Genomic sequencing of Pleistocene cave bears

James P. Noonan[1,2], Michael Hofreiter[3], Doug Smith[1], James R. Priest[2], Nadin Rohland[3], Gernot Rabeder[4], Johannes Krause[3], J. Chris Detter[1], Svante Pääbo[3] and Edward M. Rubin[1,2*]

1. US DOE Joint Genome Institute, Walnut Creek, CA
2. Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA
3. Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
4. Institute of Paleontology, University of Vienna, Vienna, Austria

One-sentence summary: Direct cloning and sequencing of genomic DNA from 40,000-year-old extinct cave bears

* Corresponding author. Email: emrubin@lbl.gov.

Abstract

Despite the information content of genomic DNA, ancient DNA studies to date have largely been limited to amplification of mitochondrial DNA due to technical hurdles such as contamination and degradation of ancient DNAs. In this study, we describe two metagenomic libraries constructed using unamplified DNA extracted from the bones of two 40,000-year-old extinct cave bears. Analysis of ~1 Mb of sequence from each library showed that, despite significant microbial contamination, 5.8% and 1.1% of clones in the libraries contain cave bear inserts, yielding 26,861 bp of cave bear genome sequence. Alignment of this sequence to the dog genome, the closest sequenced genome to cave bear in terms of evolutionary distance, revealed roughly the expected ratio of cave bear exons, repeats and conserved noncoding sequences. Only 0.04% of all clones sequenced were derived from contamination with modern human DNA. Comparison of cave bear with orthologous sequences from several modern bear species revealed the evolutionary relationship of these lineages. Using the metagenomic approach described here, we have recovered substantial quantities of mammalian genomic sequence more than twice as old as any previously reported, establishing the feasibility of ancient DNA genomic sequencing programs.

Ancient genomic DNA sequences from extinct species can help reveal the process of molecular evolution that produced modern genomes. However, the recovery of ancient DNA is technologically challenging since ancient DNAs are degraded, mixed with microbial contaminants and individual nucleotides are often chemically damaged (*1, 2*). Ancient nuclear DNA fragments containing unique gene sequences are usually present in too few copies to be amplified by PCR. In addition, ancient remains are invariably contaminated with modern DNA, which amplifies efficiently compared to ancient DNA and therefore makes it difficult to distinguish between ancient and modern genome sequences (*1, 4, 5, 7*). These factors have limited most previous studies of ancient DNA sequence to PCR amplification of mitochondrial DNAs (3-7). In exceptional cases, small amounts of single-copy nuclear DNA have been recovered from ancient remains less than 20,000 years old obtained from permafrost or dry desert environments, which are well suited to preserving ancient DNA (*8-11*). However, the remains of most ancient animals, including extinct hominid species, have not been found in such environments.

To circumvent these challenges, we developed an amplification-independent direct cloning approach to constructing metagenomic libraries from ancient DNA (Figure 1). Ancient remains are obtained from natural environments in which they have resided for thousands of years and any extracted DNAs will be a mixture of genome sequence from the ancient organism and sequences derived from the environment. A metagenomic approach, in which all genome sequences in a particular environment are anonymously cloned into a single library, is therefore a potentially powerful alternative to the targeted PCR approaches that have been used to recover ancient DNAs. We chose to explore this strategy using extinct cave bear instead of an extinct hominid to unambiguously assess

the issue of modern human contamination (*1, 2*). In addition, due to the close evolutionary relationship of bear and dog, cave bear sequences in these libraries can be identified and classified by comparison to the available, annotated dog genome. The phylogenetic relationship of cave bear and modern bear species has also been inferred from mitochondrial sequence, providing the opportunity to compare the phylogenetic information content of cave bear mitochondrial and genomic DNA (*3, 6, 12*).

We extracted DNA from a cave bear tooth recovered from Ochsenhalt Cave, Austria and a cave bear bone from Gamssulzen Cave, Austria, dated at 42,290 (+970/-870) and 44,160 (+1400/-1190) years before present (BP) respectively by accelerator mass spectrometry radiocarbon dating (Table S1). We used these ancient DNAs to construct two metagenomic libraries, designated CB1 and CB2 (Figure 1; *13*). These libraries were constructed in a laboratory into which modern bear DNA has never been introduced. Ancient DNAs were blunt end-repaired before ligation but were otherwise neither enzymatically treated nor amplified. We sequenced 9035 clones (1.06 Mb) from library CB1 and 4992 clones (1.03 Mb) from library CB2. The average insert sizes for each library were 118 bp and 207 bp respectively.

