V.15

Ancient DNA

Beth Shapiro

OUTLINE

1. Beginnings
2. The importance of being clean
3. Name that bone: Inserting extinct species into molecular phylogenies
4. Ancient population genetics and phylogeography
5. Ancient genomics
6. The future of ancient DNA

Ancient DNA is a field of molecular evolutionary biology that uses DNA sequence data recovered from poorly preserved organisms, usually deceased for hundreds to hundreds of thousands of years. Ancient DNA data can provide unique snapshots in time to better understand how populations and species evolve. The field was born in the early 1980s, when the first ancient DNA sequences were recovered from preserved muscle of a quagga, a relative of the zebra, which had been extinct for nearly 100 years. Although the early days of ancient DNA were marked by a few spectacular but flawed results, the field has matured into a robust, internally rigorous scientific pursuit with the potential to provide real insight into the mechanisms of evolution at both the species and the population level. Ancient DNA has benefited in particular from recent advances in high-throughput sequencing technologies and from the development of analytical techniques that take advantage of the evolutionary information gained by sampling genetic data over both space and time.

GLOSSARY

Ancient DNA. A field of biology that involves extracting and manipulating sequence data from samples that are old and decayed in some way.

Contaminating DNA. DNA introduced into an experiment from the preservation environment, from excavation, sample handling, or sample processing, or during the experiment itself.

Coprolite. Preserved feces.

Draft Genome. Genomes of ancient DNA published before being considered sufficient in quality to be called "complete."

Mitochondrial DNA (mtDNA). A separate DNA genome of the *mitochondria*, which are maternally inherited organelles found within every cell.

Polymerase Chain Reaction (PCR). An enzymatic technique for amplifying from one to a few copies of DNA by several orders of magnitude.

Postmortem Decay. The DNA damage that accumulates after an organism's death.

1. BEGINNINGS

In 1984, a team of researchers based mostly in Allan Wilson's laboratory at the University of California, Berkeley, cloned two short fragments of mitochondrial DNA (mtDNA) from dried muscle taken from a 140-year-old museum specimen of a quagga (*Equus quagga*), a relative of the zebra that had been extinct since 1883. This work was the first to describe DNA preserved in nonliving tissues in a mainstream scientific journal. It came three years after a Chinese-language publication reported sequences cloned from a mummified human liver, and at the same time a German-language publication described the recovery of DNA from several Egyptian mummies. The quagga work confirmed that preserved tissues contained amplifiable DNA sequences. The results captured international attention, heralding great enthusiasm for this new source of DNA and a race to sequence the oldest, most exciting extinct organism. Crucially, the quagga study also noted what remains the most pervasive problem in the field of ancient DNA: that very little DNA survives postmortem.

An early leader in the field and widely considered "the father of ancient DNA," Svante Pääbo began his work with the aim of genetically characterizing the evolutionary history of Egyptian mummies. The process of rapid

desiccation to which the bodies had been subjected immediately after death should have left the DNA molecules in a relatively intact form, making them ideal for ancient DNA analysis. In 1985, he recovered two members of the *Alu* family of human repetitive DNA sequences from a 2400-year-old Egyptian mummy. Although DNA could be recovered from only one of the 23 mummies he tested, close inspection of the data led Pääbo to conclude that few changes had occurred in the DNA postmortem.

A few years later, the polymerase chain reaction (PCR) was invented. This reaction makes millions of copies from only one or a few starting molecules of DNA; it also allows specific DNA sequences to be targeted, providing the means for focused evolutionary research. For ancient DNA, another advantage of PCR was the capability of a more thorough assessment of the ancient sequences. The enzyme used in the PCR to copy DNA was thought to read through undamaged molecules only, so that when errors were encountered, the reaction would simply end. In contrast, the enzymes that had been used during molecular cloning maintained the ability to repair damaged DNA, and this repair process could potentially introduce errors into the ancient sequences. When the same fragment of quagga DNA was amplified using the PCR, the two differences found between the quagga and its closest relative, the plains zebra, turned out to be no different at all. Damage, it seemed, was going to be a problem.

In 1989, Pääbo used the PCR to assess DNA survival in differently aged remains collected from a variety of locations. These results were instrumental in securing a place for ancient DNA as a credible scientific endeavor while warning future practitioners of the specific challenges associated with working with ancient material. He showed that ancient DNA sequences contain chemical modifications, including strand breaks, DNA crosslinks, and modified bases, that make their recovery challenging. He proposed an inverse relationship between fragment length and the number of surviving molecules of that length. He noted that DNA preservation is not determined by specimen age but by the environment in which the specimen was preserved. And crucially, he pointed out that contamination by modern DNA is likely to be the most serious challenge of working with ancient specimens. All these observations remain relevant to ancient DNA research today.

