions of a geographic isolation-by-distance model (*Figure 3*), these will be represented by the first two principal components (Engelhardt& Stephens 2010; McVean 2009; Novembre *et al.* 2008; Novembre& Stephens 2008). Another commonly used approach for studying population structure during the past decade has been model-based clustering algorithms (Pritchard *et al.* 2000a). These methods (Alexander *et al.* 2009; Chen *et al.* 2007; Corander& Marttinen 2006; Pritchard *et al.* 2000b) strive to fit multilocus genetic data from multiple individuals to a model where some fraction of each individual's ancestry is assigned to one of two or more hypothesized ancestral populations. These ancestral populations are in the model assumed to have diverged instantly, and are characterized by their predicted genotype frequencies.

The behavior and properties of these methods has been shown to share a common conceptual framework (Engelhardt& Stephens 2010), with different approaches having different properties and problems. For example, model-based clustering approaches are well suited for studies of well-defined populations, but can over-discretize continuous genetic differentiation (Engelhardt& Stephens 2010), and does presently not account for temporal sampling schemes in ancient DNA studies. Likewise, principal component analysis may be less interpretable due to the lack of an underlying model (Engelhardt& Stephens 2010). In general, model based clustering approaches perform relatively well in situations where the historical scenario is similar to the underlying assumptions of well-defined populations, whereas principal component analysis and related methods may be better suited for continuous genetic structure *e.g.* in isolation-by-distance models (Engelhardt& Stephens 2010) (*Figure 3*). For both these approaches, it remains difficult to account for the impact of sampling scheme (DeGiorgio& Rosenberg 2013; Engelhardt& Stephens 2010), and to relate observed population genetic structure directly to historical models, *e.g.* to distinguish equilibrium isolation-by-distance migration models from models of spatial population expansion (François *et al.* 2010; Novembre& Stephens 2008).

Genome-local ancestry inference

At any given segment in the genome, every human individual has inherited one chromosomal copy from each parent. If the two parents were from two differentiated populations, this corresponds to an average genome-wide ancestry of 50% from each parental population. However, in the offspring of admixed individuals, meiotic recombination breaks up the ancestral segments into successively smaller blocks, leading to an increasingly equal distribution of ancestral segments across any chromosome (except the mitochondrial genome and the non-recombining regions of the Y-chromosome). The distribution and characteristics of this mosaic in living individuals has seen a recent surge of interest, motivated by the advent of genome-wide genetic data and the prospects of admixture mapping for medical genetics (Patterson et al. 2004). Many methods have been developed for inferring and delineating stretches of the genome which derive from one of a set of putative ancestral populations (Price et al. 2009; Reich et al. 2005; Sankararaman et al. 2008; Tang et al. 2006). Note that this is relevant only in population samples where there is admixture, since if we imagine a history of two populations that instantaneously diverged without any subsequent gene flow, there may very well be incomplete lineage sorting due to processes in the ancestral population (shared ancestral variation), but no chunks of the genome have at any point in time been found in a different population from which they are now. The concepts underlying genome-local ancestry inference can for example be used to estimate the timing of recent admixture

events (Price et al. 2009; Pugach et al. 2011), since recombination acts to break down the length of contiguous ancestral segments over time. Moreover, this type of genome-local ancestry deconvolution can also be used to reconstruct complete 'virtual genomes' representing the ancestral populations in cases where most living individuals have admixed ancestry (Johnson et al. 2011), to mask segments of recent admixture *e.g.* from historical European colonization in order to be able to study more ancient gene flow events (Reich et al. 2012), or to infer the rate of recombination by studying the frequency of break points between blocks of distinct ancestry across the genome (Hinch et al. 2011; Wegmann et al. 2011). Knowing the distribution of ancestry can also be used to assess whether assortative mating has caused a deficiency in individuals with 'heterozygous' segments of different ancestry (Wall et al. 2011), to assess whether gene flow has occurred in a continuous or punctual manner through time (Gravel 2012; Pool& Nielsen 2009), and to search for signs of adaptive- or maladaptive introgression (where a region with a signature of selection in the admixed population coincides with an atypical proportion of ancestry from one of the ancestral populations compared to the genome-wide average [Paper I]).

Reconstructing models of population history

Formal tests of admixture and f-statistics

Approaches such as PCA and Structure-like clustering methods can in many cases identify recent mixture between populations, but if mixture is more ancient it can be hard to distinguish from other historical models. An alternative way to infer the presence of historical admixture in modern populations (Keinan *et al.* 2007; Patterson *et al.* 2012; Reich *et al.* 2009), is an extension of the *f*-statistics framework that forms the foundations of F_{ST} (Wright 1949). This framework relates the expected co-variances in allele frequencies between not only 2, but also 3 and 4 populations. Reich and colleagues (Reich *et al.* 2009) outlined two formal hypothesis tests of admixture: the *3-population test* and the *4-population test*, which has since been used to identify several cases of admixture in the human past, for example; between Neandertals and the ancestors of non-African human populations (Green *et al.* 2010), between an Eastern Eurasian archaic human population known as the Denisovans and the ancestors of living humans in Oceania (Reich *et al.* 2010) as well as in the history of modern humans in India (Reich *et al.* 2009), the Americas (Reich *et al.* 2012), and Europe (Patterson *et al.* 2012).

The 3-population test is the least model-based of the two and is based on a proposed unrooted history between three populations *A*, *B* and *X*, and rests on the expectation that if group *X* instantaneously diverged from one of *A*

and *B* (or the ancestral population of both *A* and *B*) and has since been genetically isolated, changes in allele frequency in the lineage of *X* due to genetic drift should not be skewed towards the allele frequency changes in the lineage of one of the other two populations (*A*, or *B*). Under this null model, the product (p_X-p_A) x (p_X-p_B), where p_X, p_A, and p_B are the frequencies of one of the two alleles at each locus in each population, is always expected to be positive averaged over all loci. However, if admixture has occurred and caused a skew of the allele frequencies in *X* towards one of the other two populations, the product can be negative. To test if the statistic is significantly positive, standard errors can be estimated by dividing the genome into large chunks in an approach known as the block jackknife (Kunsch 1989). The reason why this procedure cannot be performed for each variable site in the genome is the statistical non-independence between loci that are close enough to each other to prevent free recombination.

While the 3-population test makes no assumptions about the historical order of divergence between the three populations, it may lack statistical power if, for example, genetic drift since divergence has been strong in the target population. The intuition behind this lack of power is the fact that the overall difference in allele frequencies between population *X* and *A*, and between *X* and *B* is so large that it overwhelms any skew towards one of populations *A* or *B* that is due to admixture, leading to a positive test statistic despite of admixture. A more powerful approach, requiring *a priori* assumptions on the history of divergence between populations, is the 4-population test. Here, an unrooted population topology describing the relationship between four populations *A*, *B*, *X* and *Y*, such as (*A*,*B*),(*X*,*Y*), is proposed and the test statistic will assess whether there are any violations to such a simple tree-like history. The reasoning underlying the test is that if the population topology holds (*i.e.* there was divergence without subsequent asymmetrical gene flow), then allele frequency differences between *A* and *B* should be uncorrelated with allele frequency differences between *X* and *Y*. Computing the statistic (p_A-p_B)(p_X-p_Y) averaged across all loci, we can interpret significant deviations from 0 (assessed by block jackknife computation of the standard errors [SEs] as above) as reflecting an asymmetry in allele frequency covariance that is inconsistent with the proposed topology and can only be explained by gene flow. Moreover, the sign of the statistic contains information about the possible directions this admixture could have taken. If gene flow occurred between *A* and *X* (or *B* and *Y*), the statistic is expected to be positive, whereas if gene flow occurred between *A* and *Y* (or *B* and *X*) the statistic is expected to be negative.

A version of the 4-population test that is largely analogous but uses a different normalization is known as the *D*-test (Durand *et al.* 2011). Defined as

$$D(A,B;X,Y) = \frac{\sum_{i=1}^{n}[(p_{iA} - p_{iB})(p_{iX} - p_{iY})]}{\sum_{i=1}^{n}[(p_{iA} + p_{iB} - 2p_{iA}p_{iB})(p_{iX} + p_{iY} - 2p_{iX}p_{iY})]}$$

Where p_{iA} is the frequency of one allele (arbitrarily chosen from the two alleles present) in population *A* at marker *i* and the statistic is summed for all *n* markers. A specific case of this test is when one gene copy is sampled from each of the populations *A*, *B*, *X* and *Y*. Here, one of the populations is usually a distant outgroup so the derived and ancestral states can be inferred. Let the copy from population B provide the ancestral allele, then at SNPs in the genome where X and Y differ, we expect them to be equally likely to share the derived allele with A if the true history is (*A*,(*X*,*Y*)). In other words, we expect to see the configurations (*X*,(*A*,*Y*)) and (*Y*,(*A*,*X*)) an equal number of times, and significant deviations are only expected to arise from gene flow (Green *et al.* 2010). This test is sometimes denoted an 'ABBA-BABA' test, and the great advantage is that differences in genetic drift (for example caused by population size differences) in the different populations can be ignored owing to the fact that we are only using information from a single genetic lineage from each population, and thus no coalescences with other samples in the test can occur. However, a caveat is that for sequencing data, different error rates for different individuals can greatly affect and bias the *D*-test (Rasmussen *et al.* 2011; Reich *et al.* 2010), making it vital to assure that no correlations in allele frequencies between samples have been produced in data generation, alignment and curation.

The 4-population test framework is also highly useful since it can provide a framework to estimate admixture fractions. For example, if *X* in the example above is believed to have received ancestry from population *A* (indicated by a positive test statistic), taking the ratio of two f_4 statistics $f_4(p_O,p_B;p_X,p_Y)$ / $f_4(p_O,p_B;p_A,p_Y)$, where *O* is an outgroup population, can provide an estimate of the *A*-related ancestry fraction in X. This can be thought of as an assessment of how 'far away' from 100% admixture the allele frequency skew in *X* is, with the expected skew for 100% admixture approximated by population *A*.

The main caveat with the 3-population test, the *D*-test, and the f_4-ancestry estimation framework is that they require some prior information on which individuals may constitute a population and about the general population history of these populations. Also, while more long-standing gene flow (such as in isolation-by-distance models) is certainly detectable using 3-population tests and *D*-tests, these tests do not constitute a natural framework to investigate the possible timing or historical frequency of gene flow.

Model-based inference of population topology and admixture graphs

The f-statistic framework allows explicit hypothesis tests of population history, but how can we attempt to fit more complex models of population history to genetic data from a large number of populations? One approach to this is model-based inference of population trees (without mixture) (Cavalli-Sforza& Edwards 1967; Nielsen et al. 1998), which have the underlying assumption that the history of the included populations conform to a simple bifurcating model, where populations instantaneously become isolated and genetic drift is the sole evolutionary force acting to increase differentiation between them. A main problem here is of course that the population history of humans and many other species involve widespread gene flow and admixture after divergence, which if unaccounted for can significantly compromise relationships inferred by tree-based approaches (Lipson et al. 2013).

For this reason, we can attempt to infer graphs instead of trees to represent models of population history. One such approach, *qpgraph* (Reich *et al.* 2009), is based on the *f*-statistics framework (Reich *et al.* 2009) and uses the 4-population test in a more complex model-fitting framework where population topology, admixture events, and genetic drift along lineages are fitted to the observed *f*-statistics, and goodness of fit is assessed by investigating whether *f*-statistics predicted by the model are significantly different from those observed in empirical data. The software implementing this approach has so far not been widely used, but one study applied it to reconstruct a complex model of modern human history in Southeast Asia (Reich *et al.* 2011). The authors found that the model with the fewest predicted 4-population test statistics deviating by more than 3 SEs from the empirical data included two expansions of modern humans into Southeast Asia from an African source, with the first wave mixing with the archaic Denisovans. The approach is hampered by difficulty to conduct formal model testing, since it is hard to determine the number of degrees of freedom. More progress in the fitting of ancestry graphs to genetic data was recently attained in the software *TreeMix* (Pickrell& Pritchard 2012). This method starts by inferring a maximum-likelihood population tree without admixture based on the allele frequency covariance between populations, similar to the methods outlined above. However, the *TreeMix* approach proceeds beyond this by looking at the residuals between the allele frequency covariances predicted by the tree with that of the empirical data. Populations whose allele frequencies significantly deviate from the expected are identified using a block jackknife and "migration edges" (admixture) events are added to the model to attempt to reduce the residual variance. An area of future research might be to identify a number of admixture events beyond which the fit does not improve significantly, for example using likelihood ratio tests, or inspecting the residuals and the percentage of total covariance explained. However, a

remaining topic is the extent to which continuous gene flow between populations affects these methods. The most recent addition to these methods—*MixMapper* (Lipson *et al.* 2013)—rests on the *f*-statistics framework like *qpgraph*, but takes a two-step approach where an unadmixed tree model is first optimized using a fixed set of populations, after which additional populations can be fitted to the scaffold tree with- or without admixture. Future challenges for these methods involve modeling the high dependence on an accurate representation of reference populations, serial admixture events (where the source of gene flow is itself admixed), and how to differentiate and account for ongoing gene flow as opposed to punctual admixture.

