Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA

Hendrik N. Poinar,[1,2,3]* Carsten Schwarz,[1,2] Ji Qi,[4] Beth Shapiro,[5] Ross D. E. MacPhee,[6] Bernard Buigues,[7] Alexei Tikhonov,[8] Daniel H. Huson,[9] Lynn P. Tomsho,[4] Alexander Auch,[9] Markus Rampp,[10] Webb Miller,[4] Stephan C. Schuster[4]*

We sequenced 28 million base pairs of DNA in a metagenomics approach, using a woolly mammoth (*Mammuthus primigenius*) sample from Siberia. As a result of exceptional sample preservation and the use of a recently developed emulsion polymerase chain reaction and pyrosequencing technique, 13 million base pairs (45.4%) of the sequencing reads were identified as mammoth DNA. Sequence identity between our data and African elephant (*Loxodonta africana*) was 98.55%, consistent with a paleontologically based divergence date of 5 to 6 million years. The sample includes a surprisingly small diversity of environmental DNAs. The high percentage of endogenous DNA recoverable from this single mammoth would allow for completion of its genome, unleashing the field of paleogenomics.

Complete genome sequences of extinct species will answer long-standing questions in molecular evolution and allow us to tackle the molecular basis of speciation, temporal stages of gene evolution, and intermediates of selection during domestication. To date, fossil remains have yielded little genetic insight into evolutionary processes because of poor preservation of their DNA and our limited ability to retrieve nuclear DNA (nDNA). Most DNA extracted from fossil remains is truncated into fragments of very short length [<300 base pairs (bp)] from hydrolysis of the DNA backbone, cross-linking due to condensation (*1*, *2*), and oxidation of pyrimidines (*3*), which prevents extension by *Taq* DNA polymerase during polymerase chain reaction (PCR). In addition, DNA extracts are a mixture of bacterial, fungal, and often human contaminants, complicating the isolation of endogenous DNA. In the past, these problems could only be indirectly overcome by concentrating on the small number of genes present on the maternally inherited mitochondrial genome, which is present in high copy number in animal cells. This approach severely limits access to the storehouses of genetic information potentially available in fossils of now-extinct species. In a few rare cases, investigators have managed to isolate and characterize nuclear DNA from fossil remains preserved in arid cave deposits (*4*–*6*) or, more commonly, permafrost-dominated environments (*7*, *8*) and ice (*9*), where the average burial temperature can be as low as –10°C (*10*). Under these conditions, preservation is enhanced by reduced reaction rates: In permafrost settings, theoretical calculations predict DNA fragment survival up to 1 million years (*11*, *12*).

Although more nuclear DNA is present in cells than mitochondrial DNA, access to the nuclear genome even in well-preserved fossil material remains difficult with PCR-based approaches, which must target known sequences from specific genes. To find the ideal sample and analytical approach for paleogenomics, we screened eight of the morphologically best preserved mammoth remains in the collections of the Mammoth Museum, a dedicated permafrost "ice cave" facility in the town of Khatanga in the southeastern part of the Taimyr Peninsula, Russian Federation, and maintained by Cerpolex/Mammuthus Expeditions (CME). We extracted DNA from these samples, using ancient DNA methodology, to avoid to the greatest extent possible exogenous DNA (*13*), because we are well aware of the problems and pitfalls associated with potential contamination. These samples were screened with a quantitative PCR (qPCR) assay designed for the mammoth mitochondrial cytochrome *b* gene (*14*), with primers designed to amplify mitochondrial DNA molecules from both African and Asian elephants (*Loxodonta africana* and *Elephas maximus*). We quantitated the number of amplifiable mitochondrial DNA (mtDNA) molecules of 84 bp in length. The eight samples ranged from 1×10^6 copies per gram to 96×10^6 copies per gram, all excellent samples as judged by ancient DNA standards (*15*). However, one sample in particular, CME 2005/915, was exceptional. The specimen, an edentulous mandible dated to 27,740 ± 220 ^{14}C years before the present (uncorrected; Beta 210777), was recovered on the shore of Baikura-turku, a large bay on the southeastern side of Lake Taimyr, the largest freshwater body in Eurasia north of the Arctic Circle. The large numbers and high quality of late Quaternary fossils recovered from the Taimyr Peninsula have prompted several major investigations in recent years (*16*), including studies of ancient DNA (*14*, *17*, *18*). Taimyr's extremely cold winters, combined with short, cool summers and little annual precipitation, have ensured that conditions optimal for the preservation of bones and teeth prevailed there for most or all of the late Pleistocene.

