Prospects & Overviews

Recently in press

Analysis of ancient human genomes

Using next generation sequencing, 20-fold coverage of the genome of a 4,000-year-old human from Greenland has been obtained

*Beth Shapiro[1] and Michael Hofreiter[2]**

High-capacity sequencing technologies have dramatically reduced both the cost and time required to generate complete human genome sequences. Besides expanding our knowledge about existing diversity, the nature of these technologies makes it possible to extend knowledge in yet another dimension: time. Recently, the complete genome sequence of a 4,000-year-old human from the Saqqaq culture of Greenland was determined to 20-fold coverage. These data make it possible to investigate the population affinities of this enigmatic culture and, by identifying several phenotypic traits of this individual, provide a limited glimpse into how these people may have looked. While undoubtedly a milestone in ancient DNA research, the cost to generate an ancient genome, even from such an exceptionally preserved specimen, remains out of reach for most. Nonetheless, recently developed DNA capture methods, already applied to Neanderthal and fossil human mitochondrial DNA, may soon make large-scale genome-wide analysis of ancient human diversity a reality, providing a fresh look at human population history.

Keywords:

- Greenland; human ancient DNA; next generation sequencing; population affinities; Saqqaq

DOI 10.1002/bies.201000026

[1] Department of Biology, The Pennsylvania State University, University Park, PA, USA
[2] Evolutionary Biology and Ecology, Department of Biology, University of York, York, UK

***Corresponding author:**
Michael Hofreiter
E-mail: michi@palaeo.eu

Abbreviations:
NGS, next generation sequencing, **SNP**, single nucleotide polymorphism

Introduction

The introduction of next generation sequencing (NGS) technology has transformed the field of ancient DNA. Traditionally limited by the small number of surviving, amplifiable DNA molecules, it took more than 20 years to increase the amount of sequencing data obtained from a sample by two orders of magnitude. After the introduction of the first NGS technology, the next almost three orders of magnitude took only about 6 months, and during the time since then, the output has increased by more than another three orders of magnitude (Fig. 1).

Although NGS has significantly increased the amount of raw data that can be extracted from ancient specimens, the utility of these data has yet to be fully realized. The first report using 454 sequencing – the first NGS technology to become widely available[1] – reported 13 million base pairs of mammoth DNA from a single permafrost-preserved bone [2]. While this pioneering report demonstrated conclusively that NGS could be applied to ancient DNA, little novel biologic insight was extrapolated from the data produced. A similar conclusion can be drawn from the subsequently published draft genome of the extinct woolly mammoth [3]. Again, this work demonstrated that an enormous quantity of data could be amplified from ancient specimens, yet exactly what could be done with these data was left for future work to address. If one were to be negative, one might reflect that NGS has brought ancient DNA back to the ''bad old days'' (in which both present authors participated), when the announcement of a small sequence fragment from an extinct species was sufficient for a high-profile publication even if the biologic insight provided by these data was limited.

In addition to allowing significantly more data to be produced, NGS has helped us to become more aware of the challenges specific to our field. For example, the high potential for contamination of ancient extracts with modern human DNA has been a major obstacle for research involving human remains. Famous cases of contamination include the first reported ancient human DNA sequence from a 4,000-year-

Recently in press

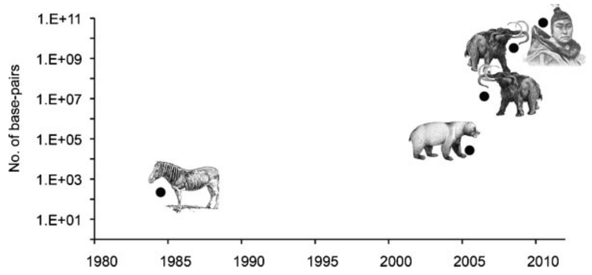

Figure 1. Increase of DNA sequence output using ancient DNA. All studies compared used a metagenomic approach. The *x*-axis depicts the years in which the respective study was published. All leaps in sequence throughput, apart from the two mammoth studies 2006 and 2008, respectively, were preceded by changes in sequencing technology. Note that the *Y*-axis is logarithmic. From left to right: quaaga 1984 [32], cave bear 2005 [25], mammoth 2006 [2], mammoth 2008 [3], and human 2010 [13].