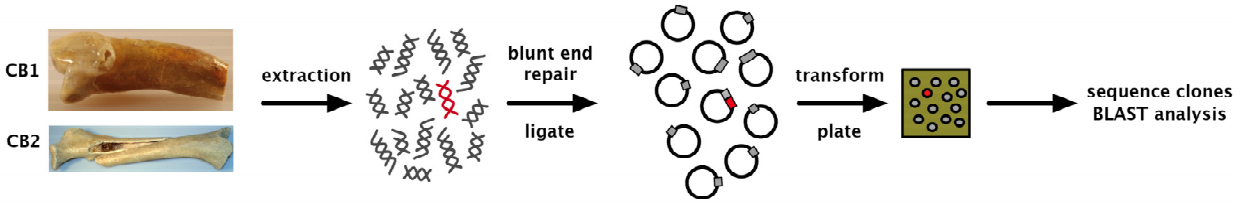

Figure 1. Schematic illustration of the ancient DNA extraction and library construction process. Extracts prepared from cave bear bone contain cave bear DNA (*red*) and a mixture of DNAs from other organisms (*grey*).

We compared each insert in these libraries to GenBank nucleotide, protein and environmental sequences, and the July 2004 dog whole genome shotgun assembly, by using BLAST with an expect value cutoff of 0.001 and a minimum hit size of 30 bp (*14, 15*). 1.1% and 5.8% of clones in library CB1 (Fig 2A) and library CB2 (Fig 2B), respectively, had significant hits to dog genome or modern bear sequences. Our direct cloning approach produces chimeric inserts, so we defined as candidate cave bear sequence only that part of the insert that had a hit to dog or bear sequence. For the 389 clones with significant hits to carnivore sequence, the average hit was 69 bp long and covered 51% of the insert (*16*). BLAST hits to the dog genome from libraries CB1 and CB2 were on average 92.4% and 92.3% identical to dog. To confirm that these sequences were indeed cave bear, we designed primers against 124 putative cave bear sequences and successfully amplified and sequenced 116 orthologous sequences from modern brown bear. All 116 of these modern bear sequences were at least 97% identical to their cave bear orthologs, verifying that all or nearly all of the carnivore sequences in both libraries are genuine cave bear genomic sequences. Only 6 of 14,027 sequenced clones had an insert identical to modern human genomic DNA (Figure 2, A and B). The average BLAST hit length to the human genome for these clones was 116 bp, and the average hit covered 76% of the insert. Although we cannot establish a formal insert length or insert coverage threshold that differentiates between ancient and modern inserts due to the limited number of modern sequences we obtained, these values are significantly greater than the corresponding values for clones with cave bear sequences ($p < 0.05$ for the difference in both average clone length and average insert coverage calculated by two-tailed t-test). This result suggests that it may be possible to discriminate between inserts