To obtain a better perspective on the preservational integrity of this sample, we quantitated the number of amplifiable mtDNA cytochrome *b* gene fragments of increasing length up to 1 Kb (fig. S1) (*13*). The DNA extract (from 58 mg bone) contained $\sim 7 \times 10^6$ copies of the 84-bp mtDNA fragment in 100 μl or 121×10^6 copies per gram of bone (*13*). On the basis of this, we estimated that the concentration of total amplifiable DNA in our sample (assuming a copy-number ratio of 1000:1, mtDNA:nDNA) was on the order of 0.73 μg of mammoth DNA per gram of bone. We extracted 1 g of bone and concentrated the DNA to 100 μl, which was subsequently used for library construction and sequencing technology that recently became available (*13*, *19*). This in vitro technique circumvents amplification or cloning biases by compartmentalizing single DNA molecules before the amplification step in a lipid vesicle, thereby maintaining the original DNA template distribution. The lipid-enclosed single DNA molecule, attached to a sepharose bead 28 μm in diameter, undergoes a PCR reaction, yielding sufficient DNA copies for sequencing. Sequencing is performed with a pyrosequencing methodology (*19*).

We obtained 302,692 sequence reads averaging 95 bp, with read length being limited by the sequencing approach rather than by the 630-bp average DNA fragment size obtained from the extract after shearing. The total sequence data produced for the bone metagenome was 28×10^6 bp. We aligned the sequencing reads with current (November 2005) assemblies of the genome sequences of African elephant (*L. africana*), human, and dog (*Canis familiaris*), downloaded from www.genome.ucsc.edu (Table 1 and Fig. 1). Alignments were computed by the program BLASTZ (*20*), with parameters chosen to identify only regions of about 90% identity or higher (*13*). As we have not detected any convincing mappings of a read to the human Y chromosome, despite random distribution across the genome, we conclude that our mammoth was a female.

A total of 137,527, or 45.4% of all reads, aligned to the African elephant genome (Fig. 1, Table 1) (*13*), currently available at 2.2-fold coverage, with an estimated number of base pairs in the genome of 2.3×10^9 bp (www.broad.mit.edu/mammals/#chart). A twofold cov-

[1]McMaster Ancient DNA Center, [2]Department of Anthropology, [3]Department of Pathology and Molecular Medicine, McMaster University, 1280 Main Street West, Hamilton ON, L8S 4L9 Canada. [4]Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, 310 Wartik Building, University Park, PA 16802, USA. [5]Henry Wellcome Ancient Biomolecules Centre, Department of Zoology, Oxford University, South Parks Road, Oxford, OX1 3PS, UK. [6]Division of Vertebrate Zoology/Mammalogy, American Museum of Natural History, 79th Street and Central Park West, New York, NY 10024, USA. [7]#2 Avenue de la Pelouse, F-94160 St. Mandé, France. [8]Zoological Institute, Russian Academy of Sciences, Universitetskaya nab.1, Saint Petersburg 199034, Russia. [9]Center for Bioinformatics (ZBIT), Institute for Computer Science, Tübingen University, 72076 Tübingen, Germany. [10]Garching Computing Center (RZG), Boltzmannstrasse 2, D-85748 Garching, Germany.

*To whom correspondence should be addressed. E-mail: poinarh@mcmaster.ca (H.N.P.); scs@bx.psu.edu (S.C.S.)

Table 1. Total percent of aligned reads and their relative identities to African elephant, human, and dog.