2. THE IMPORTANCE OF BEING CLEAN

DNA damage and contamination are the two biggest problems facing ancient DNA researchers. Initially, degradation occurs through the action of endogenous nucleases. In some circumstances, including rapid desiccation or deposition in very cold, dry, or salty environments, these enzymes will themselves be degraded before they can destroy all the DNA; however, even in ideal circumstances, environmental processes such as exposure to oxygen and water will slowly but steadily break down the surviving DNA until what remains is too damaged or fragmentary to be useful. Eventually, the continuous breakdown of DNA will result in only a few surviving, nonfragmented molecules per sample. The most common form of hydrolytic damage in ancient DNA specimens is deamination, in particular the conversion of cytosine to uracil. This results in the template DNA being read as a thymine, rather than cytosine, and in the erroneous incorporation of an adenine in the complementary strand. Although the exact numbers are still a matter of debate, this form of DNA damage is thought to account for nearly all misincorporated bases observed in amplified ancient DNA sequences. In addition to base misincorporations, double-strand breaks and DNA crosslinks are both common in ancient DNA samples; both lead to the amplification of only very short fragments of DNA. Expectations about DNA damage, and in particular the observation of cytosine deamination, are now used to distinguish authentic ancient DNA from contaminating DNA, and new phylogenetic models use information about damage to estimate the probability that certain mutations are due to decay rather than to evolution.

A variety of experiential protocols have been suggested to minimize the impact of DNA damage and contamination. These range from common laboratory sense, including wearing protective clothing and sterilizing components and work surfaces, to experimentally rigorous procedures, such as using multiple negative controls, performing independent PCRs to generate consensus sequences, and cloning PCR products to detect damage and contamination. Most laboratories comprise two separate, geographically isolated facilities: one in which the DNA extraction is performed and PCR is set up, and another for downstream (post-PCR) molecular biology work. This, in combination with a streamlined daily workflow in which researchers never move from the modern to the ancient lab, helps to ensure that amplified ancient DNA does not itself become a contaminant. These protocols have been modified over the years as technologies advance and as more is learned about how DNA degrades. For example, a requirement that was widely adopted in 2000, that each ancient DNA sequence be independently replicated in a separate laboratory, has been largely abandoned as high-throughput sequencing and population sampling have become more common, and hence contamination easier to identify. There is no doubt, however, that ancient DNA is sensitive to postmortem damage and contamination, and that care should be taken to ensure that published results are authentic.

Before these protocols were put in place and their importance made clear, the race to publish the oldest

DNA produced several results subsequently shown to be false. This period during the 1990s can be considered the "dark days" of ancient DNA. First, a 790-base-pair (bp) fragment of chloroplast DNA from a 16-million-year-old magnolia leaf was published. These data were met initially with skepticism, as it was unclear that DNA should survive for that long, even in the best possible conditions. Not surprisingly, these results were soon shown to be derived from contaminating bacteria. Next, bacterial DNA sequences and DNA from insects were reported from pieces of amber 25–120 million years old. Despite repeated attempts, these results could not be reproduced, but popular culture was already inspired, spawning movies such as *Jurassic Park* in 1993. In 1994, the first dinosaur DNA was published, purportedly isolated from a fossilized 80-million-year-old bone. To achieve this, researchers performed 2880 PCRs on two DNA extracts from the same bone, resulting in amplification of nine 170-bp fragments of mitochondrial DNA. Reanalysis of these fragments by different groups showed them to be mammalian in origin, and most likely a human sequence that, at the time, was not available in a public database for comparison.

These early missteps have not stopped ancient DNA researchers from pushing the technique's temporal limit. During and since the dark days, several hundred-million-year-old bacterial sequences from rocks excavated from salt mines have been published, as have Miocene plant DNA, and protein sequences from two different dinosaur bones. None of these results have been independently replicated, and many remain skeptical about their authenticity; nonetheless, DNA sequences have been recovered and authenticated that are very old, including sequences from 100,000-year-old bones preserved in arctic permafrost and temperate caves. The oldest authenticated DNA published so far are bacterial sequences recovered from permafrost ice cores between 450,000 and 800,000 years old. Importantly, these very old sequences have all been recovered from geographic regions where the preservation conditions favor long-term DNA survival.