Inferring explicit demographic models

Admixture graph approaches attempt to formulate a model that fits the data to a graph of admixture and genetic drift, but how can we translate a certain amount of genetic drift to investigate questions such as how long ago the ancestors of Africans and non-Africans first separated, or how the population size has fluctuated over the history of human populations? To do this, we often need to simultaneously infer how much genetic drift occurs in populations per unit of time and how much genetic drift has occurred since the time of first separation in the out-of-Africa example. Furthermore, in many cases gene flow may have been ongoing for a long time so that a graph model is not an accurate representation of history. In such a case, how can we then estimate the rate of gene flow, and whether this rate has changed through time? To achieve this, we need to have tools to infer more explicit and complex demographic models. Advance in this area actually preceded that of the more simplified (but computationally tractable) admixture graphs. From early single population models, when the focus was often to infer the parameters for the effective population size at different time points (Kuhner et al. 1998), significant advance was made into fitting so called isolation-migration models to genetic data using composite likelihood and Bayesian approaches (Nielsen& Wakeley 2001; Wakeley 1996; Wakeley& Hey 1998). In these models, two (or more) populations are assumed to have originally diverged at some point in the past, after which there may have been continual gene flow between them. Inference of demographic histories such as these have recently been implemented in more generalized statistical frameworks such as the composite likelihood inference using expectations on the site-frequency spectrum obtained using the diffusion approximation to genetic drift (Gutenkunst et al. 2009) and simulation-based approximate Bayesian computation (Beaumont et al. 2002; Pritchard et al. 1999; Tavare et al. 1997), in the latter of which the complexity of the models that can be inves-

tigated is only limited by our ability to simulate them, and identify summary statistics that are informative on their plausibility.

Detecting natural selection in genomic data

One of the most promising avenues for using genomic data in evolutionary research is for inferring the past action of natural selection using the footprints in genetic variation that different types of natural selection tend to leave behind. While many types of natural selection can be studied this way (Nielsen 2005), one of the areas that receives the greatest attention in human genetics is the study of local adaptation in human populations, such as the ability to uptake oxygen in high-altitude populations in Tibet (Simonson et al. 2010; Yi et al. 2010) and Ethiopia (Huerta-Sánchez et al. 2013), and the lactase persistence trait that allows adult individuals in *e.g.* East Africa (Tishkoff et al. 2007) and Europe (Ingram et al. 2009) to digest milk. These traits have been hypothesized to have risen to high frequency due to the action of natural selection, the process by which the genetic material in individuals that produce relatively more offspring tends to increase in frequency in the population (Darwin 1859).

There are two main genetic signatures that allow geneticists to locate the action of local adaptation in the genome (*i*) regions with unusually high differentiation between populations, and (*ii*) regions with depauperate genetic diversity and long monomorphic haplotypes (Sabeti *et al.* 2006). These are both consequences of a phenomenon known as selective sweeps, when one or more advantageous alleles pull along linked variants residing on the same chromosome (Smith& Haigh 1974). This results in long segments of genetic material being relatively monomorphic. At the same time, allele frequencies in the population for which selection is ongoing are differentiating from other populations at a rate that is faster than other parts of the genome (which are mainly subject to genetic drift). One of the most difficult aspects of these types of studies is how to distinguish between the action of natural selection and other factors such as population history. The main approaches taken so far include testing whether some loci show features too extreme to be expected under some model of population history that can either be assumed or fitted separately to genetic data, and/or taking an approach where outlier loci are directly compared to the genomic background to assess whether they show too extreme features to be accounted for by genome-wide processes.

Most presently available methods are geared towards specifically detecting very recent selection, since in this case the genetic signatures of selection may not have had time to dissipate and are thus more readily detectable.

Relatively few population genetic methods have been developed specifically for detecting signatures of selection that occurred at specific time points in the more ancient past, such as during the emergence of the first anatomically modern humans from earlier populations with archaic morphology, or during the time of the first expansions of modern humans in Eurasia, Australia, and the Americas. One approach to detect such selection was performed by the Neandertal genome consortium (Green *et al.* 2010), where they searched for regions in the genomes of 6 modern humans from different populations for regions with unusually high differentiation from Neandertals. These regions could provide clues into which genes underlie phenotypic traits that differentiate modern humans from previous populations. Some of the more interesting regions among the top candidates comprised genes involved in traits such as cognition, but in the top 20 candidates was also a region that comprised the RUNX2 gene. This region garnered quite some interest since mutations in this gene in modern individuals have been linked with several skeletal traits that also separate modern-day humans from Neandertals such as a supraorbital brow ridge, flaring breast cage, and short limbs. However, most regions identified were not as easily interpreted in relation to the archaeological and fossil record, and thus which traits were targeted by selection and what genetic architecture was underlying them in early modern evolution is still largely unknown.

Ancient genomics

Characteristic features of ancient DNA

DNA degradation commences quickly after the death of biological tissues (Pääbo 1989; Pääbo et al. 2004; Willerslev& Cooper 2005), but intact DNA molecules can be preserved in degraded form up to 100,000 years, with extremely favorable depositional conditions allowing survival upwards of a million years (Orlando et al. 2013; Valdiosera et al. 2006; Willerslev et al. 2007). However, several features of ancient DNA complicate both retrieval and downstream analyses. First, the absolute quantity of DNA in ancient tissues and environments is lower than that of fresh tissues. Besides complicating retrieval itself, the most significant consequence of this difference is that small levels of ambient DNA from present-day laboratory environments and reagents can rival the level of endogenous DNA in the fossil, thus contaminating DNA extracts and downstream genetic data. Furthermore, fossil material is invaded by microbial communities post-mortem, usually resulting in abundant presence of DNA sequences that are non-endogenous to the fossil (Noonan et al. 2005). In addition, ancient DNA is without exception severely fragmented, with mean fragment lengths generally below 100 base

pairs even in specimens only a few decades old (Sawyer et al. 2012). This fragmentation process may also show time-dependent biases towards occurring adjacent to purines (Adenine and Guanine) rather than pyrimidines (Cytosine and Thymine) (Briggs et al. 2007). Finally, ancient DNA sequences are subject to post-mortem damage in the form of nucleotide misincorporations, mainly in the form of deamination of cytosine residues into uracils, which is read as thymine during sequencing (Hansen et al. 2001; Hofreiter et al. 2001). As we shall see, these features demand careful precautions to minimize potential contamination and modified laboratory approaches for retrieval, but also substantial modifications and authentication steps to computational and statistical methods used to analyze ancient genetic data.

High-throughput sequencing of ancient DNA

Retrieval of DNA sequences from fossil material has been revolutionized in the past 7 years by the rapid surge of high-throughput sequencing technology (Poinar et al. 2006; Stoneking& Krause 2011). Whereas the previous two decades of ancient DNA research had relied on bacterial cloning or targeted PCR-amplification (Higuchi et al. 1984; Pääbo 1985), the new generation of sequencing technologies now provides an increased throughput of several orders of magnitude, as well allowing efficient sequencing of molecules across their entire length in contrast to specific priming sites. The most common approach so far is direct sequencing, in which case a sequencing library is prepared by adding artificial adapter oligos to extracted DNA and amplifying the ligated molecules without preference for specific loci (Poinar et al. 2006; Stoneking& Krause 2011). Most steps for molecular preparation of sequencing libraries are identical to those used for modern DNA, but common modifications include omitting the DNA shearing step due to the majority of ancient DNA already being fragmented to <200 bp, using polymerases able to amplify through Uracil residues, and a lower amount of amplification cycles to prevent high clonality of the resulting product. Today, DNA extraction and library preparation of ancient material is conducted at labs dedicated to sensitive material. The most important difference in the laboratory is the clean conditions, which include laboratory personnel working in protective suits, positive air pressure, and frequent decontamination of facilities and reagents using denaturing agents such as bleach and UV radiation (Gilbert et al. 2005). These facilities should also be strictly separated from those used for handling present-day DNA and PCR-amplification products (Cooper& Poinar 2000). Since amplified ancient DNA libraries must thus be taken out of the clean room for sequencing, it has also become common practice to add specific indexing oligos ("barcodes") of a few base pairs to each DNA fragment, to be able to identify and exclude contaminating sequences that may enter the sample after it is taken out of the clean

room environment (Green et al. 2010). The direct sequencing approach is now increasingly often being coupled with a pre-sequencing capture step, in order to increase the ratio of endogenous to environmental DNA (Burbano et al. 2010), and future technical developments refining genomic retrieval of ancient DNA are likely to revolutionize the field further in the coming years.

Bioinformatic challenges

The challenges associated with analyzing ancient genetic data stem from the short fragmentation of DNA templates, the increased rate of errors from post-mortem damage, and the pervasive presence of non-endogenous DNA, all complicating steps from initial processing and alignment to genotyping. In addition to the sequence of nucleotides obtained from each DNA fragment, the dominating sequencing technologies also provide a quality score associated with each base on the sequence, which can be calibrated by the inclusion of DNA from a microbial genome of known sequence (PhiX) on each run. These represent the probability that the called base is different from that of the DNA template used for sequencing, but in the case of ancient DNA the template molecule might contain post-mortem damage that differs from the sequence that was present in the individual. Thus DNA damage needs to be explicitly taken into account in downstream analyses. If paired-end sequencing is employed, each DNA fragment is read once from both ends. In applications on modern DNA, this facilitates *e.g.* identification of structural variation since the size of the fragments is often restricted to a certain size. In ancient DNA applications, this instead allows for merging overlapping parts of the two sequence reads into a mini-contig with greater base quality scores (Green et al. 2010). This property also facilitates sequencing the entirety of almost all ancient fragments even with *e.g.* 100 cycles, whereas if single-end sequencing is performed for 100 cycles, fragments with sizes > 100 bp will not be sequenced across their full length.

After sequencing, the characteristic features of ancient DNA require particular consideration in bioinformatic preparation of the sequencing data. Firstly, the short fragment length results in most sequencing reads also including the adapters that were ligated to the DNA template during library preparation. In case of single-read sequencing, these are commonly identified by comparison to the expected adapter sequencing and deleted, sometimes allowing for one or a few mismatches to the original sequence (Rasmussen *et al.* 2010). In case of paired-end sequencing, the adapter sequences are removed by aligning the two sequence-reads of opposite direction, in which case the adapters are outside the overlapping sections. After base calling, merging of paired-end reads and removal of adapters, each of the millions of obtained sequence reads from a high-throughput sequencing run on *e.g.* the Illumina

platform (Bentley *et al.* 2008) must be analyzed to investigate whether they are likely to be endogenous to the specimen (*e.g.* human) and if so, from which genomic location they are from. This is achieved by utilizing alignment programs (or 'mappers') that attempt to balance accuracy with the computational speed needed to keep up with the speed of data generation (Li& Homer 2010). The short fragment length and presence of post-mortem damage present significant challenges compared to sequence data from high-quality DNA sources (Green *et al.* 2010; Schubert *et al.* 2012). Importantly, the high rate of nucleotide misincorporations at the fragment termini would lead to many sequence reads being discarded using standard alignment parameters adopted for modern DNA. To circumvent this, alignment needs to be performed at great sensitivity and without a "seed region", a region in the beginning of each sequence at which none or only a few mismatches are allowed, and with lower penalties for mismatches across the entire sequence than usual. In general, the range of endogenous DNA from fossil bones ranges from <1% from temperate environments to around 1-5% from colder environments and up to 80% in permafrost environments. While these are typical numbers, preservation seems to be highly specific to depositional conditions even within the same locality, best exemplified by the Denisova Cave in Siberia, where samples from archaic Denisovans differ by two orders of magnitude in their fraction of endogenous DNA (Reich *et al.* 2010). A promising avenue for the future of ancient DNA is thus targeted capture, which is possible to do on a large scale for shorter regions such as mtDNA genomes (Briggs *et al.* 2009), but more difficult to implement for the large amount of loci required to make worthwhile genomic studies (Burbano *et al.* 2010; Fu *et al.* 2013a).