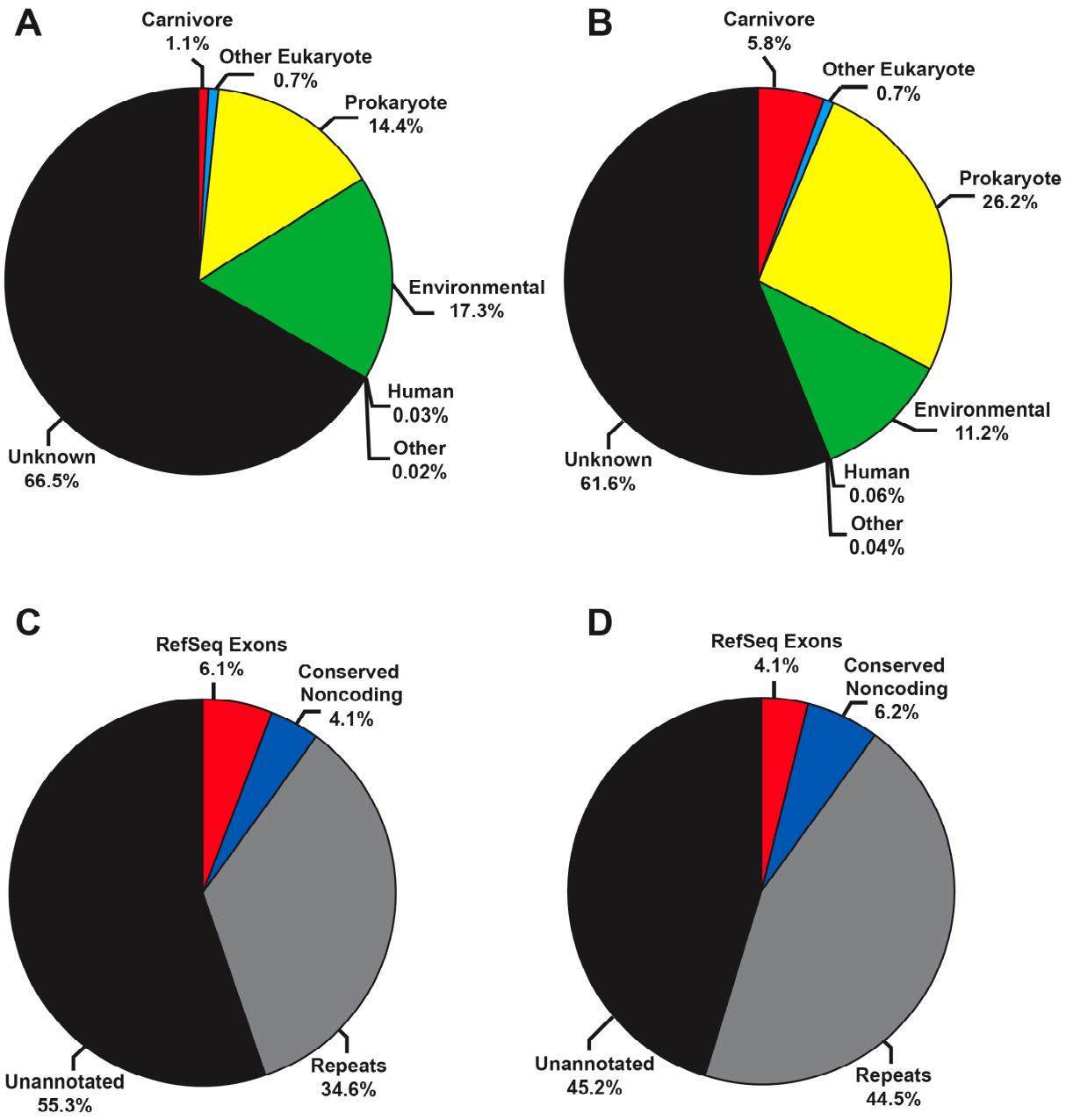

Figure 2. Characterization of two independent cave bear genomic libraries. *Top:* Predicted origin of 9035 clones from library CB1 **(A)** and 4992 clones from library CB2 **(B)** as determined by BLAST comparison to GenBank and environmental sequence databases. "Other" refers to viral or plasmid-derived DNAs. *Bottom:* Distribution of sequence annotation features in 6,775 nucleotides of carnivore sequence from library CB1 **(C)** and 20,086 nucleotides of carnivore sequence from library CB2 **(D)** aligned to the July 2004 dog genome assembly.

derived from short ancient DNAs and inserts containing modern, undamaged DNA in ancient DNA libraries. This may have relevance to the application of these methods to ancient hominids, in which the ability to distinguish ancient hominid DNA from modern contamination will be essential.

The remaining inserts with BLAST hits to sequence from known taxa were derived from other eukaryotic sources, such as plants or fungi, or from prokaryotic sources (bacteria and archaea), which provided the majority of known sequences in each library. In addition, a substantial fraction of inserts in each library (17.3% in library CB1 and 11.2% in library CB2; Figure 2) had hits only to uncharacterized environmental sequences. The majority of these clones had BLAST hits to GenBank sequences derived from a single soil sample (*17*), consistent with the contamination of each cave bear bone with soil bacteria from the recovery site. However, as in other metagenomic sequencing studies, the majority of inserts in each library had no similarity to any database sequence.

To annotate cave bear genomic sequences, we aligned each cave bear sequence to the dog genome assembly using BLAT (*18*). 6.1% of 6,775 cave bear nucleotides from library CB1 and 4.1% of 20,086 cave bear nucleotides from library CB2 aligned to predicted dog RefSeq exons, in a total of 21 genes distributed throughout the dog genome (Figure 2, C and D; Table 1). 4.1% and 6.2% of cave bear nucleotides respectively from library CB1 and library CB2 aligned to constrained nonexonic positions in the dog genome with phastCons conservation scores ≥ 0.8 (CNS, Figure 2, C and D; *13*). The majority of cave bear sequence in each library, however, aligned to dog repeats or regions of the dog genome with no discernible sequence features. These latter sequences appear to be fragments of neutrally evolving, nonrepetitive sequence from the cave bear genome.