3. NAME THAT BONE: INSERTING EXTINCT SPECIES INTO MOLECULAR PHYLOGENIES

At the same time these spectacular mistakes were captivating public and scientific attention, progress was being made in the extraction and investigation of authentic ancient DNA. For example, amplification and sequencing of two mitochondrial fragments of the recently extinct Tasmanian wolf confirmed that the Tasmanian wolf was more closely related to Australian marsupial carnivores than to more similar-looking marsupial carnivores from South America, indicating that their shared morphological features must have evolved independently. Thus began a period of genetically characterizing extinct organisms that continues today.

Ancient DNA has been used to place many extinct species in molecular phylogenies (see Section II: Phylogenetics and the History of Life). One of the earliest results was to reveal that mammoths are, rather unsurprisingly, closely related to elephants. Complete mitochondrial genomes of mammoths and mastodons later resolved this relationship further, revealing that mammoths are more closely related to Asian elephants than to African elephants (figure 1). Ancient DNA isolated from the Oxford dodo showed that this international emblem of extinction was a type of pigeon, rather than in its own evolutionary lineage as previously believed. In another revisionary discovery, ancient nuclear DNA recovered from the remains of several extinct New Zealand moa revealed that the three described species were in fact only two, and that the vast size difference used to distinguish the species from each other was actually due to pronounced sexual dimorphism.

Ancient DNA can be recovered from any element that has been shown to contain DNA, commonly bone, teeth, hair, seeds, muscle, or eggshells. In addition to these individual-specific tissues, DNA can also be recovered from mixed materials such as coprolites and soil. Despite being exposed to more damage-inducing influences than DNA preserved within bones, coprolite-recovered sequences are as reliable as those produced from bone. More interestingly, DNA extracted from coprolites provides both genetic information about the defecator and a genetic survey of that individual's last few meals. Sedimentary DNA, likely from a combination of shed cells and decaying plant material, provides the means to characterize ancient communities in the absence of macrofossil remains, circumventing potential problems with differential survival of representative members of the extinct community, wherein certain species may not leave large numbers of fossils, or certain environments may not be amenable to the long-term preservation of DNA. The DNA present in soil is generally "naked," not bound to anything and therefore not protected from bombardment by damage-inducing environmental events. This makes it difficult to distinguish damage lesions from phylogenetically informative mutations in the recovered sequences. Nonetheless, sedimentary DNA does make it possible to identify when and where species were present, potentially extending the range of locations and species that can be studied using ancient DNA.

4. ANCIENT POPULATION GENETICS AND PHYLOGEOGRAPHY

Most of the studies mentioned above use material recovered either from caves, where the ambient moisture