	Elephant	Human	Dog
Total no. reads	302,692 (100%)	302,692 (100%)	302,692 (100%)
Aligned reads	137,527 (45.4%)	4,237 (1.4%)	3,775 (1.2%)
Uniquely aligning reads	44,442 (14.7%)	3,901 (1.3%)	3,548 (1.2%)
Multiply aligned reads	93,085 (30.8%)	336 (0.1%)	227 (0.1%)
Reads with at least 95% identity	90,507 (30.0%)	1,184 (0.4%)	1,140 (0.4%)
Reads with 100% identity	21,952 (7.3%)	116 (0.04%)	142 (0.05%)
Uniquely aligning base pairs	4,332,350	318,966	291,714
Identity in unique alignments	98.55%	92.68%	92.91%
Mitochondrial reads	209	–	–
Identity in mitochondrial reads	95.93%	–	–
Mitochondrial base pairs	16,419	–	–

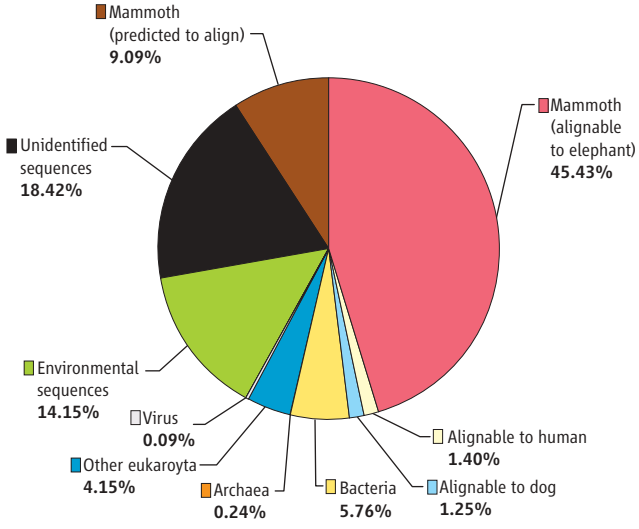

Fig. 1. Characterization of the mammoth metagenomic library, including percentage of read distributions to various taxa. Host organism prediction based on BLASTZ comparison against GenBank and environmental sequences database.

erage approximates only 80% of the total genome, so a conservative estimate is that half of our reads would align to a completed elephant sequence. Among all reads, 44,442 (14.7%) aligned to only one position in the elephant genome, and 21,952 (7.3%) exhibited a perfect (100%) match, up to a read length of 132 bp. To test whether the observed hits were more likely to be derived from endogenous mammoth DNA, as opposed to potential contaminants such as human DNA, we repeated the BLASTZ analyses as above, this time comparing our sequence reads to the currently available versions of the human and dog genomes. Only 4237 reads (1.4%) aligned to human and 3775 (1.2%) to dog (at our threshold of approximately 90% identity). Between 1% and 5% of any two distantly related mammalian genomes should align at 90% identity or greater, because roughly 0.5% of these genomes consist of protein-coding segments conserved at that level (*21*), and noncoding DNA contributes a somewhat larger fraction (*22*). Thus, the fraction of our reads that show at least 90% identity with human and dog is what is expected if only mammoth DNA were sequenced.

To further assess the possibility of contamination in our DNA sample, we explicitly considered what would be expected if contaminating human DNA sequences were present. If some of our reads were human DNA, then these reads should align with nearly 100% identity to human, and at most, 5% of these could be expected to align at 90% identity or higher to any nonprimate mammalian genome (only about 5 to 6% of the genome appears to be under negative selection (*23*), and our 90% threshold is far above the neutral level). If our data contained human reads, we could be essentially certain that a large fraction of them would align to human at or above 97% identity over at least 80% of their length and not align to either elephant or dog. Only 14 reads satisfy these criteria. Thus, there are, at most, trace levels of human contamination in the sample or the process, and the 14 potential examples may simply be regions that are well conserved between mammoth and all other mammals but that happen to be missing from the current assemblies of the dog and elephant genomes. We applied the same approach to look for potential contamination of dog DNA in our sample and found none.