old mummy [4] and mitochondrial sequences from Mungo man [5], a human fossil from Australia, both of which were most likely false positives [6–8]. As a consequence, the ancient DNA field has adopted a range of experimental protocols designed to avoid and detect contaminants prior to publication [8]. The necessity of such cautionary measures was involuntarily highlighted in the first application of NGS to a hominid remain: a fragment of Neanderthal bone [9], which claimed to have sequenced around 1 million base pairs of Neanderthal DNA. When compared to modern human genomic sequences, however, the average sequence divergence date was suspiciously young, almost identical to that obtained for the comparison of two modern humans. By comparison to a smaller data set from a second Neanderthal genome study that used traditional cloning and sequencing [10] it was later shown that the NGS study had been affected by contamination to a considerable extent, possibly up to 80% of the hominid sequences comprising modern DNA [11]. Although this first estimate of the amount of contamination may have been unreasonably high, the authors of the initial study later conceded that up to 40% of their sequences may indeed go back to contamination [12], most likely because the sequencing libraries were not constructed under clean room conditions, but in the standard laboratory of the 454 sequencing facility.

With these two problems in mind, the recent publication of a complete human genome from a 4,000-year-old sample of hair belonging to a person from the Saqqaq culture of Greenland [13] is an impressive milestone for ancient DNA. Rasmussen *et al.* explore the possibility of human contamination and provide convincing arguments as to why their results are robust to this problem. They then attempt to use these data to provide novel, biologic insights at both the population and individual level. While no doubt a technical achievement, two important questions are raised from this work. First, what do we actually learn from the addition of an ancient genome to the growing database of human genomes? And second, what is the next step for ancient DNA?

Interpreting the genome of a palaeoeskimo

The sample from which the Saqqaq genome was sequenced is a 4,000-year-old tuft of human hair found in southwestern Greenland. Although less frequently preserved than bone, hair seems to be particularly suitable for DNA preservation [14], as it is less porous and therefore less likely to ‘‘take up’’ contaminating, modern DNA. The same sample had been used previously to sequence the complete mitochondrial genome [15]. This was accomplished *via* shotgun sequencing using the 454 platform, an approach that had been used previously to successfully sequence complete mitochondrial genomes from several mammoth hair tufts [16]. From this exceptional human specimen, the authors estimate an extremely low level of contamination: 0.8% of reads. This number is derived from an analysis of the frequency of observed SNPs that should be found only in modern European populations (the potential contaminants, as no participating researcher was of Asian ancestry), and not in the Saqqaq genome. These two pieces of fortune – relatively recent, preserved hair and a genetically distinct specimen – are crucial to the success of this breakthrough project, but highlight problems that will be faced when working with less ideal ancient specimens.

To sequence the complete Saqqaq genome, Rasmussen *et al.* use Illumina sequencing technology, which has a much higher throughput per sequencing run than the Roche/454 platform and is less expensive per nucleotide to perform. However, the resulting sequencing reads are significantly shorter, and greater coverage is required to unambiguously assemble the resulting data. The Saqqaq sample was sequenced to an average coverage of 20-fold, with an average read length of about 55 base pairs and a maximum read length of around 70 base pairs. With this coverage, although 84% of the reads could be identified as being of human origin, only about half of the correctly indexed reads could be unambiguously mapped to the human genome, resulting in approximately 85% coverage of the complete genome. While the problem of unambiguously mapping very short fragments is not unique to ancient DNA, it is important to note that ancient DNA sequences are generally highly fragmented. It is therefore unlikely that a full genome will ever be completely assembled from fossil DNA.

Despite this minor shortcoming, the first sequencing of a large part of the nuclear genome to high coverage is no doubt a milestone in ancient DNA research. A more critical question, however, is what can be learned about human history from the addition of an ancient genome. Although the authors briefly discuss specific SNPs from this sample that have been associated previously with phenotypic traits, these are largely immaterial (dry earwax, brown eyes) or self-evident from the sample (thick hair). Much more interesting are two other portions of their analyses, namely SNP detection and population affinities. In this context, the term SNP is used to describe a nucleotide position that is variable in comparison to other human genome sequences without considering the population frequency of this variant, as normally done.