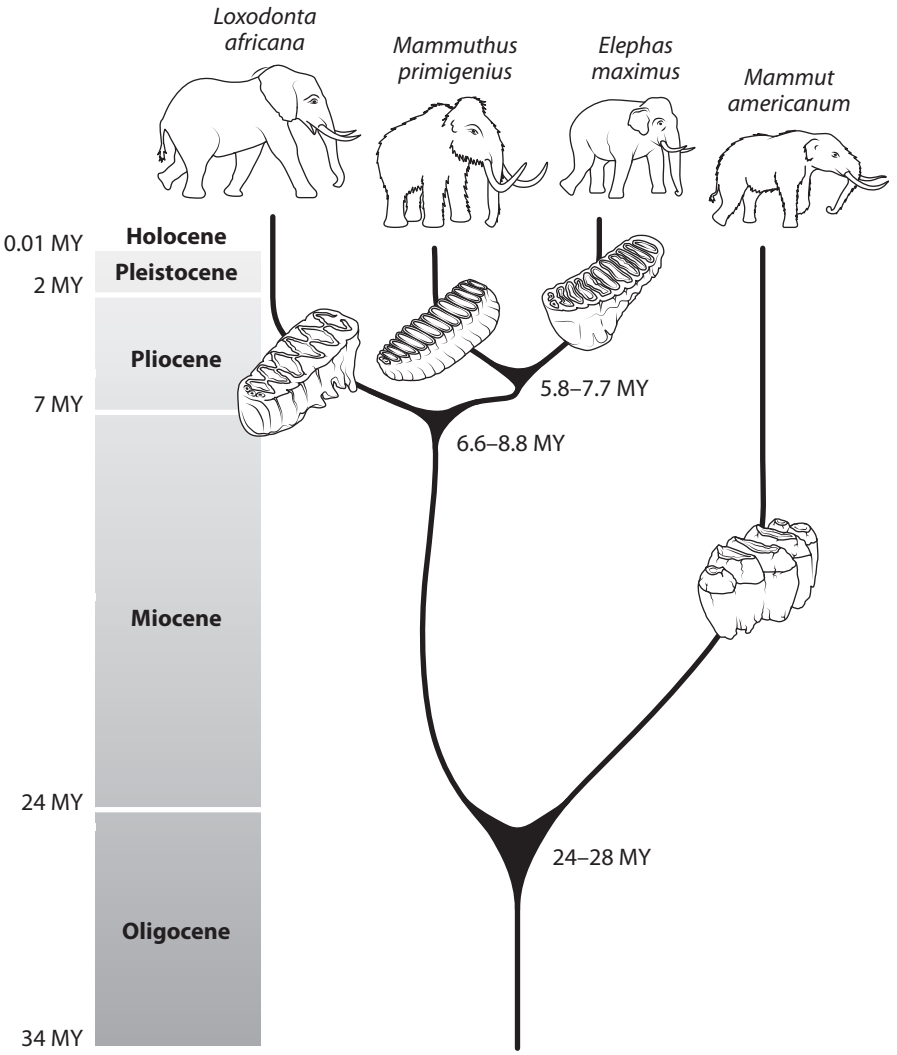

Figure 1. Complete mitochondrial genomes have now been sequenced for the extinct mammoth and mastodon, as well as for both living elephant species. These data have been used to infer both the branching order of the elephantid phylogeny (revealing that mammoths are more closely related to Asian elephants than to African elephants) and the timing of diversification between the different lineages. Elephants are only one example of lineages for which the addition of ancient DNA data has provided significant resolving power for long-standing phylogenetic and taxonomic questions. Other lineages include pigeons, ratites, cow, and even humans. (Reproduced from Rohland et al. 2007.)

and temperature tend to remain constant in both the short and long term, or from the Arctic, where remains are preserved in permanently frozen soil (permafrost). While authentic ancient DNA has been recovered from warmer climates (e.g., Florida, the Caribbean), consistently cold places (e.g., Siberia) are by far the richest source of material for ancient DNA analysis. This perhaps explains why, as the field moves toward analyzing populations of individuals rather than single individuals representing an extinct species, the focus has mostly been on Arctic and cave-dwelling species.

The first analyses of changes in genetic diversity within populations through time, however, took advantage of younger, museum-preserved skins that were more likely better preserved than ice age bones. One of the first population-level analyses amplified mtDNA from skins of three geographically isolated populations of the Panamint kangaroo rat in California collected early in the twentieth century. These data skins were compared with data collected from the same localities but in 1988; surprisingly, the populations had remained genetically isolated from each other throughout the period spanning

the sample ages. Soon after, it was shown that pocket gophers from Yellowstone National Park have been genetically isolated from nearby populations for at least 2400 years. Later, mtDNA isolated from rabbit remains from across Europe and North Africa showed that these populations had maintained genetic stability and strong population structure for at least 11,000 years.

This pattern was not found for all species, however. Sequencing of mtDNA isolated from seven permafrost-preserved, Alaskan brown bear bones showed that the existing geographic isolation between brown bear mitochondrial lineages was established 15,000 years ago, and that prior to this time a different geographic pattern prevailed. This work was later expanded to include 36 brown bears ranging from 2000 to more than 50,000 years old. The new results supported the original findings and identified four distinct temporal periods during which the geographic distribution of brown bear mitochondrial lineages had remained stable, with rapid changes occurring between these. Climate change, in particular that linked to the last ice age, was implicated as the driver of most of these demographic changes.

The analysis of population genetic data sampled over time became more common as statistical tools capable of taking advantage of this kind of data were developed. Two complementary approaches were introduced approximately simultaneously, and both have been used to test the role of environmental change in driving changes in population genetic diversity. A major question within the reach of ancient DNA is, What caused the extinction of the ice age megafauna 8000 years ago? New genetic analysis methods allow a full-probabilistic estimation of the demographic history of a set of sampled DNA sequences, enabling the first attempt to answer to this question. The first large, ancient DNA population data set contained more than 600 sequences from North American bison, ranging from only 100 to more than 55,000 years old. These data showed clear evidence for a peak in bison diversity around 35,000 years ago, followed by a rapid decline toward extinction. The timing of the beginning of this decline was surprising, as it predated both the peak of the last ice age and the first appearance of large numbers of humans in North America, two competing hypotheses about the cause of the mass extinction. Later, more sophisticated demographic models further resolved the bison demographic history, revealing that around 13,000 years ago, bison narrowly escaped extinction; this bottleneck was followed by rapid recovery of the genetic diversity that persists today.