Authenticating ancient genomic data

Despite extensive precautions during sample preparation for sequencing, contamination is present in many archaeological specimens (Malmström et al. 2005; Green et al. 2006). Since contamination cannot be ruled out on the basis of laboratory procedures alone, it is essential that authenticity is established for the genetic data itself. For rare or extinct organisms, contamination from present-day sources is implausible. For example, the first mtDNA sequence isolated from a Neandertal was different from any mtDNA sequence obtained from modern humans, suggesting that it was unlikely to represent contamination. However, this type of argument is not applicable for example when studying ancient DNA from anatomically modern humans from Europe, or ancient DNA studies of the domesticated animals whose DNA frequently contaminates common reagents (Leonard et al. 2007). In addition, contamination of present-day nuclear DNA in *e.g.* Neandertals is much more difficult to detect, and can be misidentified as representing gene flow (Wall& Kim 2007). While fragmentation appears to occur also in contami-

nating sequences of a few decades, substantial post-mortem nucleotide misincorporations appear to be a distinguishing feature for ancient DNA, suggesting that it can be used to establish that retrieved sequences are endogenous (Krause et al. 2010a; Sawyer et al. 2012). However, present-day contamination can still be substantial despite a clear signal of post-mortem damage. For high-coverage data, a useful approach is to follow up the observation of post-mortem damage by investigating whether there is evidence of genetic material from more than one individual in the data (Green et al. 2010; Krause et al. 2010a), but this is generally difficult for low-coverage data sets. Also, this approach only allows investigation of whether contamination is present, and cannot reduce the level of contamination. This has restricted large-scale ancient DNA studies to specimens with very low levels of contamination.

Challenges for population genetic analyses

The power of the direct sequencing approach is that it enables the quick generation of relatively large amounts of genomic data directly from fossils. However, unless exceptionally well-preserved specimens are available (Rasmussen et al. 2010), moderate sequencing efforts will only result in low genomic coverage. This makes comparisons between individuals sequenced in this way very difficult, as each individual will be represented by its own random set of genomic loci. Therefore, ancient genomic studies have so far focused on single individuals (Blow et al. 2008; Green et al. 2010; Keller et al. 2012; Miller et al. 2008; Rasmussen et al. 2010; Reich et al. 2010; Skoglund et al. 2011) or on a handful of individuals with highly similar ancestry (Green et al. 2010; Sánchez-Quinto et al. 2012). Another complication arises from the high-frequency (up to 40% at sequence termini) of post-mortem damage in ancient DNA sequences (Briggs et al. 2007). Experimental solutions to this problem includes using uracil-DNA glycosylase to repair most damaged residues (Briggs et al. 2010), but this also reduces the use of damage for authentication purposes. In practice, transition SNPs (C/T and G/A) are many times excluded from population genetic analyses, but a venue for future research may be to deal with post-mortem damage probabilistically during genotyping (Jónsson et al. 2013; Rasmussen et al. 2010). For low coverage studies, errors arising from the sequencing reaction itself are also a concern, with the most common solution being to impose strict quality filters, and to restrict population genetic analyses to loci that are already known to be polymorphic. Finally, temporal differences between ancient samples, and the ancient samples and modern-day populations complicate interpretations of population genetic structure. Even in the absence of population structure, genetic drift is expected to produce genetic differences between genetic data from different points in time (Depaulis et al. 2009; Nord-

borg 1998; Nyström et al. 2012), which in practice makes separating historical scenarios of replacement and genetic drift difficult (Castroviejo-Fisher et al. 2011; Malmström et al. 2009; Nordborg 1998; Serre et al. 2004a). The higher error rate resulting from postmortem damage can also bias population genetic analyses by inflating the number of rare alleles, which in addition to biasing estimates of genetic diversity and drift through time (Axelsson et al. 2008) can make ancient individuals appear more closely related to genetically distant individuals (Rasmussen et al. 2011).

Research Aims

The main objective of this thesis was to use a wide array of genome-wide data from understudied ancient and modern populations to improve our understanding of human population history and evolution. To do this, a central goal was to develop new approaches for moving the field of ancient DNA beyond single-locus markers into the era of genomics. Specific aims were to:

I Obtain a complete view of human genetic variation by the inclusion and analysis of previously underrepresented southern African groups.

II Reconstruct models of population history in Africa to learn about early human prehistory.

III Develop methods for identifying possible targets of adaptation in the human genome during the emergence of early modern humans.

IV Investigate the distribution of archaic human ancestry in Eurasia, Oceania and the Americas using previously published ancient genomes from Neandertals and Denisovans.

V Directly test hypotheses about population structure during the Neolithic transition in Europe using ancient genome sequencing, and investigate the impact on living populations.

VI Develop statistical methods of authentication and contamination removal for large-scale sequencing studies using ancient DNA.

VII Investigate the theoretical properties of temporal sampling, and its use for learning about population history.

Results and Discussion

The origin and spread of anatomically modern humans (Papers I and II)

One of the main arguments that have been taken as evidence for the out-of-Africa model of human origins is an observed decline in allelic and haplotypic diversity with distance from Africa (Jakobsson et al. 2008; Ramachandran et al. 2005). Based on similar patterns, it has also been suggested that such a gradient exists in Africa (Henn et al. 2011), supporting an origin of modern humans in southern Africa. This hypothesis has seemed attractive since southern African hunter-gatherers, speakers of Khoisan click-languages traditionally referred to as Khoe-San by outsiders, have appeared in preliminary studies to represent the now-living human population that is the most diverged from the others (Tishkoff et al. 2009). To address questions surrounding human origins in Africa and African population history, we genotyped more than 2 million SNPs in 220 southern African individuals from 7 Khoe-San groups, and combined this data with individuals from a large set of other African populations (Paper I). Since more recent colonization of southern Africa from Europe, Asia, and Central Africa has greatly impacted the ancestry of the indigenous African populations, we first identified individuals of atypical ancestry compared to others of the same group, and excluded these from many analyses in order to better capture more ancient population history.

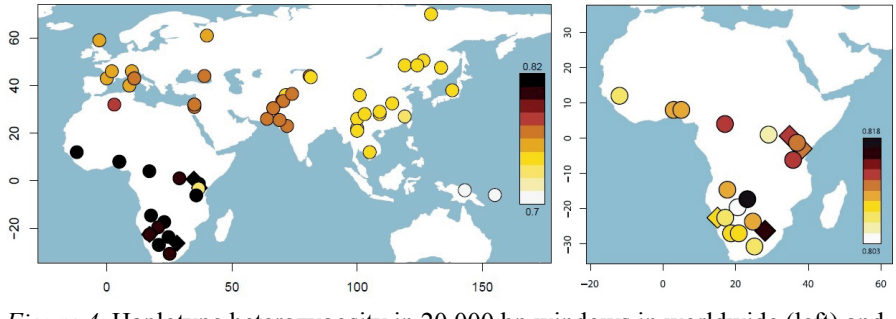

Figure 4. Haplotype heterozygosity in 20,000 bp windows in worldwide (left) and African populations (right). Whereas there is a strong worldwide pattern of decreasing heterozygosity with distance from Africa worldwide, no such clear pattern exists within Africa.

Analysing patterns of haplotype heterozygosity, haplotype richness, linkage disequilibrium, and runs of homozygosity, we found no clear geographical pattern or cline in genetic diversity within Africa (*Figure 4*). Instead, we find that genetic diversity even within groups which self-identify as Khoe-San range from among the lowest to among the greatest observed in Africa. In contrast, these summary statistics of genetic diversity—of which genome-wide estimates of haplotype heterozygosity and haplotype richness have not previously been obtained for worldwide populations—all show a strong correlation with distance from Africa in Eurasian, Oceanian and American populations. Among some of the populations with the highest observed genetic diversity are those with a high degree of admixture between divergent populations, such as the Khwe which have nearly equal amounts of Khoe-San and Bantu-speaking ancestry. This suggests that current genetic diversity may be more strongly shaped by recent genetic history than population dynamics around the time of modern human origins.

However, patterns of genetic diversity can never provide strong evidence for the origin of expansions. Better information on the origin of expansions might instead be gained by analyzing the genetic relationships of populations. For example, if all African populations where found to share a more recent genetic history with some Khoe-San groups than other Khoe-San groups, that could be taken as evidence of an expansion out of southern Africa. Analyses of population history such as this are problematic mainly because of post-divergence gene flow affecting the genetic affinities of different populations. To circumvent this and investigate the relationship between Khoe-San and other African populations such as Central African Pygmies, West African speakers of Niger-Kordofanian languages, the widespread speakers of Bantu-related languages, and click-speaking East African populations such as the Hadza and Sandawe, we used a new approach to reconstruct tree-like models of population history based on sampling single gene copies from quartets of populations ('concordance tests'). We found that this method is robust to substantial amounts of gene flow after divergence, which makes it more suitable for investigating African population history than distance-based methods. We found that six out of seven Khoe-San populations formed a separate lineage that appears to have been involved in the earliest diversification of modern human populations (the remaining population, the Khwe, trace more than half of their ancestry to Bantu-speaking populations), which we estimate to at least 100 thousand years ago. Since the Khoe-San share a common origin, this leaves the question of modern human origins in Africa unanswered. However, several features of more recent African population history can be discerned, such as pastoralist populations in Africa other than Khoe groups appearing to share a common origin to the exclusion of diverse hunter-gatherer groups such as Mbuti and Biaka Pygmies of Central Africa, as well as the Hadza and Sandawe of eastern Africa. We also used *D*-

tests and admixture graph inference to test for admixture between these hunter-gatherer groups, and found several cases of gene flow, most notably between the ancestors of East African Hadza and southern African San (Ju/'hoansi). Notably, we also detected significant population structure between Northern and Southern Khoe-San groups, and estimate that the population divergence times between them might date back as far as 25,000-43,000 years ago.

Since we find support for a model where all the ancestors of most Khoe-San groups diverged from other human populations 100,000 years ago, we devised an approach that utilizes this early divergence to search for signals of selective sweeps, zooming in specifically on the period prior to the divergence of human populations such as the Bantu-speakers from the Khoe-San. To achieve this we designed a statistic, *aPBS*, that quantifies whether two populations have a high and balanced frequency of derived alleles, and identified regions that had segments of consecutively high values for this statistic. To estimate *aPBS*, we used the F_{ST}-based framework of Yi *et al.* (2010) in the hypothesized history (chimpanzee, (Khoe-San, Bantu-speakers)). More generally, we first estimated F_{ST} using a commonly applied estimator (Weir 1996) between each pairwise combination of two populations (A and B), and an outgroup (O), and transformed each value to a scale that is proportional to genetic drift under neutrality T=-log(1- F_{ST}). Then *aPBS* was defined as

$$aPBS = (T_{A-O} + T_{B-O} - T_{A-B})/2$$

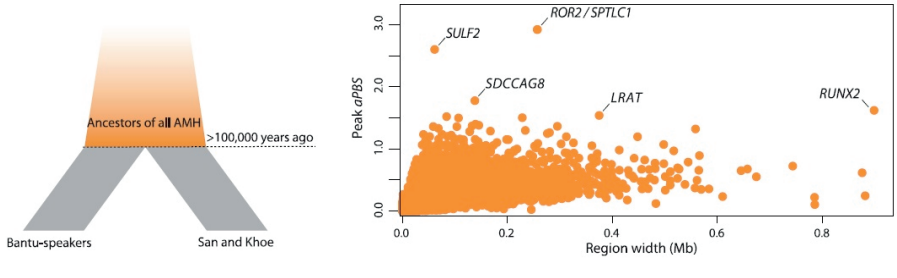

Figure 5. We found that all 7 Khoe-San groups investigated share a common origin, with divergence from Bantu-speakers more than 100,000 years ago (left). This allowed us to devise a statistic, *aPBS*, that is tuned to detect regions with high-frequency derived alleles in both populations. The regions are candidates for being under selection during the emergence of early modern humans (right).

The reasoning behind looking for this pattern is that selective sweeps in the ancestral population of the Khoe-San and Bantu-speakers are expected to have dragged long segments of the genome to high- or fixed frequencies, and this signal is not likely to be detected using more standard statistics aimed at detecting recent selective sweeps. If we were to look for regions of high derived allele frequencies only in a single population, the top candidates would many times be more recent selective sweeps in that population, and we would be prone to miss more ancient selection events. However, two other processes that could produce high *aPBS* values in the genome could be convergent selection or more recent selection in both populations where variation is shared through gene flow.

The longest consecutive region we identified harbored the *RUNX2* gene (*Figure 5*), which was also identified as one of the top 20 candidates in a similar study using the Neandertal genome (Green *et al.* 2010). *RUNX2* mutations cause cladiocranial dysplasia, resulting in morphological features such as a supraorbital brow ridge, a flaring rib cage, and hypertelomeric limbs, all features known to have changed rapidly during the emergence of early modern humans. The region with the greatest observed *aPBS* value comprised *ROR2*, a gene involved in bone and cartilage development, for which mutations are known to cause *e.g.* shorter limb development. All in all, three of the top five regions in our genome scan contained genes involved in skeletal development. This is noteworthy since the present-day gracile human morphology appeared relatively abruptly compared to previous rates of morphological evolution in the human lineage (Tattersall 2009). This provides genetic evidence that natural selection targeting skeletal morphology coincided with the emergence of anatomical modernity.