Constrained sequences are slightly overrepresented in our set: only 1.7% of bases in the dog genome assembly are annotated as RefSeq exons, and ~10% of the cave bear sequences we obtain appear to be constrained overall, versus an estimated 5-8% in sequenced mammalian genomes. (*19*). This discrepancy may be due to our use of BLAST sequence similarity to identify cave bear sequences, an approach biased in favor of more constrained sequences. Nevertheless, coding sequences, conserved noncoding sequences and repeats appear in both cave bear genomic libraries at frequencies roughly proportional to what has been observed in modern mammalian genomes

Table 1. Cave bear gene sequences identified in CB1 and CB2 ancient DNA libraries.

Gene	Hit	Dog chr	Start	Stop	Description
DHX29	CDS	chr2	43213733	43213804	Predicted DEAH-box helicase
LRRK1	UTR	chr3	42958681	42958735	Protein kinase
FLJ13231	CDS	chr4	74653046	74653076	Hypothetical protein
FXYD2	UTR	chr5	18786890	18786924	Ion transport
AARS	CDS	chr5	79372718	79372836	Alanyl-tRNA synthesis
GLUL	UTR	chr7	18351343	18351418	Glutamate biosynthesis
KSR	CDS	chr9	34690953	34691011	Part of MEK/RAF signaling cascade
ABCA1	CDS	chr11	62139936	62140037	Cholesterol efflux
OR2T33	CDS	chr14	4698505	4698593	Olfactory receptor
GNL2	CDS	chr15	7914165	7914209	Nuclear export of 60S ribosomal subunits
TAL1	UTR	chr15	16435499	16435567	Involved in acute lymphoblastic leukemia
DF	CDS	chr20	60815187	60815275	Complement pathway
FLJ32752	CDS	chr21	32149512	32149550	Hypothetical protein
NKIRAS1	CDS	chr23	22882941	22883002	NF-kappaB cascade
IGSF10	CDS	chr23	49064640	49064697	Platelet-derived growth factor receptor
TNFRSF19	CDS	chr25	18027935	18027952	TNF receptor
CABP7	CDS	chr26	26062041	26062116	Calcium ion binding
GGTLA1	CDS	chr26	30381270	30381396	Glutathione metabolism
VCIP135	CDS	chr29	19241197	19241240	Reassembly of Golgi stacks after mitosis
MYO9A	CDS	chr30	38342316	38342352	Cytoskeleton remodeling
CBS	CDS	chr31	39530760	39530792	Defects cause homocystinuria

.

To determine whether the cave bear sequences we obtained contain sufficient information to reconstruct the phylogeny of cave bear and modern bears, we generated and aligned 3201 bp of orthologous sequences from cave bear and modern black, polar and brown bears and estimated their phylogeny by maximum likelihood (Figure 3A, *20*). This phylogeny is topologically equivalent to phylogenies previously obtained using cave bear and modern bear mitochondrial DNA (*3, 6, 12*). This result further indicates that our libraries contain genuine cave bear sequences, and demonstrates that we can obtain sufficient ancient sequences from those libraries to estimate the evolutionary relationships between ancient and modern lineages. Interestingly, the substitution rate we estimate for cave bear is higher than that in any other bear lineage. Based on results from PCR-amplified ancient mitochondrial DNAs, cytosines in ancient DNA can undergo deamination to uracil, which results in an excess of G to A and C to T (GC-AT) transitions (*21*). The inflated substitution rate in cave bear is likely to be due to an excess of such events, since a large proportion of the substitutions assigned to the cave bear lineage are GC-AT transitions (Figure 3A). These presumably damage-induced substitutions complicate phylogenetic reconstruction and the identification of functional sequence differences between extinct and modern species using ancient DNAs. However, these substitutions are not randomly distributed among all cave bear sequences, but are clustered on a few clones from library CB2. Thus, two clones from library CB2 have three and four GC-AT transitions specific to cave bear, an observation that is extremely unlikely given that the occurrence of true, randomly arising substitutions