Work on this question continues, and population data sets now exist for six herbivores and several carnivores. All of these seem to respond to changes in climate differently, depending on their particular habitat requirements. Horses, for example, peak in genetic diversity slightly after bison peak in North America, probably because they were better able to survive once the steppe grasslands began to disappear at the onset of the last ice age. Although the jury is out regarding the ultimate cause of these extinctions, it is clear that climate change played a major role.

The second approach to analyzing ancient population genetic data takes advantage of the approximate Bayesian computation (ABC) framework. In this framework, genetic data are simulated under proposed models of population evolution and compared to those estimated from real data to identify the most likely demographic scenario. A major breakthrough came with another program that allowed simulated data sets to mimic ancient DNA data sets, in that samples could be drawn from an evolving population at different points in time. This approach was used to show that 3000 years ago, the Argentinean colonial tuco-tuco, a subterranean rodent, suffered a severe population bottleneck in which it lost around 99.7 percent of its mitochondrial genetic diversity.

5. ANCIENT GENOMICS

The recent technological advances collectively known as "next-generation sequencing" have been embraced by the ancient DNA community. These technologies allow millions of sequencing reactions to happen in parallel by creating microreactors and/or attaching DNA molecules to solid surfaces or beads prior to sequencing. These technologies provide a means to explore more fully the amount and quality of DNA preserved in ancient specimens. They also make it feasible to obtain larger amounts of ancient data in a much less time-consuming and often less expensive way than using traditional approaches.

The first complete ancient genomes were published in 2001, long before next-generation sequencing was state of the art. Two teams working independently both published mitochondrial genomes from two species of moa. Each 17,000-bp mitochondrial genome was painstakingly pieced together from overlapping 350–600 bp fragments amplified via PCR. These genomes were proof that ancient genomics is feasible, and could provide useful evolutionary information. The moa genomes were used to estimate the timing of the divergence between ratite birds and provide a temporal framework for the breakup of Gondwana into smaller continental fragments (these eventually became most of the landmasses found in today's Southern Hemisphere, as well as a few landmasses that migrated further north).

Five years later, two complete mitochondrial genomes of the mammoth were published using different techniques. One group pieced together the mammoth mitochondrial genome by targeting only longer, intact fragments, between 1200 and 1600 base pairs in length. A second developed a multiplex PCR approach to coamplify

nonoverlapping fragments of mammoth mitochondrial DNA in a single PCR, greatly speeding up the process of data generation and significantly reducing the amount of sample required to perform the experiment. In the same year, a third group took mammoth mitochondrial-genome sequencing into next generation. They used the Roche 454 technology to shotgun sequence a permafrost-preserved mammoth bone. Of the 13 million base pairs of mammoth DNA they recovered, 222 reads, each around 89 bp long, mapped to the mammoth mitochondrial genome.

In a shotgun-sequencing approach, all the DNA extracted from a particular specimen is made into a library, and that DNA library is then sequenced. As a result, sequences are generated not only from the target specimen but also from any bacteria or other organisms that may have colonized the sample during its preservation history, and any DNA that may have contaminated the sample during processing. The sample used in this first study was remarkably well preserved: 45.4 percent of the sequences from the genomic library were identified as mammoth DNA, the remainder likely coming from organisms colonizing the sample after its deposition. In contrast, the libraries that were later used to sequence the complete nuclear genome of the Neanderthal contained only 1–5 percent Neanderthal DNA. In this case, enzymes targeting specific sequences present in bacterial DNA (and absent from Neanderthal DNA) were used to chop up bacterial DNA in the DNA libraries, thereby increasing the ratio of Neanderthal to contaminating DNA. Draft ancient genomes have now been published for a mammoth, a 4000-year-old Paleoeskimo from Greenland, a Neanderthal, and a previously unknown hominin from Denisova Cave in Siberia.

6. THE FUTURE OF ANCIENT DNA

Although the field of ancient DNA is now more than 25 years old, its potential is only beginning to be realized. After the excitement of simply generating complete ancient genomes fades, a new era of ancient DNA research is likely to emerge, in which the unique perspective allowed by ancient DNA is fully recognized. While it is impossible to know what the next discovery will be, two questions stand out.