Some time after the emergence of early modern humans in Africa, their descendants expanded into Eurasia, and it has been suggested that they absorbed gene flow from archaic humans on at least two occasions: from Neandertals and from Denisovans further in the east (Green *et al.* 2010; Reich *et al.* 2010). However, the distribution and extent of this ancestry has not been investigated in a large set of worldwide human populations. To address this, we overlapped genomic data from the Neandertal and Denisovan draft genomes with genotypes from more than a thousand individuals from across the globe from which we had information on half a million genetic markers in the genome (Altshuler *et al.* 2010; Li *et al.* 2008b; Surakka *et al.* 2010). Previous analyses had suggested that Neandertal ancestry was uniformly distributed in Eurasia and the Americas, whereas Denisova ancestry was restricted to Oceania (Reich *et al.* 2010). In **Paper II**, we performed principal component analysis on the chimpanzee, Neandertal, and Denisova genomes, and projected the present-day human populations on the resulting axes of variation. We found that genetic affinity to the two archaic humans

increased from west to east in Eurasia, in contrast to previous suggestions. However, using simulations we also showed that a serial founder model of increasing genetic drift from west to east could produce a similar pattern in data with ascertainment bias for common alleles (such as the empirical data analyzed), but some southeast Asian populations affinities to archaic humans (in particular Denisovans) were difficult to reconcile with this expectation (see below). Recent analysis of the complete high-coverage Denisova genome have partly confirmed this view, arguing for a higher degree of archaic Neandertal-related ancestry in East Eurasia (~1.7%) compared to West Eurasia (~1.0%) (Meyer *et al.* 2012). In our analysis we also found that on the East Asian mainland, southern populations showed significantly higher degree of genomic affinity to the Denisova genome compared to northern populations. In the light of the Central Siberian location of Denisova cave, this casts new light on the previously puzzling observations that Denisovan ancestry was completely restricted to the Oceanian islands, and absent from mainland East Asia (Reich *et al.* 2010; Reich *et al.* 2011). Our study suggests that the original admixture event could have taken place on the mainland, but the genetic traces may have been subsequently diluted by later migrations.

Denisovans share a common origin with Neandertals, but were genetically differentiated from Neandertals in Europe and the Caucasus (Reich *et al.* 2010). In **Paper V**, we obtained a complete mitochondrial genome sequence from a Neandertal in Okladnikov cave, close to the Denisova cave, and show that its mtDNA lineage is basal to all published complete Neandertal mtDNA sequences from Central and Western Europe. However, robust conclusions about Neandertal population structure and their genetic history with Denisovans will require more genomic information from both populations. These studies will also provide more information on when and how admixture with modern humans took place.

Prehistoric interactions between hunter-gatherers and agriculturalists (Papers I, III, and IV)

One of the most important events in human history is the transition from foraging lifestyles to sedentary lifestyles that utilize the resources of domesticated animals and/or plants. This process occurred independently in many different parts of the world with the onset of warmer climate approximately 10,000 years ago. In Europe, it spread from the Near East, and one of the most contentious issues surround whether it spread by means of migration or as an idea that was taken up by resident populations (Ammerman& Cavalli-Sforza 1984), or perhaps rather what the relative contributions of the two processes were. Scandinavia was one of the last places in Europe to be reached by agriculture, and for a period, populations living primarily as hunter-gatherers and farmers were living in the same region at the same time. A null hypothesis can thus be posited where under complete cultural transmission of agriculture, individuals from farming and hunter-gatherer groups are not expected to have different genetic ancestry.

To test this hypothesis directly, we used shotgun sequencing to generate low-coverage genetic data from three Neolithic hunter-gatherers associated with the Pitted Ware Culture (PWC) and one Neolithic farmer associated with the Funnel Beaker culture (TRB), all from present-day Sweden and dated to approximately 5,000 years ago, to a total amount of 249 million base pairs (**Paper III**). We identified ancient DNA sequences in the data that overlapped with genetic markers typed in individuals from modern-day populations from across Europe and Western Eurasia, and compared the ancient individuals to the genetic variation of modern-day humans. We found that all three individuals associated with the foraging PWC were outside the population structure of modern-day Europeans but most closely related to modern-day Northern Europeans (*Figure 6*), whereas the individual associated with TRB was most closely related to Southern Europeans. This suggests strong population structure in Scandinavia 5,000 years ago, and is consistent with a model where the spread of the Neolithic was facilitated by population migrations from the south. We also observed evidence for admixture related to the Neolithic farmer individual in modern-day populations, and found that ancestry related to the Neolithic individual increased from South to North in Europe and in local Swedish populations when these populations are fitted as two-way admixtures in a simplified model of population history where Finnish populations are taken as most closely related to the alternative source of ancestry.

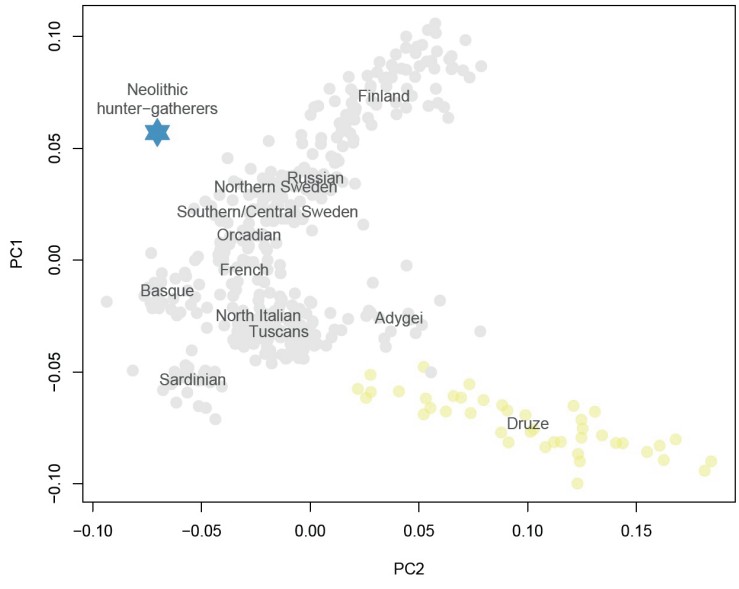

Figure 6. Neolithic hunter-gatherers are outside the population structure of modern-day Europe. A principal component analysis was performed on 14,113 SNPs obtained from pooled data from the 3 Neolithic hunter-gatherers in Paper III and present-day European and Middle Eastern populations.

In a follow up study (**Paper IV**), we obtained more sequence data from three of the previous individuals, and also sequenced 3 additional individuals associated with the TRB and 3 additional individuals associated with the PWC. Under the hypothesis that they represent different populations, we pooled the data from each group, and found that the closest modern-day groups to the TRB data and the PWC data are Sardinians, and Lithuanians, respectively. We then tested for each separate individual if they were closer to either northern Europeans (Lithuanians) or southern Europeans (Sardinians), or none. We found that all TRB-associated individuals are significantly closer to Sardinians, and all PWC-associated individuals are significantly closer to Lithuanians. In addition, we performed the same test for the 5,300 year old Tyrolean Iceman from the Italian Alps (Keller *et al.* 2012) and two Mesolithic Iberian individuals from La Braña, Spain (Sánchez-Quinto *et al.* 2012). We find that the Tyrolean Iceman is significantly closer to Sardinians whereas the Mesolithic Iberians are significantly closer to Lithuanians. This demonstrates that hunter-gatherer and farmer individuals sampled so far share genetic signatures with individuals with the same mode of subsistence from other parts of Europe, and that the Neolithic PWC individuals trace their ancestry to Mesolithic European populations. We also confirmed this by reconstructing maximum likelihood trees (Pickrell& Pritchard 2012), and by restricting the analysis to sequences with postmortem damage pattern using the methods in **Paper V**.

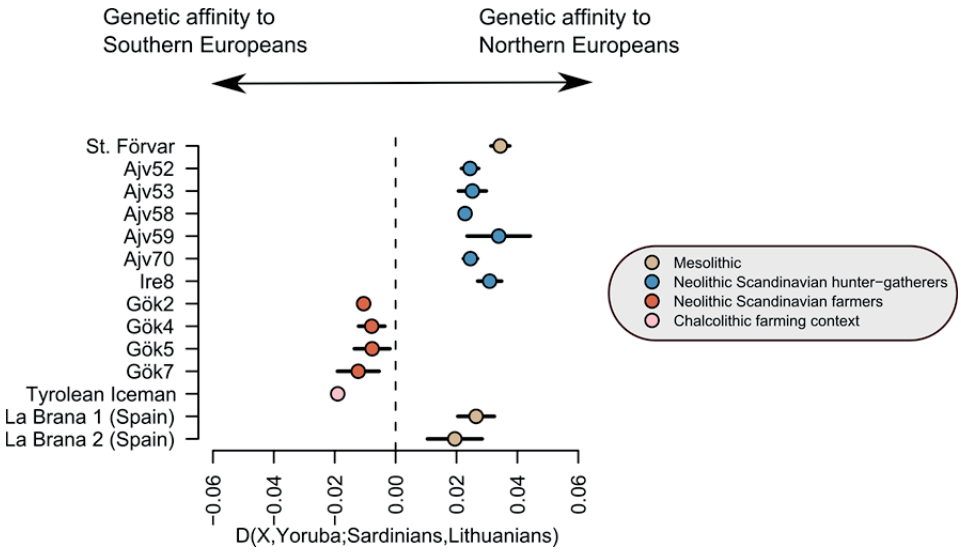

Figure 7. Test for whether each of 14 ancient individuals is closer to the genomic variation of Southern Europeans (Sardinians) or Northern Europeans (Lithuanians), or neither. Error bars represent 1 standard error.

To investigate the shared signal of the Neolithic hunter-gatherers of the PWC with Mesolithic individuals in further detail, we have also analyzed a Mesolithic individual from Stora Förvar cave on the Stora Karlsö Island in the Baltic Sea (Götherström A, Jakobsson M, Malmström H, Omrak A, Skoglund P, Storå J, *unpublished results*), which is very close to the locality where the PWC individuals in our studies were excavated. The individual was directly radiocarbon dated to 7,952 ± 53 BP (6,873 ± 119 cal BC). We used the Illumina HiSeq platform to obtain a total of 2,250,928 sequences that could be confidently aligned to the human genome, and found strongly elevated levels of C→T and G→A mismatches indicative of post-mortem nucleotide misincorporations (Briggs *et al.* 2007). To assess whether there in addition to endogenous DNA was evidence of modern-day contamination, we exploited the approximately 12-fold coverage of the mitochondrial genome to estimate contamination (Krause *et al.* 2010a), and found that all 61 informative sequences were consistent with the consensus, which yields a point estimate of 0.0% modern contamination. We also found that the mtDNA sequence belonged to haplogroup U4b1, consistent with the notion that mitochondrial lineages belonging to hg U were fixed or nearly fixed in pre-Neolithic Europe (Fu *et al.* 2013b; Pinhasi *et al.* 2012). To assess whether this individual showed a similar genetic signature to other ancient individuals from the Neolithic period, we repeated the test for if the individual was closer to Northern Europeans (Lithuanians), Southern Europeans (Sardinians), or neither. We found that the Stora Förvar individual was significantly closer to Lithuanians in this analysis (*Figure 7*), which is consistent with a

high degree of continuity from the Mesolithic in the individuals associated with the Pitted Ware culture.

In contrast to the apparently strong population structure observed between ancient farmers and hunter-gatherers in Europe, Khoe pastoralists (the Nama) in southern Africa were genetically closely related to the San hunter-gatherers in our study of modern southern African genetic variation (**Paper I**). However, we did observe a significant but minor affinity to East Africans in the Nama pastoralists, which suggests that they may have absorbed gene flow from incoming pastoralists that have been hypothesized to also have sparked the appearance of Khoe-Kwadi languages in the region (Güldemann 2008). The relative minor fraction of this ancestry that we observed suggests that part of this process involved cultural diffusion (in addition to demic diffusion) of the practice of pastoralism, but more studies are needed to resolve this question further. Our analyses thus suggest that the genetic makeup of modern-day southern Africa mainly stems from three major ancestral populations. The first of these were likely the ancestors of the Khoe-San peoples, and our estimates of historical divergences between different Khoe-San groups into the Pleistocene suggest some form of continuity in the region from that time. The second major identified source of ancestry of the present-day population is associated with the expansion of iron-working Bantu-speaking agropastoralists from their equatorial homeland. This ancestry is present in many Khoe-San populations, but it is noteworthy that Khoe-San ancestry is also present in all southeastern Bantu-speakers in our study (representing Zulu, Xhosa and Tswana). The third is the large-scale colonization of European and Asian migrants, the ancestry of which is present not only in the Coloured populations, but which we also observed in many Khoe-San groups (with a great degree of inter-individual variation).