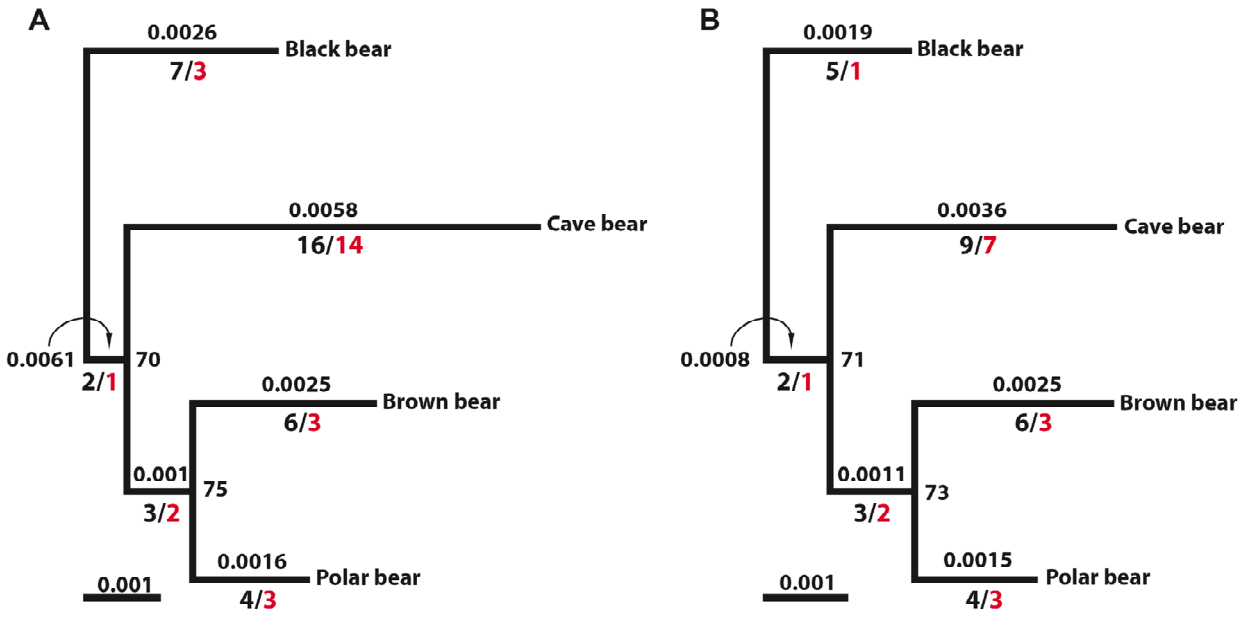

Figure 3. (A) Phylogenetic relationship of cave bear and modern bear sequences obtained by maximum likelihood estimation using 3201 aligned sites. (B) Phylogeny obtained using all sites from (A) excluding cave bear and orthologous modern bear sites corresponding to two heavily damaged cave bear clones (B; Table S2). Substitution rates, total substitutions (*black*) and GC-AT transitions (*red*) for each branch are shown. Orthologous dog sequence was used to root the trees. Percent support for internal nodes is also shown.

on cave bear clones should follow a Poisson distribution (Table S2; *22*). When these two clones are excluded from the analysis (Figure 3B), the apparent substitution rate and the excess of GC-AT transitions in cave bear is drastically reduced, with little impact on substitution rate estimates in the modern bear lineages. Although we cannot distinguish individual GC-AT substitutions from deamination-induced damage, these observations provide a quantitative means to identify heavily damaged clones of ancient DNAs.

In this study we have recovered significantly greater amounts of authentic ancient mammalian genomic DNA than has ever been obtained. The cave bear metagenomic libraries we constructed yielded sufficient sequence, including gene sequences, to reconstruct a bear phylogeny consistent with that obtained from mitochondrial surveys. In addition, our direct cloning strategy has pushed the threshold of ancient mammalian genomic DNA studies at least 20,000 years further into the past than had previously been achieved (*11*). Ancient DNA sequencing programs for extinct Pleistocene species, including hominids, are therefore feasible using the metagenomic methods described here, and by revealing the phylogenomic terrain of recent mammalian evolution, these efforts should help identify the molecular events underlying adaptive differences among modern species.

References

1. S. Pääbo *et al*., *Ann. Rev. Genet.* **38**, 645 (2004).

2. M. Hofreiter, D. Serre, H. N. Poinar, M. Kuch, S. Pääbo, *Nat. Rev. Genet.* **2**, 353 (2001).

3. C. Hänni, V. Laudet, D. Stehelin, P. Taberlet, *Proc. Natl. Acad. Sci. USA* **91**, 12336 (1994).

4. M. Krings *et al*., *Cell* **90**, 19 (1997).

5. M. Krings, H. Geisert, R. W. Schmitz, H. Krainitzki, S. Pääbo, *Proc. Natl. Acad. Sci. USA* **96**, 5581(1999).

6. M. Hofreiter *et al.*, *Mol. Biol. Evol.* **19**, 1244 (2002).

7. D. Serre *et al.*, *PLoS Biol.* **2**, 313 (2004).

8. P. Goloubinoff, S. Pääbo, A.C. Wilson, *Proc. Natl. Acad. Sci. USA* **90**, 1997 (1993).

9. A.D. Greenwood, C. Capelli, G. Possnert, S. Pääbo, *Mol. Biol. Evol.* **16**, 1466 (1999).

10. V. Jaenicke-Després *et al.*, *Science* **302**, 1206 (2003).

11. H. Poinar, M. Kuch, G. McDonald, P. Martin, S. Pääbo, *Curr. Biol.* **13**, 1150 (2003).

12. O. Loreille *et al.*, *Curr. Biol.* **11**, 200 (2001).

13. Materials and methods are available as supporting material on *Science* Online.

14. Dog genome sequence generated by the Broad Institute and Agincourt Biosciences was obtained from the UCSC Dog Genome Browser (http://genome.ucsc.edu).

15. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).

16. Cave bear sequences and related statistics are available as supporting material on *Science* Online.

17. S. G. Tringe *et al.*, *Science*, in press.

18. W. J. Kent, *Genome Res* **12**, 656 (2002).

19. G. M. Cooper *et al.*, *Genome Res* **14**, 539 (2004).

20. We cannot formally exclude alternative topologies given the small number of nucleotide sites in the analysis and the limited divergence among these bear species. Damage-induced substitutions in cave bear also perturb the phylogeny. The topology with the next-best likelihood score has cave bear as the outgroup to the modern bears due to excess of C to T and G to A substitutions in several library CB2 sequences.

21. M. Hofreiter, V. Jaenicke, D. Serre, A. von Haeseler, S. Pääbo, *Nucleic Acids Res.* **29**, 4793 (2001).

22. A detailed description of the distribution of C to T and G to A events in cave bear clones and orthologous modern bear PCR amplicons is provided as supplemental material.

23. We thank members of the Rubin and Pääbo laboratories for insightful discussions and support. This work was conducted at the Max Planck Institute for Evolutionary Anthropology, and at the E. O. Lawrence Berkeley National Laboratory and the Joint Genome Institute with support from the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

Supporting Online Material
www.sciencemag.org
Materials and Methods
Tables S1 and S2
Supporting text for Table S2

Supplemental Online Material

Tables

Dating Number	Location (material)	weight (g)	collagen (%)	Radiocarbon date	$\delta^{13}C$(‰)
KIA 25283	Ochsenhalt (CB1)	0.258	16.1	42290 + 970 / -870 BP	-22.18 ± 0.10
KIA 25287	Gamssulzen (CB2)	1.149	23.0	44160 + 1400 / -1190 BP	-20.38 ± 0.12

Table S1. AMS dating data for the cave bear samples used in this study.

Library		Cave Bear		Modern Bear	
	# C-T/G-A	Clones Predicted	Clones Observed	Clones Predicted	Clones Observed
CB1	0	28.10	28	26.46	27
	1	3.65	4	5.03	4
	2	0.24	0	0.48	1
	3	0.01	0	0.03	0
	4	0	0	0	0
CB2	0	54.54	60	58.49	58
	1	8.73	1	5.26	6
	2	0.70	1	0.24	0
	3	**0.04**	****1**	**0.01**	**0**
	4	**0**	*****1**	**0**	**0**

Poisson probability: ** < 0.0005 * < 0.00005**

Table S2. Poisson-predicted and observed distribution of C to T and G to A substitutions across orthologous cave bear and modern bear genomic clones.

Supplemental material for Table S3

The observation that the GC-AT transitions in Figure 3A are clustered on particular clones suggests that some cave bear genomic DNA fragments are more damaged than others. We hypothesized that the distribution across clones of genuine C to T and G to A substitutions between cave and modern bear sequences would follow a Poisson process. Under this expectation, most clones would have no substitution events, several clones would have one substitution, and clones with more than one substitution

would be extremely rare. In this analysis we consider only those sequence differences between cave bear and modern bear that can be unambiguously assigned to the cave bear or a single modern bear lineage (including ancestral lineages). The distribution of such lineage-specific C to T and G to A substitutions in all sequenced modern bear PCR products indeed follows a Poisson distribution (Table S3). C to T and G to A events are also distributed by Poisson in cave bear sequences from library CB1, while two fragments from library CB2 have three and four C to T or G to A substitutions, which are extremely unlikely under a Poisson-distributed stochastic substitution model. We marked these fragments as heavily damaged and excluded them, and the modern bear sequences orthologous to them, from the alignments that generated the tree in Figure 3B.

