Letter

# The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*)

Webb Miller,[1,10] Daniela I. Drautz,[1] Jan E. Janecka,[2] Arthur M. Lesk,[1] Aakrosh Ratan,[1] Lynn P. Tomsho,[1] Mike Packard,[1] Yeting Zhang,[1] Lindsay R. McClellan,[1] Ji Qi,[1] Fangqing Zhao,[1] M. Thomas P. Gilbert,[3] Love Dalén,[4] Juan Luis Arsuaga,[5] Per G.P. Ericson,[6] Daniel H. Huson,[7] Kristofer M. Helgen,[8] William J. Murphy,[2] Anders Götherström,[9] and Stephan C. Schuster[1,10]

[1]*Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, University Park, Pennsylvania 16802, USA;* [2]*Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, Texas 77843, USA;* [3]*Centre for Ancient Genetics, University of Copenhagen, DK-2100 Copenhagen, Denmark;* [4]*School of Biological Sciences, Royal Holloway, University of London, Egham TW20 0EX, United Kingdom;* [5]*Centro Mixto UCM-ISCIII de Evolución y Comportamiento Humanos, c/Sinesio Delgado 4 Pabellon 14, 28029 d, Spain;* [6]*Department of Vertebrate Zoology, Swedish Museum of Natural History, S-10405 Stockholm, Sweden;* [7]*Center for Bioinformatics Tübingen, Tübingen University, Tübingen 72076, Germany;* [8]*Division of Mammals, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20013-7012, USA;* [9]*Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, S-752 36 Uppsala, Sweden*

We report the first two complete mitochondrial genome sequences of the thylacine (*Thylacinus cynocephalus*), or so-called Tasmanian tiger, extinct since 1936. The thylacine's phylogenetic position within australidelphian marsupials has long been debated, and here we provide strong support for the thylacine's basal position in Dasyuromorphia, aided by mitochondrial genome sequence that we generated from the extant numbat (*Myrmecobius fasciatus*). Surprisingly, both of our thylacine sequences differ by 11%–15% from putative thylacine mitochondrial genes in GenBank, with one of our samples originating from a direct offspring of the previously sequenced individual. Our data sample each mitochondrial nucleotide an average of 50 times, thereby providing the first high-fidelity reference sequence for thylacine population genetics. Our two sequences differ in only five nucleotides out of 15,452, hinting at a very low genetic diversity shortly before extinction. Despite the samples' heavy contamination with bacterial and human DNA and their temperate storage history, we estimate that as much as one-third of the total DNA in each sample is from the thylacine. The microbial content of the two thylacine samples was subjected to metagenomic analysis, and showed striking differences between a wild-captured individual and a born-in-captivity one. This study therefore adds to the growing evidence that extensive sequencing of museum collections is both feasible and desirable, and can yield complete genomes.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. FJ515780–FJ515782. See also http://thylacine.psu.edu.]

The thylacine has attracted considerable attention from biologists, for several reasons. Morphologically unique and phylogenetically isolated in its own taxonomic family (Thylacinidae), it also provides one of the most striking examples of convergent evolution in mammals, showing remarkable eco-morphological similarities with members of the placental carnivore family Canidae (i.e., wolves and dogs). The causes and timing of its extinction across both its historical range (midland woods and coastal heath habitats in Tasmania) and its Holocene distribution (extending to continental Australia and New Guinea) are of great interest to many researchers (Johnson and Wroe 2003). It has become a focal point for discussions about large-mammal extinction that include pondering whether the species can be resurrected through a combination of ancient DNA research and modern reproductive medicine (Paddle 2000). Until very recently (Pask et al. 2008), extractions from museum specimens have not yielded usable amounts of endogenous DNA, but rather bacterial or human contaminants (see Supplemental Material), despite the availability of hundreds of thylacine specimens in public and private collections. Because available thylacine DNA is degraded and of low quantity, most phylogenetic studies have focused on short segments of widely used mitochondrial genes, e.g., 12S rRNA and cytochrome b (*cytb*) (Thomas et al. 1989; Krajewski et al. 1992, 1997, 2000).

Several studies have explored the phylogenetic relationships of the thylacine. Some early morphological studies allied it with South American marsupials (e.g., Sinclair 1906). Later morphological phylogenies placed it within the Australian marsupial radiation in Dasyuromorphia, a clade also comprising the numbat (*Myrmecobius fasciatus*) and the family Dasyuridae (insectivorous and carnivorous marsupials) (e.g., Sarich et al. 1982). An initial molecular study used a 219-bp fragment from the 12S ribosomal RNA gene and placed the thylacine in Dasyuridae (Thomas et al. 1989), consistent with an earlier serological study (Sarich et al. 1982). A more extensive analysis using 12S and *cytb* genes, and the

**[10]Corresponding authors.**
**E-mail webb@bx.psu.edu; fax (814) 863-6699.**
**E-mail scs@bx.psu.edu; fax (814) 863-6699.**
Article published online before print. Article and publication date are available at http://www.genome.org/cgi/doi/10.1101/gr.082628.108. Freely available online through the *Genome Research* Open Access option.

nuclear protamine P1 gene, suggested a sister Thylacinidae + Dasyuridae relationship with a basal position for the numbat, though this topology lacked strong statistical support (Krajewski et al. 1992, 1997, 2000).