To search for evidence of local adaptation between southern African populations using the more than 2 million SNPs in our data, we computed F_{ST} for every SNP and between every pair of populations, and contrasted the highest observed value in each comparison with the genome-wide average. We found that the greatest differences between the genome-wide average F_{ST} and the most divergent marker were all in comparisons between the Nama and other Khoe-San, and all in a small region on chromosome 16. To further scrutinize this interesting region, we performed genome-local ancestry deconvolution, and found that the same region had a highly unusual degree of Bantu-related ancestry. While no coding regions reside inside the identified region, our results suggest that it could harbor adaptive variants in regulatory elements, possibly associated with the pastoralist lifestyle of the Nama.

New approaches for ancient genomics (Papers II through VI)

Due to the susceptibility of DNA analysis of ancient human remains to present-day contamination, the most important step for this type of studies is to investigate the authenticity of the genetic data and the results derived from them. While strict precautions during sample preparation for sequencing are standard, including the addition of sample-specific DNA barcodes, we can never exclude contamination fully using laboratory methods. This is especially the case for ancient material that has been handled in other osteological analyses where no precautions against DNA contamination were taken. The main feature distinguishing DNA from several thousand year old remains from more recent DNA is the presence of nucleotide misincorporations arising from DNA damage (cytosine deamination) (Hansen *et al.* 2001; Hofreiter *et al.* 2001), and a first step towards authentication of ancient genomic data is the observation of these damage patterns. However, if genomic data is a mix of ancient DNA and more recent contamination, it could still show a pattern of DNA damage, only diminished by the presence of molecules with little damage. In **Paper III**, we first introduced a new approach to authentication especially geared towards low-coverage genome sequencing. We identified and isolated specific sequences with mismatches to the human reference genome consistent with ancient DNA damage, and compared our population genetic results for that set of sequences to the set that did not display evidence of ancient DNA damage. We found that our conclusions about population structure between Neolithic farmers and hunter-gatherers were robust in both these subsections of the data. If our results were caused by contamination we would have expected the sequences without evidence of damage to be a mixture of endogenous and contaminating sequences, whereas the sequences with damage would have consisted almost exclusively of endogenous sequences.

In **Paper V** we extend this approach significantly by developing a statistical framework to compare the likelihood of a sequence being endogenous or contaminant, resulting in a likelihood ratio of two different models that we term a 'PMD score'. We show that this method is able to reduce even high contamination fractions to negligible levels, and is more powerful than the simple approach first used in Paper III. To validate the method, we use artificial contamination experiments of both high-coverage Neandertal mitochondrial data and low-coverage autosomal genomic data from Neolithic humans. We also validate the efficiency of method on Neandertal DNA sequence data sets known to be affected by genuine contamination, since present-day contamination can be identified by mitochondrial variants that are fixed between Neandertals and modern humans. Finally, we use the method

on a novel mtDNA data set generated from a Neandertal individual from Okladnikov cave in Siberia that contains substantial amounts of present-day human contamination, and use our method to reconstruct its mitochondrial genome sequence.

Another main problem in this early generation of ancient genomic studies is the difficulty associated with obtaining high-coverage data from ancient remains. Indeed, most studies of ancient DNA still focus on single markers such as mitochondrial DNA, but direct sequencing is able to obtain a much larger set of loci without much increased laboratory effort per sample. The challenge is that these sequences will be spread out in random locations of the genome, making comparisons between ancient individuals difficult. In **Paper III**, we addressed this by comparing each low-coverage ancient individual to high-quality population data sets from modern groups separately using principal component analysis. In order to facilitate interpretation of the separate analyses, we then used Procrustes transformation, a statistical technique to compare multidimensional configurations (Wang *et al.* 2010), to rotate the PC1-PC2 configuration of each individual to that of the reference data set (*Figure 8*). This results in a 'merged' comparison, revealing that hunter-gatherer individuals are all closer to Northern Europeans than the farmer individual.

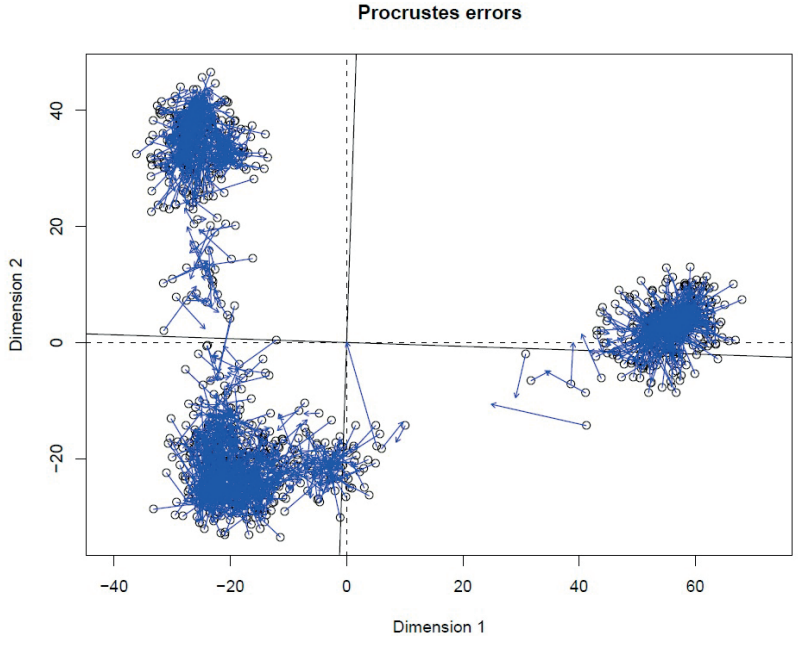

Figure 8. Illustration of Procrustes transformation of one two-dimensional configuration to another, with residuals indicated in blue.

When one or a few ancient individuals from a particular geographic and temporal context are analyzed, questions are often focused on their relationships with modern-day populations. In **Paper III**, we computed a pairwise allele sharing statistic, and showed that in European populations, a high degree of allele sharing with hunter-gatherers predicted a lower degree of allele sharing with farmers. In **Paper VI**, we used simulations to show that standard methods such as principal component analysis are sensitive to different degrees of genetic drift in the time period that separates the ancient individual(s) and the modern populations. To overcome this issue, we present an approach in **Paper IV** where we compute an f_3 statistic (Patterson *et al.* 2012) not to detect admixture but using an outgroup, a focus population, and a population for comparison. The approach is designed to reflect the shared genetic history of the focus population and the population for comparison. This is highly useful as conventional statistics measuring divergences between populations are heavily affected by post-divergence genetic drift. It also allows different types of these tests to be contrasted, and we show that shared genetic drift between modern populations and farmers and hunter-gatherers, respectively, refines the remarkable pattern in Europe captured by the allele sharing statistic.

In **Paper III**, we also introduced an alternative approach for inferring tree-like population histories that is insensitive to lineage-specific genetic drift. We sample a random gene copy from each of four populations. If a population history that conforms to the tree $(A,B),(X,Y)$ where A,B,X, and Y are different populations, is correct, we expect to observe more loci with allele configurations concordant with the proposed topology than any other configuration. We thus use the counts N_{conc} for the number of loci with a SNP configuration where A and B have the same allele and X and Y have the same allele. We use N_{disc1} for the number of loci that shows the second most frequent tree-like topology. Then our statistic C is

$$C = (N_{conc} - N_{disc1})/(N_{conc} + N_{disc1})$$

which we compute for all proposed topologies and assess which one is supported (if any) by the data. This approach has similarities to previous frameworks for asking similar questions (Green *et al.* 2010; Reich *et al.* 2010), and in **Paper I** we show that it is highly robust to recent admixture.

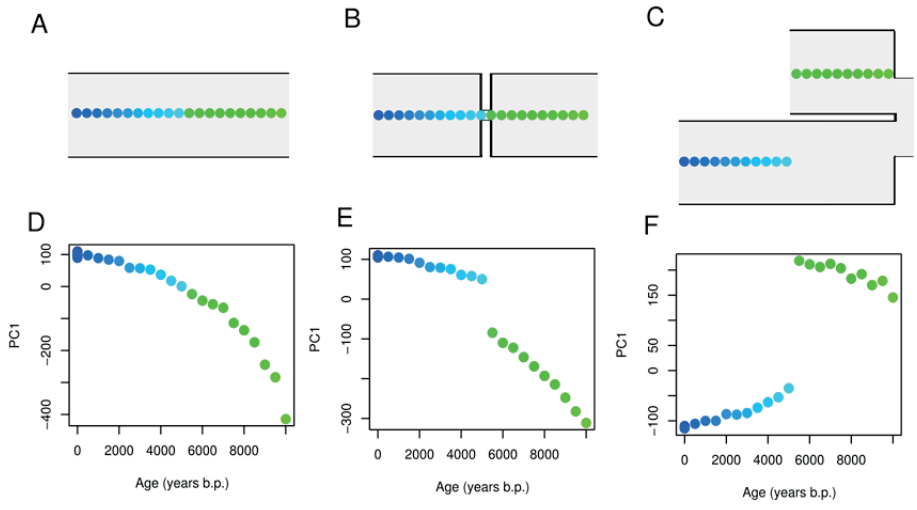

Figure 9. Temporal sampling schemes help distinguishing genetic drift from population replacement. A) Constant size model. B) Bottleneck model. C) Replacement model. D) PC1 stratified by sample-time under the constant size model. E) PC1 stratified by sample-time under the bottleneck model. F) PC1 stratified by sample-time under the replacement model.

The 4-population tests and *D*-tests have been found to be robust to ascertainment bias, the bias that arises when discovering a set of genetic markers in a limited panel, in the sense that it is not expected to create false positive signals of admixture (Durand *et al.* 2011; Reich *et al.* 2009). However, in **Paper II** we show that in the presence of admixture and ascertainment bias, populations will display deviations in the 4-population test statistic which are correlated to the amount of genetic drift in their history, perhaps due to the process of filtering out rare alleles leading to false negatives. For example, since a 1-4% admixture with Neandertals occurred in the ancestral population of all non-Africans (potentially in the Middle East), subsequent founder effects during the expansion in Eurasia and the Americas could have resulted in a gradient of apparent Neandertal ancestry, which increases with distance from Africa in panels of common SNPs.

Another large difference between analyses of ancient DNA and that of only present-day individuals is that individuals from different time points are inescapably separated by genetic drift. This genetic drift makes it difficult to interpret differences between ancient and modern individuals, or even ancient individuals from different time points, since the effects of genetic drift and possible population replacement by migrants are confounded. In **Paper VI**, we investigate the properties of Wright's F_{ST} in relation to samples from different time points, and show that separation by time and population structure share mathematical properties in the genetic differentiation they predict. In this paper, we also show that the best remedy for the analytical difficulty

of distinguishing between genetic drift caused by *e.g.* bottlenecks and population replacements when comparing data from different time points is in fact to incorporate a large set of samples from a range of time points. When samples from such a study-design are analyzed, the observed trajectories in genetic affinities in the putatively ancestral populations can be used to gauge whether a population replacement is likely to have taken place (*Figure 9*).

Conclusions and future prospects

In this thesis, I have shown that the integration of ancient and modern large-scale genomic data can provide powerful means for learning about human population history. Specifically, I have shown that direct sequencing of ancient human remains provides substantial power for population genetic analyses despite challenges such as low genomic coverage, high rate of postmortem DNA damage, present-day contamination and genetic differentiation arising from temporal differences. I present several methodological approaches to alleviate the problems associated with these features, as well as how to utilize the information they provide. First, I present solutions to combine the analyses made from several low-coverage data sets from different ancient remains to allow inferences about the differentiation and genetic affinities of the ancient individuals. These approaches, such as Procrustes transformation of principal components, concordance tests, outgroup f3 statistics, and allele sharing, are also robust to postmortem damage and sequence errors. Secondly, I show that post-mortem damage can be used to remove present-day human contamination from ancient DNA data sets in silico. Using a new probabilistic framework for this, we isolate endogenous Neandertal DNA from a large-scale data set affected by substantial present-day human contamination and reconstruct the near-complete mitochondrial genome sequence of the Neandertal individual. I also explore the population genetic properties of temporal genetic sampling schemes, and show that temporal genetic differentiation shares many conceptual features with spatial sampling, but that temporal sampling provides valuable information about population history by tracing the trajectory of allele frequency changes through time.

I use these approaches to analyze new genomic data from Neolithic farmers and hunter-gatherer contexts in Scandinavia 5,000 years ago. I find that individuals from the two contexts show starkly different genetic signature, with the Neolithic hunter-gatherers being outside the variation of modern-day humans but most closely related to populations from the Baltic area today, whereas Neolithic farming individuals where most closely related to Southern European populations such as Sardinians. Genomic data from hunter-gatherers and farmers from Southern Europe show similar genetic signatures associated with their mode of subsistence, which together with evidence for

admixture in modern-day populations suggests major northward expansions of agriculturalists during the Neolithic.