Using this definition, a total of 2.2 million SNPs were identified in the Saqqaq genome, of which 13.8% (approximately 300,000) were new when compared to the public SNP database (dbSNPv130). However, a set of filtering steps reduces the 2.2 million to 350,000 high-quality SNPs. This reduced data set has a 93.2% overlap with dbSNP, reducing the number of high-quality, novel SNPs to around 24,000. If we assume this to be the true ratio of as yet unknown SNPs for the Saqqaq sample, the complete Saqqaq genome would actually contain only 150,000 new SNPs. In other words, as many as half of the originally reported 300,000 new SNPs could be due to various types of sequencing errors. This highlights another major limitation of ancient genome projects. Not only does the short read length prevent the reconstruction of a complete nuclear genome, even where assembly is feasible, verification of a truly high-quality sequence becomes prohibitively expensive. These issues will be of importance for the interpretation of any ancient genome, including a potential Neanderthal genome sequence, for which a onefold coverage has been announced [17] and for which it has been claimed that a 12-fold coverage would be sufficient to obtain a reliable high-quality sequence [18].

What does the addition of temporal depth add to the quest for novel human diversity? In another, recently published human genome study, complete genome sequences were obtained for hunter-gatherers from southern Africa [19]. Here, more than 700,000 new SNPs were detected from a single individual. Given that human genetic diversity is by far greatest in Africa [20], this result is not entirely surprising. However, it underlines the fact that, given that modern humans originated in Africa, the diversity that can be found by probing African population is most likely always larger than what can be found by sequencing fossils from regions outside Africa.

While ancient non-African genomes may not be informative with regard to the total genetic diversity in humans, the data are clearly useful to investigate the population affinities of enigmatic cultural groups such as the Saqqaq. The previously published mitochondrial genome from this individual suggested a close affinity with either Aleut people from the Commander Islands or the Sireniki and Yuit from Siberia, rather than with modern Inuit [15]. However, mitochondrial DNA provides only a single, maternally inherited locus, and nuclear data are required to investigate population relationships on a finer scale. By comparing SNP diversity within the Saqqaq genome to a worldwide human population sample (the CEPH panel) and 16 North American and northern Asia individuals, it was again shown that this individual was more closely related to northern Asian than to New World populations, including modern Inuit. A fine-scale analysis of the population affinities suggests that the Saqqaq individual lacked admixture from western Eurasia, and, in contrast to modern Inuit, also lacked admixture from Amerindians. However, their data suggested that modern Inuit and Saqqaq shared a ‘‘Beringian’’ component that is also shared by modern Chukchi and Koryak, suggesting that Saqqaq and Inuit shared a common ancestor in Asia prior to the Inuit dispersal into North America. Finally, the authors estimated the most likely date of divergence between the Saqqaq individual and the closest modern population, the Chukchi. Their estimate of around 5,500 years ago had a wide confidence interval (ca. 2,000 years), but coincides with the first archaeologic evidence of the Saqqaq culture.

What next?

As noted above, the most interesting part of the analysis – apart from the achievement of sequencing an extinct human genome in itself – lies in the population analysis. The population affinities of the Saqqaq culture have been debated ever since the first evidence of the culture were discovered. But only now, using modern genetics has it become possible to address questions about the evolutionary history of the Saqqaq people. The obvious next step would be to extend this analysis to other, ancient human fossils whose evolutionary history might be revealed using genomic analysis. Genomic data from ancient remains could provide key insights into questions about the evolutionary origins of enigmatic early American fossils [21] or about the genetic continuity between Palaeolithic hunter-gatherers, Neolithic farmers, and modern European populations [22–24]. The next, much more difficult challenge will therefore be to sequence a human genome from a temperate environment. In fact such a sequence is under way in the form of the Neanderthal genome [17]. However, unlike remains from anatomically modern humans, the Neanderthal sequencing effort will also benefit from genetic dissimilarity from modern humans as a means to detect and estimate contamination.