To determine substitution patterns between mammoth and other species (table S2), we used the subset of reads that aligned to only one position. We observed 98.55% identity between mammoth and elephant. As this number does not correct for alterations of the sequence due to damage caused by base decomposition, we are likely underestimating the amount of sequence similarity. Base damage in fossil matrices can result in a myriad of base changes; however, the most commonly observed change has been deamination of cytosine to uracil (*24*, *25*) resulting in C-to-T and G-to-A substitutions. We therefore looked for asymmetries in base composition between modern and ancient genomes (table S2). We reasoned that C-to-T transitions due to DNA damage would manifest as an excess of elephant C aligned to mammoth T over elephant T aligned to mammoth C. Indeed, the ratio of the rates to and from mammoth relative to elephant is 1.91 (15297/7990) for C-to-T and 1.15 (9530/8252) for G-to-A [C-to-T human 0.96 (3849/4027); dog 1.11 (3622/3276); G-to-A human 0.87 (3541/4060); and dog 0.97 (3321/3440)]. Thus, we find noticeable deamination of cytosines in our extract. In addition, we analyzed the frequencies with which substitutions likely to be attributable to postmortem damage occurred in the amplified fragments of mtDNA by comparison with the publicly available mammoth mitochondrion genome (GenBank accession #NC_007596, DQ188829) (*13*). We found 222 reads that aligned to the public GenBank mammoth mitochondrial genome (Fig. 2). Two hundred nine reads gave a total of 18,581 bp, 7617 bp of which were overlapping, resulting in a total coverage of 10,964 bp of a possible 16,770 bp (65%). One hundred fourteen of the 209 reads (55%) matched the previously published sequence exactly. One hundred nine base differences were observed between the reads and the published sequence, 49 of which were not supported by overlapping reads. The majority of these substitutions (84%) were either C-to-T or G-to-A transitions, as is expected if the substitutions were due to postmortem DNA deamination. The remaining 13 reads differed significantly from the published data and may be evidence of potential nuclear inserts of DNA from the mitochondrion (*13*), which have been reported previously to be common in elephants and mammoths (*26*).

These findings are well within the predicted levels of damage for ancient DNA and demonstrate the feasibility and benefits of ancient whole-genome sequencing without previous amplification, as overlapping reads from a multifold coverage would easily correct for decompositional base changes accrued during the sample's depositional history and serve as a DNA damage correction filter. The ratio of mtDNA to nuclear DNA for our sample was 1:658, which agrees with what one would expect given a 1:1000 copy-number ratio for nDNA versus mtDNA.

Despite the presence in our sample of an exceptionally high percentage (54.5%, including reads predicted to align to elephant) of mammoth DNA, relative to environmental contaminants, 45.5% of the total DNA derives from

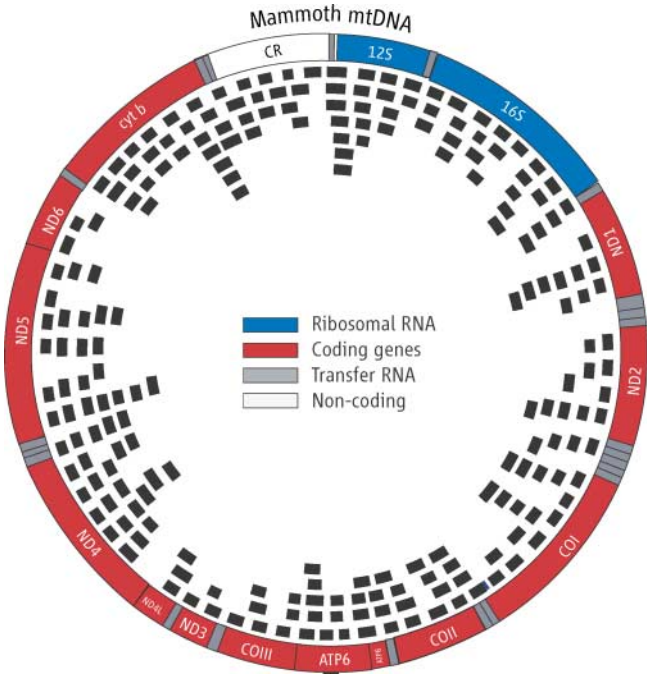

Fig. 2. Distribution of 209 reads along the mammoth mitochondrial genome (GenBank accession #NC_007596, DQ188829). Average fragment length was 89 bp; not shown to scale. To determine whether the reads were randomly distributed along the genome, we compared the distribution of fragment lengths that would result from cutting the genome at the 5′ end of each read against that resulting from 100 million randomly generated distributions of 209 reads. Despite multiple overlapping regions, the real distribution was not significantly different from the empirical distribution of fragment lengths (P = 0.069).