I also use a novel genome-wide SNP data set from southern African Khoe-San populations to update our current view on genetic variation in living human populations. Reconstructing the population history of present-day African populations, I find that the ancestors of the Khoe-San were involved in one of the earliest diversification events of modern humans 100,000 years ago. I use this observation to search for evidence of selective sweeps in the ancestral population of Khoe-San and Bantu-speaking populations, which would have lived during the emergence of anatomical modernity between 100 and 300 thousand years ago. In remarkable agreement with the observation of an acceleration of human anatomical evolution in the African fossil record during this period, I find that the major candidate regions for selective sweeps overlap genes involved in skeletal development. In more recent African history, we also find evidence of shared ancestry with East Africa and local adaptions in the Nama, a Khoe-San population that practices a pastoralist lifestyle.

Together these results show that the direct sequencing approach together with the power of large genetic data sets from modern-day humans holds the promise to revolutionize our understanding of the human past. Future developments in the field of statistical population genetic inference that specifically use the information provided by the age of multiple ancient individuals should be especially promising. Since bioinformatic approaches are now in place for dealing with present-day human contamination, technological advances that will allow retrieval of large-scale DNA sequence data from ancient remains also from environments with less favorable preservation conditions will allow innumerable palaeoanthropological, archaeological, and historical hypotheses to be tested using genetics.

Svensk sammanfattning

På samma sätt som att människans rika historia har lämnat spår i form av arkeologiskt material och språks utbredning över världen så har historien också lämnat spår i DNA-variationen hos dagens individer. Trots att variationen mellan individer från samma grupp är nästan lika stor som variationen mellan individer från olika grupper, så kan genomets miljontals variabla positioner användas till att kartlägga genetiska mönster i tid och rum och rekonstruera de historiska processer som har format deras variation. Svårigheterna med att analysera DNA direkt från arkeologiskt skelettmaterial har dock länge inneburit att genomiska studier av populationshistoria har varit begränsade till att tolka mönster av variation i nu levande människor, utan tillgång till direkta historiska observationer. I den här avhandlingen utarbetar jag nya metoder för isolering, analys och tolkning av stora genomiska data från forntida populationer, inklusive metoder för att kunna analysera data med begränsad genomtäckning, DNA-kontamination från personer som arbetat med materialet, stora åldersskillnader mellan olika prover, och postmortem DNA-skador. Detta gör det möjligt att integrera storskaliga genomiska data från arkeologiska lämningar med genetisk variation från nu levande människor för att spåra nutida och forntida människopopulationers historia.

Genom att först rekonstruera de tidigaste diversifieringarna av tidiga moderna människor med hjälp av nutida variation från Khoe-San populationer i södra Afrika utarbetades en ny metod för att söka efter gener som kan ha varit viktiga för formandet av anatomiskt moderna människor för några hundra tusen år sedan. De gener som identifieras i denna analys indikerar att skelettets utveckling påverkades av naturligt urval under den tidsperiod då de första anatomiskt modern människorna levde. Genom att jämföra tillgängliga genomdata från arkaiska människor såsom Neandertalare och Denisova-människor med nu levande människor var det också möjligt att visa att andelen genetiskt material som kommer från arkaiska människor varierar över Eurasien, i motsats till tidigare studier som föreslagit en homogen fördelning i denna region.

I den första direkta genomiska studien av populationsstruktur i forntida populationer visar jag att ca. 5000 år gamla individer associerade med jordbruk- och jägar-samlar grupper i Skandinavien under yngre stenåldern var starkt genetiskt differentierade. Direkta jämförelser med nutida populationer

och andra förhistoriska individer från södra Europa visar att detta mönster återkommer i andra delar av Europa och att den genetiska strukturen troligen har sin grund i att jordbrukande populationer från södra Europa migrerade norrut. Slutligen utvecklar jag en bioinformatisk metod för att ta bort kontamination från nutida personer från DNA-sekvenser isolerade från antikt mänskligt material, och använder denna metoden att rekonstruera en komplett mitokondrie-DNA-sekvens från en sibirisk Neandertalare.

Acknowledgements

First of all I want to thank Mattias Jakobsson for accepting me as a graduate student, always being available for stimulating discussions, pushing along the projects with never-ending energy and creating a great group environment over the years. I also want to thank my co- supervisor Anders Götherström for being a constant source of suggestions, insight and inspiration. Thanks also to Hanna Johannesson and Hans Ellegren for hosting me at the department and for creating a great research environment.

Big thanks to all past members and visitors of Mattias' lab. Carina Schlebusch, role model of anthropological genetics. Helena Malmström, ancient DNA lab overlord. Per Sjödin, Lucie Gattepaille and Sen Li, mathematical and computational wizards. Thanks also for interesting collaborations and discussions to Tobias Skoglund, Agnes Sjöstrand, Magdalena Fraser, Joakim Karlsson, Nicolas Duforet-Frebourg, Hiba Babiker, Gwenna Breton, Nina Hollfelder, Solenn Stoeckel, Christoff Erasmus, Patrik Båtelsson, Ricardo Rodriguez Varela, Katie Owers, Torsten Günther, and TJ Naidoo. Also big thanks to past crazy- and non-crazy members and visitors in Anders' ancient DNA group. Special thanks to Ayça Omrak and Helena Malmström for their contributions to the work behind this thesis but also to Emma Svensson, Eva Daskalaki, Oddny Sverrisdottir, Maja Krzewinska, Linus Girdland-Flink, Irene Ureña, Frida Johnsson, and Carolin Johansson.

Thanks to the Copenhagen-Berkeley crew for always interesting collaborations and discussions. Special thanks to Maanasa Rhagavan for her contributions to the work underlying this thesis, but also Eske Willerslev, Tom Gilbert, Rasmus Nielsen, Ludovic Orlando, Morten Rasmussen, Maria Avila-Arcos, Ida Moltke, Anders Albrechtsen, Simon Rasmussen, Mike DeGeorgio, Mait Metspalu, Amhed Missael Vargas Velasquez, and Hannes Schröder.

I also am very grateful to Johannes Krause, Svante Pääbo, Bernd Nordhoff and Professors Shunkov and Derevianko for the collaborations on rescuing ancient mtDNA and very illuminating discussions.

I am also fortunate to have been able to collaborate with several other people: Jan Storå in Stockholm; Karl-Göran Sjögren in Göteborg; Cris Valdiosera in Melbourne; Michael Blum and Flora Jay in Grenoble; Himla Soodyall and group in Johannesburg; Love Dalén and Veronica Nyström in Stockholm; Santiago-Castroviejo-Fisher, Carles Vilá, and Jennifer Leonard by way of Sevilla; Jacob Höglund on the floor above; and all other co-authors and collaborators.

For hopefully continuing discussions about population genetics and other things I want to thank Padraic Corcoran, Michael Stocks, Martin Lascoux, Thomas Källman, and Jelmer Poelstra. Thanks also to other people I have been bugging for help with computational issues: Benoit Nabholz, Pall Olason, Tobias Skoglund, Axel Künstner, Jochen Wolf, Björn Rogell, Tanja Slotte, Linnea Smeds and Udo Stenzel. Thanks also to the master students in the evolutionary biology and MEME programs for great journal club discussions over the years. Thanks to Frida, Signe and Emma for helping with administration, and thanks to everyone else at the evolutionary biology department and the EBC for a fun and inspiring environment!

Thanks to people who have been kind enough to host me abroad: Michael Blum and group in Grenoble; Himla Soodyall and group in Johannesburg; Susanna Sawyer, Svante Pääbo and group in Leipzig; David Reich and group in Boston; Maanasa Rhagavan, Eske Willerslev and group in Copenhagen.

I also want to acknowledge the individuals who donated DNA used in study about southern African genetic variation. More information on Khoe-San peoples in southern Africa and their current situation can be found at www.san.org.za.

The Royal Swedish Academy of Sciences, the Sven and Lilly Lawski foundation, the Helge Ax:son foundation, the Nilsson-Ehle foundation, and the Lars Hierta foundation provided funding for work presented in this thesis. I am also grateful to Hans Ellegren, Mattias Jakobsson, Ted Morrow, and Jochen Wolf for organizing the graduate school on genomes and phenotypes for the travel opportunities that it has provided.

Special thanks also to Mattias Jakobsson, Carina Schlebusch, Helena Malmström, Jan Storå, and Eleftheria Palkopoulou for suggestions and comments on the kappa of this thesis.

Finally, I am very grateful for the almost overwhelming support of Eleftheria, mamma och pappa, Tobias och Elisabeth, and friends in Uppsala, Stockholm, Umeå, Malmö, Thessaloniki, Kalmar, Öland and beyond. Tack!

References

Ahlström T (1997a) Den exogama gränsen: Kring interaktionen mellan jägare-samlare och bönder-boskapsskötare under mellanneolitisk tid. In: *Till Gunborg: arkeologiska samtal* (eds. Bergh S, Nordbladh J, Taffinder J, Åkerlund A), pp. 325-338. Stockholm University, Stockholm.

Ahlström T (1997b) Pitted-Ware skeletons and boreal temperatures. *Lund archaeological review* **3**, 37-48.

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* **27**, 2534-2547.

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

Altshuler DM, Gibbs RA, Peltonen L*, et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52.

Ammerman AJ, Cavalli-Sforza LL (1984) *The Neolithic transition and the genetics of populations in Europe* Princeton University Press Princeton.

Arenas M, François O, Currat M, Ray N, Excoffier L (2013) Influence of admixture and paleolithic range contractions on current European diversity gradients. *Molecular Biology and Evolution* **30**, 57-61.

Axelsson E, Willerslev E, Gilbert MTP, Nielsen R (2008) The effect of ancient DNA damage on inferences of demographic histories. *Molecular Biology and Evolution* **25**, 2181-2187.

Barbieri C, Vicente M, Rocha J*, et al.* (2013) Ancient Substructure in Early mtDNA Lineages of Southern Africa. *The American Journal of Human Genetics* **92**, 285-292.

Barbujani G, Di Benedetto G (2001) Genetic variances within and between human groups. *Genes, Fossils and Behaviour*, 63-77.

Barham L, Mitchell P (2008) *The first Africans: African archaeology from the earliest tool makers to most recent foragers* Cambridge University Press Cambridge.

Barker G (1985) *Prehistoric farming in Europe* CUP Archive.

Beaumont MA, Zhang WY, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025 - 2035.

Behar DM, Villems R, Soodyall H*, et al.* (2008) The Dawn of Human Matrilineal Diversity. *The American Journal of Human Genetics* **82**, 1130-1140.

Bentley DR, Balasubramanian S, Swerdlow HP*, et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59.

Blow MJ, Zhang T, Woyke T*, et al.* (2008) Identification of ancient remains through genomic sequencing. *Genome research* **18**, 1347-1353.

Bramanti B, Thomas M, Haak W*, et al.* (2009) Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *science* **326**, 137-140.
Briggs AW, Good JM, Green RE*, et al.* (2009) Targeted Retrieval and Analysis of Five Neandertal mtDNA Genomes. *science* **325**, 318-321.
Briggs AW, Stenzel U, Johnson PLF*, et al.* (2007) Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences* **104**, 14616-14621.
Briggs AW, Stenzel U, Meyer M*, et al.* (2010) Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic acids research* **38**, e87-e87.
Brown P, Sutikna T, Morwood MJ, Soejono RP (2004) A new small-bodied hominin from the Late Pleistocene of Flores, Indonesia. *Nature* **431**, 1055-1061.
Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics* **81**, 1084-1097.
Bryc K, Auton A, Nelson MR*, et al.* (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences* **107**, 786-791.
Burbano HA, Hodges E, Green RE*, et al.* (2010) Targeted Investigation of the Neandertal Genome by Array-Based Sequence Capture. *science* **328**, 723-725.
Burger J, Kirchner M, Bramanti B, Haak W, Thomas MG (2007) Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proceedings of the National Academy of Sciences* **104**, 3736-3741.
Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* **325**, 31-36.
Castroviejo-Fisher S, Skoglund P, Valadez R, Vila C, Leonard J (2011) Vanishing native American dog lineages. *BMC Evolutionary Biology* **11**, 73.
Cavalli-Sforza LL, Edwards AW (1967) Phylogenetic analysis. Models and estimation procedures. *American journal of human genetics* **19**, 233.
Cavalli-Sforza LLL, Menozzi P, Piazza A (1994) *The history and geography of human genes* Princeton university press.
Champion T, Gamble C, Shennan S, Whittle A (2009) *Prehistoric Europe* Left Coast Press.
Chen C, Durand E, Forbes F, François O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes* **7**, 747-756.
Childe VG (1925) The Dawn of European Civilization (Kegan Paul, London). *AMERICAN ANTIQUITY* **674**.
Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome research* **15**, 1496-1502.
Conrad DF, Jakobsson M, Coop G*, et al.* (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature genetics* **38**, 1251-1260.
Cooper A, Poinar HN (2000) Ancient DNA: Do It Right or Not at All. *science* **289**, 1139b-.