Another problem for future ancient human genome sequencing efforts will be the expense. As fossils from temperate environments generally have a much lower proportion of endogenous DNA compared to fossils from very cold regions [2, 25], more sequencing will be required to produce equivalent genomic coverage. This is demonstrated by the $6.4 million cost to generate onefold coverage of the Neanderthal genome [17], compared to $500,000 for 20-fold coverage of the Saqqaq genome [26] (although some of this difference is due to different sequencing approaches). Given that the proportion of endogenous DNA in temperate samples is usually at least 10 but often more than 50 times lower than that estimated for the Saqqaq hair specimen, it is unlikely that we will see large-scale population analyses using a complete genome sequencing approach in the near future. Even if we assume a relatively high percentage of endogenous DNA of 4% and another drop in sequencing costs by a factor of 10, which is realistic, each ancient, temperate-climate genome would still cost about $1 million. This is not a price for which one would tackle 20 or 50 samples.

But then, for most purposes, complete genome sequences are also not necessary. The population analysis using the Saqqaq sample was performed using a mere fraction of the total data set – and for the comparative samples, genomic sequencing was not even attempted. Therefore, a real alternative to complete genome sequencing is to sequence only a sufficiently large genetic sample to generate an informative data set. This has been the standard approach for population analyses restricted to modern data. When a sufficient number of polymorphic positions is probed, such as in a recent analysis of 1 million SNPs sequenced from 3,000 Europeans [27], a remarkably detailed picture can emerge. Because only short fragments are required, SNP typing may prove to be highly amenable to ancient DNA data. Problematically, SNP analyses may be affected by ascertainment bias, emphasized by the recently published African genomes within which a large number of so-far-unknown SNPs were discovered. While this is likely to be less of a problem in regions outside Africa, it remains unknown how increasing the temporal diversity of human samples will be affected by ascertainment bias.

An alternative to using SNP typing arrays is to target a subset of the genome so that homologous data are obtained for all samples. While this is not possible using PCR, DNA capture *via* hybridization coupled to NGS is a feasible approach [28, 29]. In fact this approach has been used to investigate the protein coding regions of additional individuals in the African genome publication [19]. Although initially developed for studying modern DNA, some variants of hybridization capture have already been shown to be effective for the analysis of ancient DNA, at least with regard to the mitochondrial genome [30, 31]. However, as the sensitivity of this approach seems to surpass that of PCR, most likely because, in contrast to PCR, it allows the analysis of very short fragments, there is every reason to believe that it should work on ancient nuclear DNA as well.

Conclusion

The sequencing of a complete human genome using ancient DNA – in fact the first high coverage genome obtained from a fossil – by Rasmussen *et al.* marks another milestone in the analysis of ancient DNA. However, as at the beginning of ancient DNA research a quarter of a century ago, there remains the pressing question as to how the technologic leaps achieved can be turned into progress in biologic knowledge – at an affordable price. Despite the rapidly increasing DNA sequencing throughput achievable using NGS technologies, obtaining complete genomes from a large number of ancient samples still does not seem to be a viable option for the near future. To tackle questions about the affinities of extinct populations to modern humans, further developments will be required. For ancient human research, a major challenge will be to develop methods to detect and account for sample contamination by modern human DNA, in particular where the ancient specimen is expected to have a genome sequence that closely resembles a modern human genome. Most likely, these advances lie in the adoption of DNA capture methods coupled to NGS. Such an approach, feasible even with the characteristically short sequence fragments of ancient specimens, will allow homologous data to be obtained for multiple individuals from broad spatial and temporal samples. Thus, the field of ancient DNA may finally reach its promised potential and allow unique insights into the evolutionary history both of our own and of other species.