First, what makes species unique? With the publication of the Neanderthal genome, we now have much more information about precisely which mutations distinguish us from our closest living relative, the chimpanzee (see chapter II.18). Prior to 2010, Neanderthals and humans were known to share a derived allele at the FOXP2 locus, which is involved in speech and language, suggesting that a selective sweep (see chapter V.14) occurred in this region prior to the divergence between Neanderthals and humans. The draft Neanderthal genome revealed large genomic regions that have been under positive selection *since* our divergence from Neanderthals. These regions include genes associated with human-specific maladies, including autism spectrum disorder and type 2 diabetes. Learning more about these genomic regions may reveal much about what it means to be human. Methods to target and capture specific regions of DNA provide a promising route to refining these observations and improving our understanding of what makes species look and act the way they do.

Second, what is the role of environmental change in the maintenance and distribution of genetic diversity? Shotgun sequencing hundreds of individuals for population genomic analyses is still too expensive; however, approaches are in development to capture specific fragments of DNA from DNA libraries. These captured fragments can then be bar-coded, pooled, and sequenced together on a next-generation platform. This approach allows hundreds or thousands of loci to be sequenced simultaneously from hundreds of individuals. It provides a solution to the matrilineal bias of using only mitochondrial DNA, and much more power to detect changes in genetic diversity associated with either particular environmental events or episodes of natural selection.

The results of analyses incorporating ancient DNA data have ranged from obvious (that a mammoth is closely related to an elephant) to surprising (that all non-African humans still contain some Neanderthal DNA). Regardless of what happens in the next 25 years, it is clear that the perspective gained from these data has benefited many aspects of evolutionary research. We know much more about the evolution of life on earth, about how populations respond to climate change, and about our own, recent evolutionary history than we could have known without ancient DNA.

FURTHER READING

Bunce, M., T. H. Worthy, T. Ford, W. Hoppitt, E. Willerslev, A. Drummond, and A. Cooper. 2003. Extreme reversed sexual size dimorphism in the extinct New Zealand moa *Dinornis*. Nature 425: 172–175. *This paper shows that ancient DNA can be used to resolve long-standing taxonomic issues. What were once thought three species of Dinornis are actually only one, with very large females and smaller males.*

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, et al. 2010. A draft sequence of the Neandertal genome. Science 328: 710–722. *This paper provide details of the draft genome sequence of a female Neanderthal and provides evidence of interbreeding between Neanderthals and anatomically modern* Homo sapiens *after* H. sapiens *left Africa.*

Lorenzen, E. D., D. Nogués-Bravo, L. Orlando, J. Weinstock, J. Binladen, K. A. Marske, A. Ugan, et al. 2011. Species specific responses of Late Quaternary megafauna to

climate and humans. Nature 479: 359–364. *This paper provides a detailed assessment of changes in distribution and abundance of six large herbivores (mammoth, bison, horse, caribou, musk ox, and woolly rhino) over the last 50,000 years, including both full-probabilistic and ABC-based inference of changes in population size and an analysis of changes in the distribution of appropriate habitat for each species over this time frame.*

Pääbo, S., H. Poinar, D. Serre, V. Jaenicke-Després, J. Hebler, N. Rohland, M. Kuch, et al. 2004. Genetic analyses from ancient DNA. Annual Review of Genetics 38: 645–678. *This paper contains a concise review of the history and challenges associated with working with ancient DNA specimens and data.*

Poinar, H, C. Schwarz, J. Qi, B. Shapiro, R.D.E. MacPhee, B. Buigues, A. Tikhonov, et al. 2006. Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. Science 311: 392–394. *This manuscript provides the first example of use of next-generation sequencing technology to sequence an ancient specimen.*

Rohland, N., A.-S. Malaspinas, J. L. Pollack, M. Slatkin, P. Matheus, and M. Hofreiter. 2007. Proboscidean mitogenomics: Chronology and mode of elephant evolution using mastodon as out-group. PLoS Biology 5: e207. *This manuscript uses ancient DNA data to infer the phylogeny of the elephants, and demonstrates conclusively that the addition of ancient DNA data can considerably improve knowledge of the rate and timing of species divergence.*

Willerslev, E., A. J. Hansen, J. Binladen, T. B. Brand, M.T.P. Gilbert, B. Shapiro, M. Bunce, et al. 2003. Diverse plant and animal genetic records from Holocene and Pleistocene sediments. Science 300: 791–795. *This manuscript describes the amplification of both plant and animal ancient DNA directly from Alaskan permafrost sediments.*