Corander J, Marttinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. *Molecular Ecology* **15**, 2833-2843.
d'Errico F, Backwell L, Villa P*, et al.* (2012) Early evidence of San material culture represented by organic artifacts from Border Cave, South Africa. *Proceedings of the National Academy of Sciences* **109**, 13214-13219.
Dalén L, Orlando L, Shapiro B*, et al.* (2012) Partial genetic turnover in Neandertals: Continuity in the east and population replacement in the west. *Molecular Biology and Evolution* **29**, 1893-1897.
Darwin CR (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* John Murray, London.
DeGiorgio M, Jakobsson M, Rosenberg NA (2009) Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proceedings of the National Academy of Sciences* **106**, 16057-16062.
DeGiorgio M, Rosenberg NA (2013) Geographic sampling scheme as a determinant of the major axis of genetic variation in principal components analysis. *Molecular Biology and Evolution* **30**, 480-488.
Deguilloux MF, Leahy R, Pemonge MH, Rottier S (2012) European neolithization and ancient DNA: an assessment. *Evolutionary Anthropology: Issues, News, and Reviews* **21**, 24-37.
Deguilloux MF, Soler L, Pemonge MH*, et al.* (2011) News from the west: Ancient DNA from a French megalithic burial chamber. *American Journal of Physical Anthropology* **144**, 108-118.
Depaulis F, Orlando L, Hanni C (2009) Using classical population genetics tools with heterochroneous data: time matters! *PloS one* **4**, e5541.
Der Sarkissian C, Balanovsky O, Brandt G*, et al.* (2013) Ancient DNA reveals prehistoric gene-flow from siberia in the complex human population history of north East europe. *PLoS Genetics* **9**, e1003296.
Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* **28**, 2239-2252.
Edwards AWF (2003) Human genetic diversity: Lewontin's fallacy. *BioEssays* **25**, 798-801.
Engelhardt BE, Stephens M (2010) Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics* **6**, e1001117.
Eriksson G, Linderholm A, Fornander E*, et al.* (2008) Same island, different diet: Cultural evolution of food practice on Öland, Sweden, from the Mesolithic to the Roman Period. *Journal of Anthropological Archaeology* **27**, 520-543.
Fischer A (2002) Food for feasting. *The neolithisation of Denmark* **150**, 341e395.
François O, Currat M, Ray N*, et al.* (2010) Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution* **27**, 1257-1268.
Fu Q, Meyer M, Gao X*, et al.* (2013a) DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences* **110**, 2223-2227.
Fu Q, Mittnik A, Johnson Philip LF*, et al.* (2013b) A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. *Current Biology* **23**, 553-559.

Fu W, O/'Connor TD, Jun G*, et al.* (2013c) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216-220.
Gamba C, Fernández E, Tirado M*, et al.* (2012) Ancient DNA from an Early Neolithic Iberian population supports a pioneer colonization by first farmers. *Molecular Ecology* **21**, 45-56.
Gattepaille LM, Jakobsson M (2012) Combining markers into haplotypes can improve population structure inference. *Genetics* **190**, 159-174.
Gilbert MTP, Bandelt HJ, Hofreiter M, Barnes I (2005) Assessing ancient DNA studies. *Trends in Ecology & Evolution* **20**, 541-544.
Gravel S (2012) Population genetics models of local ancestry. *Genetics* **191**, 607-619.
Gravel S, Henn BM, Gutenkunst RN*, et al.* (2011) Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences* **108**, 11983-11988.
Green RE, Briggs AW, Krause J*, et al.* (2009) The Neandertal genome and ancient DNA authenticity. *The EMBO journal* **28**, 2494-2502.
Green RE, Krause J, Briggs AW*, et al.* (2010) A Draft Sequence of the Neandertal Genome. *science* **328**, 710-722.
Green RE, Krause J, Ptak SE*, et al.* (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330-336.
Green RE, Malaspinas A-S, Krause J*, et al.* (2008) A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**, 416-426.
Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A (2011) Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics* **43**, 1031-1034.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics* **5**, e1000695.
Güldemann T (2008) A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. *Southern African Humanities* **20**, 93-132.
Götherström A, Collins M, Angerbjörn A, Lidén K (2002) Bone preservation and DNA amplification. *Archaeometry* **44**, 395-404.
Götherström A, Liden K, Ahlström T, Källersjö M, Brown T (1997) Osteology, DNA and sex identification: Morphological and molecular sex identifications of five Neolithic individuals from Ajvide, Gotland. *International Journal of Osteoarchaeology* **7**, 71-81.
Haak W, Forster P, Bramanti B*, et al.* (2005) Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *science* **310**, 1016-1018.
Hallgren F (2008) *Identitet i praktik: lokala, regionala och överregionala sociala sammanhang inom nordlig trattbägarkultur*, Uppsala University.
Hansen AJ, Willerslev E, Wiuf C, Mourier T, Arctander P (2001) Statistical evidence for miscoding lesions in ancient DNA templates. *Molecular Biology and Evolution* **18**, 262-265.
Hawks J (2012) Longer time scale for human evolution. *Proceedings of the National Academy of Sciences* **109**, 15531-15532.
Henn BM, Gignoux C, Lin AA*, et al.* (2008) Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proceedings of the National Academy of Sciences* **105**, 10693-10698.

Henn BM, Gignoux CR, Jobin M*, et al.* (2011) Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences* **108**, 5154-5162.
Henshilwood CS, d'Errico F, Yates R*, et al.* (2002) Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *science* **295**, 1278-1280.
Hervella M, Izagirre N, Alonso S*, et al.* (2012) Ancient DNA from hunter-gatherer and farmer groups from Northern Spain supports a random dispersion model for the Neolithic expansion into Europe. *PloS one* **7**, e34417.
Higuchi R, Bowman B, Freiberg M, Ryder OA, Wilson AC (1984) DNA sequence from the quagga, an extinct member of the horse family. *Nature* **312**, 282-284.
Hinch AG, Tandon A, Patterson N*, et al.* (2011) The landscape of recombination in African Americans. *Nature* **476**, 170-175.
Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Paabo S (2001) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic acids research* **29**, 4793-4799.
Hublin J-J (2009) The origin of Neandertals. *Proceedings of the National Academy of Sciences* **106**, 16022-16027.
Hudson RR (2002) Generating samples under a Wrightâ€“Fisher neutral model of genetic variation. *Bioinformatics* **18***,* 337-338.
Huerta-Sánchez E, DeGiorgio M, Pagani L*, et al.* (2013) Genetic Signatures Reveal High-Altitude Adaptation in a Set of Ethiopian Populations. *Molecular Biology and Evolution* **30**, 1877-1888.
Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708-713.
Ingram CJ, Mulcare CA, Itan Y, Thomas MG, Swallow DM (2009) Lactose digestion and the evolutionary genetics of lactase persistence. *Human genetics* **124**, 579-591.
Jakobsson M, Scholz SW, Scheet P*, et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998-1003.
Johnson NA, Coram MA, Shriver MD*, et al.* (2011) Ancestral components of admixed genomes in a Mexican cohort. *PLoS Genetics* **7**, e1002410.
Jónsson H, Ginolhac A, Schubert M, Johnson P, Orlando L (2013) mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics*.
Jungers WL, Harcourt-Smith W, Wunderlich R*, et al.* (2009) The foot of Homo floresiensis. *Nature* **459**, 81-84.
Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *science* **336**, 740-743.
Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature genetics* **39**, 1251-1255.
Keller A, Graefen A, Ball M*, et al.* (2012) New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun* **3**, 698.
Kingman JFC (1982a) The coalescent. *Stochastic processes and their applications* **13**, 235-248.

Kingman JFC (1982b) Exchangeability and the evolution of large populations.
Kingman JFC (1982c) On the genealogy of large populations. *Journal of Applied Probability*, 27-43.
Kong A, Frigge ML, Masson G*, et al.* (2012) Rate of de novo mutations and the importance of father/'s age to disease risk. *Nature* **488**, 471-475.
Krause J, Briggs AW, Kircher M*, et al.* (2010a) A Complete mtDNA Genome of an Early Modern Human from Kostenki, Russia. *Current biology : CB* **20**, 231-236.
Krause J, Fu QM, Good JM*, et al.* (2010b) The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464**, 894-897.
Krings M, Capelli C, Tschentscher F*, et al.* (2000) A view of Neandertal genetic diversity. *Nature genetics* **26**, 144-146.
Krings M, Stone A, Schmitz RW*, et al.* (1997) Neandertal DNA sequences and the origin of modern humans. *Cell* **90**, 19-30.
Kuhner MK, Yamato J, Felsenstein J (1998) Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**, 429-434.
Kunsch HR (1989) The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* **17**, 1217-1241.
Lacan M, Keyser C, Crubézy E, Ludes B (2012) Ancestry of modern Europeans: contributions of ancient DNA. *Cellular and Molecular Life Sciences*, 1-15.
Lacan M, Keyser C, Ricaut F-X*, et al.* (2011) Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. *Proceedings of the National Academy of Sciences* **108**, 9788-9791.
Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. *PLoS Genetics* **8**, e1002453.
Leonard JA, Shanks O, Hofreiter M*, et al.* (2007) Animal DNA in PCR reagents plagues ancient DNA research. *Journal of Archaeological Science* **34**, 1361-1366.
Lewontin RC (1972) The Apportionment of Human Diversity. In: *Evolutionary Biology* (eds. Dobzhansky T, Hecht M, Steere W), pp. 381-398. Springer US.
Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics* **11**, 473-483.
Li JZ, Absher DM, Tang H*, et al.* (2008a) Worldwide human relationships inferred from genome-wide patterns of variation. *science* **319**, 1100-1104.
Li JZ, Absher DM, Tang H*, et al.* (2008b) Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *science* **319**, 1100-1104.
Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-2233.
Lidén K (1996) A dietary perspective on Swedish hunter-gatherer and Neolithic populations. *Laborativ Arkeologi* **9**, 5-23.
Lidén K, Eriksson G, Nordqvist B*, et al.* (2004) The wet and the wild followed by the dry and the tame-or did they occur at the same time?: Diet in Mesolithic Neolithic southern Sweden. *Antiquity*, 23-33.

Linderholm A (2008) *Migration in prehistory: DNA and stable isotope analyses of Swedish skeletal material*, Stockholm.
Linderholm A, Malmström H, Lidén K, Holmlund G, Götherström A (2008) Cryptic contamination and phylogenetic nonsense. *PloS one* **3**, e2316.
Lindqvist C, Possnert G (1997) The subsistence economy and diet at Jakobs/Ajvide, Eksta parish and other prehistoric dwelling and burial sites on Gotland in long-term perspective. *Remote sensing* **1**, 29-90.
Lipson M, Loh P-R, Levin A*, et al.* (2013) Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow. *Molecular Biology and Evolution* **30**, 1788-1802.
Loh P-R, Lipson M, Patterson N*, et al.* (2012) Inference of admixture parameters in human populations using weighted linkage disequilibrium.
Malmer MP (2002) *The Neolithic of south Sweden: TRB, GRK, and STR* Royal Academy of Letters &.
Malmström H, Gilbert MTP, Thomas MG*, et al.* (2009) Ancient DNA Reveals Lack of Continuity between Neolithic Hunter-Gatherers and Contemporary Scandinavians. *Current Biology* **19**, 1758.
Malmström H, Linderholm A, Lidén K*, et al.* (2010) High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern Europe. *BMC Evolutionary Biology* **10**, 89.
Malmström H, Storå J, Dalén L, Holmlund G, Götherström A (2005) Extensive human DNA contamination in extracts from ancient dog bones and teeth. *Molecular Biology and Evolution* **22**, 2040-2047.
McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics* **5**, e1000686.
Mellars P (2006) Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *science* **313**, 796-800.
Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in Europeans. *science* **201**, 786-792.
Meyer M, Kircher M, Gansauge M-T*, et al.* (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *science* **338**, 222-226.
Midgley MS (2008) *The Megaliths of Northern Europe* Routledge.
Mikkelsen TS, Hillier LW, Eichler EE*, et al.* (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87.
Miller W, Drautz DI, Ratan A*, et al.* (2008) Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**, 387-390.
Moorjani P, Patterson N, Hirschhorn JN*, et al.* (2011) The history of African gene flow into Southern Europeans, Levantines, and Jews. *PLoS Genetics* **7**, e1001373.
Mounier A, Condemi S, Manzi G (2011) The stem species of our species: a place for the archaic human cranium from Ceprano, Italy. *PloS one* **6**, e18821.
Nelson MR, Wegmann D, Ehm MG*, et al.* (2012) An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. *science* **337**, 100-104.
Nielsen R (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197-218.
Nielsen R, Beaumont MA (2009) Statistical inferences in phylogeography. *Molecular Ecology* **18**, 1034-1047.