References

1. **Margulies M**, **Egholm M**, **Altman WE**, *et al*. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376–80.
2. **Poinar HN**, **Schwarz C**, **Qi J**, *et al*. 2006. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311**: 392–4.
3. **Miller W**, **Drautz DI**, **Ratan A**, *et al*. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**: 387–90.
4. **Pääbo S.** 1985. Molecular cloning of ancient Egyptian mummy DNA. *Nature* **314**: 644–5.
5. **Adcock GJ**, **Dennis ES**, **Easteal S**, *et al*. 2001. Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proc Natl Acad Sci USA* **98**: 537–42.
6. **Cooper A**, **Rambaut A**, **Macaulay V**, *et al*. 2001. Human origins and ancient human DNA. *Science* **292**: 1655–6.
7. **Del Pozzo G**, **Guardiola J.** 1989. Mummy DNA fragment identified. *Nature* **339**: 431–2.
8. **Pääbo S**, **Poinar H**, **Serre D**, *et al*. 2004. Genetic analyses from ancient DNA. *Annu Rev Genet* **38**: 645–79.
9. **Green RE**, **Krause J**, **Ptak SE**, *et al*. 2006. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**: 330–6.
10. **Noonan JP**, **Coop G**, **Kudaravalli S**, *et al*. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* **314**: 1113–8.
11. **Wall JD**, **Kim SK.** 2007. Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genet* **3**: 1862–6.
12. **Green RE**, **Briggs AW**, **Krause J**, *et al*. 2009. The Neandertal genome and ancient DNA authenticity. *EMBO J* **28**: 2494–502.
13. **Rasmussen M**, **Li Y**, **Lindgreen S**, *et al*. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* **463**: 757–62.
14. **Gilbert MTP**, **Wilson AS**, **Bunce M**, *et al*. 2004. Ancient mitochondrial DNA from hair. *Curr Biol* **14**: R463–4.
15. **Gilbert MT**, **Kivisild T**, **Gronnow B**, *et al*. 2008. Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. *Science* **320**: 1787–9.
16. **Gilbert MT**, **Tomsho LP**, **Rendulic S**, *et al*. 2007. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* **317**: 1927–30.
17. **Pennisi E.** 2009. Neandertal genomics. Tales of a prehistoric human genome. *Science* **323**: 866–71.
18. **Green RE**, **Malaspinas AS**, **Krause J**, *et al*. 2008. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**: 416–26.
19. **Schuster SC**, **Miller W**, **Ratan A**, *et al*. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**: 943–7.
20. **Tishkoff SA**, **Reed FA**, **Friedlaender FR**, *et al*. 2009. The genetic structure and history of Africans and African Americans. *Science* **324**: 1035–44.
21. **Gonzalez S**, **Jimenez-Lopez JC**, **Hedges R**, *et al*. 2003. Earliest humans in the Americas: new evidence from Mexico. *J Hum Evol* **44**: 379–87.
22. **Bramanti B**, **Thomas MG**, **Haak W**, *et al*. 2009. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *Science* **326**: 137–40.
23. **Haak W**, **Forster P**, **Bramanti B**, *et al*. 2005. Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* **310**: 1016–8.
24. **Malmstrom H**, **Gilbert MT**, **Thomas MG**, *et al*. 2009. Ancient DNA reveals lack of continuity between neolithic hunter-gatherers and contemporary Scandinavians. *Curr Biol* **19**: 1758–62.
25. **Noonan JP**, **Hofreiter M**, **Smith D**, *et al*. 2005. Genomic sequencing of Pleistocene cave bears. *Science* **309**: 597–9.
26. **Dalton R.** 2010. Palaeogenetics: Icy resolve. *Nature* **463**: 724–5.
27. **Novembre J**, **Johnson T**, **Bryc K**, *et al*. 2008. Genes mirror geography within Europe. *Nature* **456**: 98–101.
28. **Hodges E**, **Rooks M**, **Xuan Z**, *et al*. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc* **4**: 960–74.
29. **Hodges E**, **Xuan Z**, **Balija V**, *et al*. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat Genet* **39**: 1522–7.
30. **Briggs AW**, **Good JM**, **Green RE**, *et al*. 2009. Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325**: 318–21.
31. **Krause J**, **Briggs AW**, **Kircher M**, *et al*. 2010. A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr Biol* **20**: 231–6.
32. **Higuchi R**, **Bowman B**, **Freiberger M**, *et al*. 1984. DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312**: 282–4.