endogenous bacteria and nonelephantid environmental contaminants. In addition to ubiquitous contaminants resulting from handling or conditions of storage, these exogenous species are likely to represent taxa present at or immediately after the time of the mammoth's death, thereby contributing to the decomposition of the remains. To acquire a glimpse of the biodiversity of these communities, we have devised software (GenomeTaxonomyBrowser) (*27*, *28*) that allows for the taxonomic identification of various species on the basis of sequence comparison and current phylogenetic classification at the National Center for Biotechnology Information (NCBI) taxonomy browser as of November 2005 (www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html). We compared 302,692 reads (100%) against the nonredundant (nr) and the environmental database (env_nr). Using an adjustable factor for bitscore (*13*), we classified the reads simultaneously to the individual kingdom, phylum, class, order, family, and genus down to the species level wherever possible (fig. S2 and table S3), excluding all hits matching Gnathostomata (jawed vertebrates).

The remaining 12,563 hits within the Eukaryota (4.15%) were only surpassed by the number of bacterial hits, 17,425 (5.76%), and hits against the environmental database, 42,816 (14.15%). The kingdom Archaea was hit infrequently with only 736 hits (0.24%). Within this group, the Euryarchaeota dominated the Crenarchaeota by a ratio of 16:1. In the bacterial superkingdom, the most prevalent species were found to be proteobacteria, 5282 (1.75%); Firmicutes (gram-positives), 940 (0.31%), mostly Bacilli and Chlostridia; Actinobacteria, 2740 (0.91%); Bacteroidetes, 497 (0.16%); and the group of the Chlorobi bacteria, 248 (0.08%). Other identified microorganisms included the fungal taxa *Ashbya*, *Aspergillus*, and *Neurospora*/*Magnaporte* with 440 hits (0.14%). We also found 278 hits against viral sequences, which could be assigned to dsDNA viruses, 193 hits (0.06%); retrotranscribing viruses, 20 hits (0.01%); and ssRNA viruses, 46 hits (0.02%).

The soil-inhabiting eukaryotic species, *Dictyosteliida*, with 127 hits (0.04%), and *Entamoeba*, with 64 hits (0.02%), were found to be underrepresented, as were hits against the two nematode genomes, 277 hits (0.09%). A detailed identification of plant species is handicapped, because presently only the two plant genomes, *Arabidopsis thaliana* (mouse-ear cress) and *Oryza sativa* (rice), are publicly available. We found the hits against grass species to outnumber the ones from Brassicales by a ratio of 3:1, which could be indicative of ancient pastures on which the mammoth is believed to have grazed.

From this classification, it is evident that nonvertebrate eukaryotic and prokaryotic species occur at approximately equal ratios, with the mammalian fraction dominating the identifiable fraction of the metagenome. The paucity of fungal species is surprising, as is the low number of reads from nematodes.

Recently, a whole-genome approach was attempted from DNA of the extinct cave bear *Ursus spelaeus*, yielding ~27,000 bp of endogenous genetic material from 1.1 to 5.8% of all DNA reads (*29*). We have produced 13 million bp of endogenous genetic material from 45% of all DNA reads, some 480 times as much DNA sequence and 15 times the percentage. The ability to obtain this level of genetic information from extinct species makes it possible to consider detailed analysis of functional genes and fine-scale refinement of mutation rates. A rapid identification assay of single nucleotide polymorphisms (SNPs) would be of great value for studying population genetics of Pleistocene mammals and plants, which in turn could help elucidate their responses to climate changes during late glacial and early postglacial time and ultimately shed new light on the cause and consequences of late Quaternary extinctions.