Nielsen R, Mountain JL, Huelsenbeck JP, Slatkin M (1998) Maximum-likelihood estimation of population divergence times and population phylogeny in models without mutation. *Evolution*, 669-677.
Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**, 443-451.
Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* **158**, 885-896.
Noonan JP, Coop G, Kudaravalli S*, et al.* (2006) Sequencing and analysis of Neanderthal genomic DNA. *science* **314**, 1113-1118.
Noonan JP, Hofreiter M, Smith D*, et al.* (2005) Genomic sequencing of Pleistocene cave bears. *science* **309**, 597-600.
Nordborg M (1998) On the probability of Neanderthal ancestry. *American journal of human genetics* **63**, 1237.
Novembre J, Johnson T, Bryc K*, et al.* (2008) Genes mirror geography within Europe. *Nature* **456**, 274-274.
Novembre J, Ramachandran S (2011) Perspectives on Human Population Structure at the Cusp of the Sequencing Era. *Annual review of genomics and human genetics* **12**, 245-274.
Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nature genetics* **40**, 646-649.
Nyström V, Humphrey J, Skoglund P*, et al.* (2012) Microsatellite genotyping reveals end-Pleistocene decline in mammoth autosomal genetic variation. *Molecular Ecology* **In press**.
Orlando L, Ginolhac A, Zhang G*, et al.* (2013) Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74-78.
Ovchinnikov IV, Götherström A, Romanova GP*, et al.* (2000) Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* **404**, 490-493.
Parra EJ, Kittles RA, Argyropoulos G*, et al.* (2001) Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *American Journal of Physical Anthropology* **114**, 18-29.
Patterson N, Hattangadi N, Lane B*, et al.* (2004) Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics* **74**, 979-1000.
Patterson N, Moorjani P, Luo Y*, et al.* (2012) Ancient admixture in human history. *Genetics* **192**, 1065-1093.
Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* **2**, e190.
Pickrell JK, Patterson N, Barbieri C*, et al.* (2012) The genetic prehistory of southern Africa. *Nat Commun* **3**, 1143.
Pickrell JK, Patterson N, Loh P-R*, et al.* (2013) Ancient west Eurasian ancestry in southern and eastern Africa. *arXiv preprint arXiv:1307.8014*.
Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* **8**, e1002967.
Pinhasi R, Thomas MG, Hofreiter M, Currat M, Burger J (2012) The genetic history of Europeans. *Trends in Genetics*.
Pleurdeau D, Imalwa E, Détroit F*, et al.* (2012) "Of Sheep and Men": Earliest Direct Evidence of Caprine Domestication in Southern Africa at Leopard Cave (Erongo, Namibia). *PloS one* **7**, e40340.

Poinar HN, Schwarz C, Qi J*, et al.* (2006) Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *science* **311**, 392-394.
Pool JE, Nielsen R (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics* **181**, 711-719.
Prado-Martinez J, Sudmant PH, Kidd JM*, et al.* (2013) Great ape genetic diversity and population history. *Nature* **499**, 471-475.
Price AL, Tandon A, Patterson N*, et al.* (2009) Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. *PLoS Genet* **5**, e1000519.
Pritchard J, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945 - 959.
Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution* **16**, 1791-1798.
Pritchard JK, Stephens M, Donnelly P (2000b) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959.
Prugnolle F, Manica A, Balloux F (2005) Geography predicts neutral genetic diversity of human populations. *Current Biology* **15**, R159-R160.
Pugach I, Matveyev R, Wollstein A, Kayser M, Stoneking M (2011) Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biol* **12**, R19.
Pääbo S (1985) Molecular cloning of ancient Egyptian mummy DNA.
Pääbo S (1989) Ancient DNA; extraction, characterizaation, molecular cloning and enzymatic amplification. *Proceedings of the National Academy of Sciences* **86**.
Pääbo S, Poinar H, Serre D*, et al.* (2004) Genetic analyses from ancient DNA. *Annual Review of Genetics* **38**, 645-679.
Ramachandran S, Deshpande O, Roseman CC*, et al.* (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15942-15947.
Ramírez-Soriano A, Nielsen R (2009) Correcting estimators of θ and Tajima's D for ascertainment biases caused by the single-nucleotide polymorphism discovery process. *Genetics* **181**, 701-710.
Rasmussen M, Guo X, Wang Y*, et al.* (2011) An Aboriginal Australian genome reveals separate human dispersals into Asia. *science* **334**, 94-98.
Rasmussen M, Li Y, Lindgreen S*, et al.* (2010) Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**, 757-762.
Reich D, Green RE, Kircher M*, et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053-1060.
Reich D, Patterson N, Campbell D*, et al.* (2012) Reconstructing native American population history. *Nature* **488**, 370-374.
Reich D, Patterson N, De Jager PL*, et al.* (2005) A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nature genetics* **37**, 1113-1118.
Reich D, Patterson N, Kircher M*, et al.* (2011) Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *The American Journal of Human Genetics* **89**, 516-528.
Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* **461**, 489-494.

Relethford JH, Harpending HC (1995) Ancient differences in population size can mimic a recent African origin of modern humans. *Current Anthropology* **36**, 667-674.
Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet* **3**, 380-390.
Rosenberg NA, Pritchard JK, Weber JL*, et al.* (2002) Genetic structure of human populations. *science* **298**, 2381-2385.
Sabeti P, Schaffner S, Fry B*, et al.* (2006) Positive natural selection in the human lineage. *science* **312**, 1614-1620.
Sadr K (1998) The first herders at the Cape of Good Hope. *African Archaeological Review* **15**, 101-132.
Sánchez-Quinto F, Schroeder H, Ramirez O*, et al.* (2012) Genomic Affinities of Two 7,000-Year-Old Iberian Hunter-Gatherers. *Current biology : CB* **22**, 1494-1499.
Sankararaman S, Patterson N, Li H, Pääbo S, Reich D (2012) The date of interbreeding between Neandertals and modern humans. *PLoS Genetics* **8**, e1002947.
Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. *The American Journal of Human Genetics* **82**, 290-303.
Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S (2012) Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PloS one* **7**, e34131.
Scally A, Durbin R (2012) Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics* **13**, 745-753.
Scally A, Dutheil JY, Hillier LW*, et al.* (2012a) Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169-175.
Scally A, Dutheil JY, Hillier LW*, et al.* (2012b) Insights into hominid evolution from the gorilla genome sequence. *Nature* **483**, 169-175.
Scarre C (2009) *The Human Past: World Prehistory & the Development of Human Societies* Thames & Hudson New York.
Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics* **78**, 629-644.
Scheinfeldt LB, Soi S, Tishkoff SA (2010) Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. *Proceedings of the National Academy of Sciences* **107**, 8931-8938.
Schlebusch C (2010) Issues raised by use of ethnic-group names in genome study. *Nature* **464**, 487-487.
Schlebusch CM, Lombard M, Soodyall H (2013) MtDNA control region variation affirms diversity and deep sub-structure in populations from Southern Africa. *BMC Evolutionary Biology* **13**, 56.
Schubert M, Ginolhac A, Lindgreen S*, et al.* (2012) Improving ancient DNA read mapping against modern reference genomes. *BMC genomics* **13**, 178.
Schuster SC, Miller W, Ratan A*, et al.* (2010) Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943-947.
Serre D, Langaney A, Chech M*, et al.* (2004a) No evidence of neandertal mtDNA contribution to early modern humans. *PLoS biology* **2**, 313-317.

Serre D, Langaney A, Chech M*, et al.* (2004b) No evidence of Neandertal mtDNA contribution to early modern humans. *PLoS biology* **2**, e57.
Simonson TS, Yang Y, Huff CD*, et al.* (2010) Genetic Evidence for High-Altitude Adaptation in Tibet. *science* **329**, 72-75.
Skoglund P, Götherström A, Jakobsson M (2011) Estimation of population divergence times from non-overlapping genomic sequences: examples from dogs and wolves. *Molecular Biology and Evolution* **28**, 1505-1517.
Skoglund P, Storå J, Götherström A, Jakobsson M (2013) Accurate sex identification of ancient human remains using DNA shotgun sequencing. *Journal of Archaeological Science* **40**, 4477-4482.
Slatkin M (1991) Inbreeding Coefficients and Coalescence Times. *Genetical research* **58**, 167-175.
Smith AB (1992) Origins and spread of pastoralism in Africa. *Annual Review of Anthropology* **21**, 125-141.
Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. *Genet Res* **23**, 23-35.
Stoneking M, Krause J (2011) Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics* **12**, 603-614.
Stringer C (2002) Modern human origins: progress and prospects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **357**, 563-579.
Stringer CB, Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. *science* **239**, 1263-1268.
Stynder D, Ackermann R, Sealy J (2007a) Early to mid-Holocene South African Later Stone Age human crania exhibit a distinctly Khoesan morphological pattern. *South African Journal of Science* **103**, 349-352.
Stynder DD, Ackermann RR, Sealy JC (2007b) Craniofacial variation and population continuity during the South African Holocene. *American Journal of Physical Anthropology* **134**, 489-500.
Sun JX, Helgason A, Masson G*, et al.* (2012) A direct characterization of human mutation based on microsatellites. *Nature genetics*.
Surakka I, Kristiansson K, Anttila V*, et al.* (2010) Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome research* **20**, 1344-1351.
Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics* **79**, 1-12.
Tattersall I (2009) Human origins: out of Africa. *Proceedings of the National Academy of Sciences* **106**, 16018-16021.
Tavare S, Balding DJ, Griffiths R, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505-518.
Tennessen JA, Bigham AW, O'Connor TD*, et al.* (2012) Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *science* **337**, 64-69.
The 1000 genomes project consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65.
Tishkoff SA, Reed FA, Friedlaender FR*, et al.* (2009) The Genetic Structure and History of Africans and African Americans. *science* **324**, 1035-1044.
Tishkoff SA, Reed FA, Ranciaro A*, et al.* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39**, 31-40.

Tocheri MW, Orr CM, Larson SG*, et al.* (2007) The primitive wrist of Homo floresiensis and its implications for hominin evolution. *science* **317**, 1743-1745.
Trinkaus E (2005) Early modern humans. *Annual Review of Anthropology* **34**, 207-230.
Wakeley J (1996) Distinguishing migration from isolation using the variance of pairwise differences. *Theoretical population biology* **49**, 369-386.
Wakeley J, Hey J (1998) Testing speciation models with DNA sequence data. In: *Molecular approaches to ecology and evolution*, pp. 157-175. Springer.
Valdiosera C, García N, Dalén L*, et al.* (2006) Typing single polymorphic nucleotides in mitochondrial DNA as a way to access Middle Pleistocene DNA. *Biology letters* **2**, 601-603.
Wall JD, Jiang R, Gignoux C*, et al.* (2011) Genetic variation in Native Americans, inferred from Latino SNP and resequencing data. *Molecular Biology and Evolution* **28**, 2231-2237.
Wall JD, Kim SK (2007) Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genetics* **3**, e175.
Wang C, Szpiech ZA, Degnan JH*, et al.* (2010) Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Statistical applications in genetics and molecular biology* **9**.
Wang Y, Nielsen R (2012) Estimating population divergence time and phylogeny from single-nucleotide polymorphisms data with outgroup ascertainment bias. *Molecular Ecology* **21**, 974-986.
Veeramah KR, Wegmann D, Woerner A*, et al.* (2012) An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. *Molecular Biology and Evolution* **29**, 617-630.
Wegmann D, Kessner DE, Veeramah KR*, et al.* (2011) Recombination rates in admixed individuals identified by ancestry-based inference. *Nat Genet* **43**, 847-853.
Weir BS (1996) *Genetic Data Analysis II* Sinauer Associates, Inc., Sunderland.
White TD, Asfaw B, DeGusta D*, et al.* (2003) Pleistocene homo sapiens from middle awash, ethiopia. *Nature* **423**, 742-747.
Willerslev E, Cappellini E, Boomsma W*, et al.* (2007) Ancient Biomolecules from Deep Ice Cores Reveal a Forested Southern Greenland. *science* **317**, 111-114.
Willerslev E, Cooper A (2005) Ancient DNA. *Proceedings of the Royal Society of London Series B-Biological Sciences* **272**, 3-16.
Wollstein A, Lao O, Becker C*, et al.* (2010) Demographic history of Oceania inferred from genome-wide data. *Current Biology* **20**, 1983-1992.
Wolpoff MH, Hawks J, Caspari R (2000) Multiregional, not multiple origins.
Wright S (1949) Population structure in evolution. *Proc Am Philos Soc* **93**, 471-478.
Yi X, Liang Y, Huerta-Sanchez E*, et al.* (2010) Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *science* **329**, 75-78.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 1069

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology.

Distribution: publications.uu.se
urn:nbn:se:uu:diva-206787

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2013