Materials and Methods

Extraction and dating of cave bear nuclear DNAs

The cave bear DNA used to construct library CB1 was extracted as described (1), using 125mg of the tooth root and eluting the DNA in 100 µl 1x TE buffer. The cave bear DNA for library CB2 was extracted as follows. 38 g of the compact part of a femur was ground under liquid nitrogen using a freezer mill (Spex freezer mill 6750). The bone powder was incubated in 1 L buffer consisting of 0.5 M EDTA, pH 8 and 250 µg/ml proteinase-K overnight at room temperature. The supernatant was concentrated to 50 ml using the Vivaflow 200 system (Vivascience, Germany) with a polyethersulfone membrane with a molecular weight cutoff of 30,000. The DNA in the solution was purified by binding to silica in 5 aliquots of 10 ml and eluted in 50 µl 1x TE as described (1). The number of copies of mitochondrial DNA of length 100 bp in the extract was estimated to be 60,000 copies per µl of DNA solution, or 15 million copies in the total extract, by quantitative PCR (ABI 7700, Applied Biosystems). If the ratio of mitochondrial to nuclear DNA copies is assumed to be between 1000:1 and 5,000:1 (2, 3), this amount corresponds to 12 – 60 copies of nuclear DNA per microliter or 3,000 – 15,000 copies (10 – 50 ng) in total. Note that these numbers represent only the part of the DNA that is amplifiable by PCR and the amount of directly clonable DNA may in fact be larger.

Both cave bear bones were dated using accelerator mass spectroscopy at the Leibniz Labor für Altersbestimmung und Isotopenforschung, Christian-Albrechts-Universität Kiel, Germany (Table S1). All collagen fractions contained more than 1 mg carbon as recommended for AMS dating. The δ^{13}C-values are in the normal range for bone samples.

Construction of cave bear DNA libraries

End-repair of ancient DNAs

Extracted ancient DNAs were end-repaired using the Epicentre End-It kit (Epicentre, Madison, WI) as follows: 60 µl of each ancient DNA extract were added to 9 µl 10x reaction buffer, 9 µl 2.5mM dNTPs, 9 µl 10 mM ATP and 3 µl enzyme mix and incubated at 25 degrees for 45 minutes. The reactions were heat inactivated at 70 degrees for ten minutes. Following the inactivation, the samples were cooled for a minimum of ten minutes on ice. Reactions were purified using the Qiagen QIAquick PCR Purification kit (Qiagen, Valencia, CA). 450 µl of buffer PB were added to 90 µl of the end-repaired sample. This volume was loaded onto a QIAquick column by centrifugation at 13,000 RPM for 1 minute. Columns were washed using 700 µl of buffer PE. The column was allowed to sit for 5 minutes with buffer PE loaded before centrifugation. Columns were then spun at 13,000 RPM for 1 minute, the flow-through was discarded and columns were spun again (13,000 RPM, 1 min) to remove residual buffer. DNAs were eluted by a 1 minute incubation with 30 µl of warm (50 degrees) elution buffer EB followed by centrifugation for 1 minute at 13,000 RPM.

Ligation of end-repaired DNAs

The end-repaired samples were blunt-end ligated into pMCL200 (http://www.jgi.doe.gov/sequencing/protocols) without any size-selection as follows: 9.8 µl of end-repaired ancient DNAs were added to a reaction mix containing 0.3 µl pMCL200 (~27.5 ng/µl), 1.6 µl 10x ligation buffer, 0.7 µl nuclease-free water. 1.2 µl T4 DNA ligase was added after these were combined in order to reduce the occurrence of vector self-ligation. After these components were combined and mixed, 2.4ul of 30% w/v PEG was added. Further mixing was done by vortexing and hitting of tubes together followed by spinning down. Multiple cycles of mixing were done to ensure thorough mixing. The ligation was incubated at 16 degrees overnight (at least 16 hours) in a water bath.

The ligation reaction was cleaned up by phenol extraction followed by ethanol precipitation. 34 µl of T0.1E (10 mM Tris-Cl, pH 8.0, 0.1 mM EDTA) was added to the 16 µl ligation reaction for a total volume of 50 µl. 50 µl of phenol was added to the sample and mixed by vortexing for 15 seconds. The phenol/sample mixture was then added to a prepared Eppendorf phase-lock tube (Hamburg, Germany). The sample was spun at 12,500 RPM for 5 minutes to separate the phases. The upper aqueous layer was removed. 1/10 volume of 1M NaCl, 2.5 volumes of 100% ethanol and 0.75 µl of pellet paint were added and the sample chilled at -80 degrees for at least 30 minutes. Precipitated samples were pelleted in a refrigerated centrifuge at 4 degrees for 20 minutes at 13,500 RPM. Pellets were washed with 400 µl of 100% ethanol, spun for 5 minutes,

and dried by vacuum centrifugation without heating. Pellets were resuspended in 15 µl T0.1E by mixing on a heated shaker at 40 degrees for 15 minutes.
Library Transformation and Plating