References and Notes

1. S. Pääbo *et al.*, *Annu. Rev. Genet.* **38**, 645 (2004).
2. H. N. Poinar *et al.*, *Science* **281**, 402 (1998).
3. M. Höss, P. Jaruga, T. H. Zastawny, M. Dizdaroglu, S. Pääbo, *Nucleic Acids Res.* **24**, 1304 (1996).
4. H. N. Poinar, M. Kuch, G. McDonald, P. S. Martin, S. Pääbo, *Curr. Biol.* **13**, 1150 (2003).
5. V. Jaenicke-Despres *et al.*, *Science* **302**, 1206 (2003).
6. M. Bunce *et al.*, *Nature* **425**, 172 (2003).
7. M. Höss, S. Pääbo, N. K. Vereshchagin, *Nature* **370**, 333 (1994).
8. A. D. Greenwood, C. Capelli, G. Possnert, S. Pääbo, *Mol. Biol. Evol.* **16**, 1466 (1999).
9. E. Willerslev, A. J. Hansen, B. Christensen, J. P. Steffensen, P. Arctander, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 8017 (1999).
10. E. Willerslev, A. J. Hansen, H. N. Poinar, *Trends Ecol. Evol.* **19**, 140 (2003).
11. C. I. Smith, A. T. Chamberlain, M. S. Riley, C. Stringer, M. J. Collins, *J. Hum. Evol.* **45**, 203 (2003).
12. H. N. Poinar, M. Hoss, J. L. Bada, S. Pääbo, *Science* **272**, 864 (1996).
13. Materials and methods are available as supporting material on *Science* Online.
14. R. Debruyne, V. Barriel, P. Tassy, *Mol. Phylogenet. Evol.* **26**, 421 (2003).
15. O. Handt, M. Krings, R. H. Ward, S. Pääbo, *Am. J. Hum. Genet.* **59**, 368 (1996).
16. R. D. E. MacPhee *et al.*, *J. Arch. Sci.* **29**, 1017 (2002).
17. R. MacPhee, A. Tikhonov, D. Mol, A. D. Greenwood, *BMC Evol. Biol.* **5**, 49 (2005).
18. A. D. Greenwood, C. Capelli, G. Possnert, S. Pääbo, *Mol. Biol. Evol.* **16**, 1466 (1999).
19. M. Margulies *et al.*, *Nature* **437**, 376 (2005).
20. S. Schwartz *et al.*, *Genome Res.* **13**, 103 (2003).
21. W. Makalowski, J. H. Zhang, M. S. Boguski, *Genome Res.* **6**, 846 (1996).
22. A. Siepel *et al.*, *Genome Res.* **15**, 1034 (2005).
23. R. H. Waterston *et al.*, *Nature* **420**, 520 (2002).
24. T. Lindahl, *Nature* **362**, 709 (1993).
25. M. Hofreiter, V. Jaenicke, D. Serre, A. von Haeseler, S. Pääbo, *Nucleic Acids Res.* **29**, 4793 (2001).
26. A. D. Greenwood, S. Pääbo, *Mol. Ecol.* **8**, 133 (1999).
27. D. H. Huson, A. Auch, J. Qi, S. C. Schuster, in preparation.
28. GenomeTaxonomyBrowser will be made available to readers upon request.
29. J. P. Noonan *et al.*, *Science* **309**, 597 (2005).

30. We thank D. Poinar, C. Flemming, and E. Willerslev for help in mammoth sampling; N. E. Wittekindt and A. Rambaut for help with the manuscript; and two anonymous reviewers. We also thank the Natural Sciences and Environmental Research Council of Canada (299103-2004) for a grant to H.N.P. and McMaster University for financial support. R.D.E.M. was supported by NSF OPP 0117400, B.S. was supported by the Wellcome Trust, and W.M. was supported by NIH grant HG02238. S.C.S. thanks The Pennsylvania State University for initial funding.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1123360/DC1
Materials and Methods
Figs. S1 and S2
Tables S1 to S4
References

2 December 2005; accepted 15 December 2005
Published online 20 December 2005;
10.1126/science.1123360
Include this information when citing this paper.