1 µl of resuspended ligation reaction was electroporated into DH10B Electromax™ cells (Invitrogen, Carlsbad, CA) using the GENE PULSER® II electroporator (Bio-Rad, Hercules, CA). Transformed cells were transferred into 1000 µl of SOC and incubated at 37 degrees in a rotating wheel for 1 hour. Cells (5-50 µl) were spread on 22 x 22 cm LB agar plates containing 20 µg/ml of chloramphenicol and 50 mg/ml of x-gal. Colonies were grown for 16 hours at 37 degrees. Individual white recombinant colonies were selected and picked into 384-well microtiter plates containing LB/glycerol (7.5%) media containing 20 µg/ml of chloramphenicol using the Q-Bot™ multitasking robot (Genetix, Dorset, U.K.). To test the quality of the library, 24 colonies were directly PCR amplified with pUC m13 -28 and -40 primers using standard protocols. For details on production sequencing protocols see http://www.jgi.doe.gov/sequencing/protocols/.

Characterization of cave bear libraries

CB1 and CB2 library clones were assigned to the categories shown in Figure 2A by a hierarchical process. Any clone with a significant hit (as defined in the text) to the dog genome or to modern bear sequences was exclusively classified as a cave bear clone if it did not have a better hit to a non-carnivore sequence (defined as a longer hit than the hit to carnivore, with either a higher percent identity or a lower expect value) that overlapped the carnivore hit. Any clone with a hit to the human genome that was > 30 bp and > 98% identical was classified as human contamination and removed from subsequent analysis. Clones without significant carnivore hits were then classified according to their most significant hit in the nr database. Clones without significant hits in nr were classified according to their most significant hit in the nt database. Clones that only had hits to GenBank environmental sequences were classified as such (Figure 2A). For clones with no significant hits to carnivore or human sequences, we only considered the best BLAST hit for each clone, without regard to other regions of the same clone that may have had weaker hits to other sequences.

Cave bear sequences were aligned to the dog genome assembly by BLAT via the UCSC Dog Genome Browser. We created a custom track in the dog genome for each dog-cave bear BLAT alignment, and determined the number of nucleotides in all custom tracks that overlapped with any predicted dog RefSeq genes or repeats using the Table Browser. To estimate the degree of conservation at each nonrepetitive, nonexonic position aligned between dog and cave bear, we used phastCons scores, which are calculated using a phylogenetic hidden Markov model applied to an alignment of human, dog, and mouse genome sequences (4). The phastCons score for each position was obtained, and a position was considered constrained if its score was equal to or greater than 0.8.

Recovery of modern bear DNAs

Primers were designed using Primer3 (5) against 124 nonrepetitive CB1 and CB2 cave bear sequences > 50 bp. Modern black (*Ursus thibetanus*), brown (*Ursus arctos*) and polar (*Ursus maritimus*) sequences orthologous to cave bear sequences were recovered by PCR followed by TA cloning of the products (TA Cloning Kit, Invitrogen, Carlsbad, CA). 99 sequences were obtained for all three modern bears. Three clones were sequenced for each product from each bear species, and only substitutions present in all clones sequenced were included in the analysis. Modern bear DNAs were kindly provided by Dr. Oliver Ryder of the San Diego Zoo.

Cave bear and modern bear sequences (excluding primer sites) were aligned using CLUSTALW (6). Tree topologies for these alignments were constructed using Tree-Puzzle v5.2 under the Tamura-Nei model of nucleotide substitution and the assumption that substitution rates are distributed uniformly among sites (7). Branch lengths and lineage-specific substitutions for the best topologies obtained were re-estimated using baseml under the REV model excluding gapped sites (8).

Supplemental References

1. Hofreiter M, Rabeder G, Jaenicke-Despres V, Withalm G, Nagel D, Paunovic M, Jambresic G and Paabo S. (2004) Evidence for reproductive isolation between cave bear populations. *Curr Biol* **14:** 40-43.
2. Bogenhagen D and Clayton DA. (1974) The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. *J Biol Chem* **249:** 7991-7995.
3. Moraes CT (2001) What regulates mitochondrial copy number in animal cells? *Trends Genet* **17:** 199-205.
4. Siepel A and Haussler D (2003) Combining phylogenetic and hidden Markov models in biosequence analysis. In: *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology* (*RECOMB 2003*), pp. 277-286.
5. Rozen S and Skaletsky HJ (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp 365-386.
6. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG and Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*. **31(13):** 3497-500.
7. Yang Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* **13:** 555-556.
8. Schmidt HA, Strimmer K, Vingron M, and von Haeseler A. (2002) TREE-PUZZLE:

maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*. **18:** 502-504.