In this study, we have utilized "next-generation" sequencing technology and improved methods for extraction of ancient DNA to generate two highly accurate thylacine mitochondrial genomes. The new data allow us to determine more accurately the thylacine's phylogenetic position among marsupials. More generally, the success of the current project shows that museum specimens preserved under a variety of conditions are amenable to genomic sequencing.

## Results and Discussion

### Sequencing the thylacine and numbat mitochondrial genomes

Four specimens of thylacine were sampled, three stored as dried skins, and the other an almost complete animal placed in ethanol. From each specimen, we attempted to extract DNA from hair shafts (Gilbert et al. 2007a). One of the dry specimens and the sample stored in ethanol resulted in enough DNA to attempt genomic analysis. The successfully processed dried skin was from an individual that died in the National Zoo, Washington D.C., in 1905 and has been kept at room temperature since then. This individual is the direct offspring of the female specimen sequenced by Krajewski et al. (1997, 2000), and the animal's history is well documented (see Supplemental Material). The ethanol-preserved specimen died in the London Zoo in 1893 and was acquired by Stockholm University ("Zootomiska Institutet" at the time). It was transferred to the Swedish Museum of Natural History (NRM), Stockholm, probably in the early 1970s and has been kept there since. It is a gutted specimen assumed to be a female.

From the dry specimen (Fig. 1) and the ethanol-preserved specimen (Fig. 2) we generated 104,956 and 1,037,369 reads, respectively, using our Roche GS FLX sequencer (454 Life Sciences). The average lengths of the reads were only 87.5 and 67.0 bp, respectively, far below that observed in most modern samples. A major cause is the damaged condition of the DNA, which includes a high rate of double-strand breakage. (However, note in Table 1 that read lengths from modern hair can be much shorter than from other tissues, perhaps due to damage during the keratinization process of hair formation [Linch et al. 2001], though differences in the sample-preparation protocols between the two numbat samples are also a factor.) From the total metagenome, we extracted the reads that aligned with the available northern quoll (*Dasyurus hallucatus*) mitochondrial genome or to contigs created from such reads, and assembled them into independent consensus mitochondrial genomes with 66.8- and 49.7-fold mean coverage (i.e., read depth), for the two thylacine specimens. We trimmed the ends of these assemblies to remove regions that included long, nearly identical repeats, which cannot be confidently sequenced with short-read technologies, leaving contiguous intervals each of 15,492 bases that contain all of the known mitochondrial genes and tRNAs. At a dozen positions, the GS FLX was unable to resolve the precise length of a homopolymer run, which is a well-known limitation of that technology. We resolved all but four of these using PCR and Sanger sequencing, and have annotated the remaining intervals to indicate that, e.g., the run consists of either 6 or 7 Ts. Except for those uncertainties, there are only 5 nucleotide (nt) differences (substitutions) between the sequences from the two thylacine specimens.

To help resolve the phylogenetic position of the thylacine within Dasyuromorphia, we also sequenced the mitochondrial genome of the numbat using the same approach. Two samples were sequenced, one from liver and one from hair shafts, which we assembled into mitochondrial genome sequences with 41.2- and 206.5-fold mean coverage, respectively.

### Properties of the thylacine samples

We identified a substantial amount of human DNA contamination in the thylacine samples, unlike in our earlier analysis of woolly mammoth bone (Poinar et al. 2006; Gilbert et al. 2007b) and hair shafts (Gilbert et al. 2007a, 2008; Miller et al. 2008). For example, from the ethanol-preserved specimen (thylacine 2 in Table 1), we identified 44,493 reads (4.3%) that originated from the human nuclear genome and 136 reads from the human mitochondrial genome. The distribution of read lengths for the human DNA is similar to that for the thylacine DNA (Supplemental Fig. S3), suggesting that the contamination was introduced long ago. The ratio of nuclear reads to mitochondrial reads (327:1) suggests that the contamination is from a human tissue or tissues not particularly rich in mitochondrial DNA. (Compare with the nu/mt ratios in Table 1.) For thylacine 1 (the dry skin), 8.9% of the reads were human, and their average length was 131.9 bp, suggesting a more recent origin. We speculate that the high level of contamination in both thylacine samples, and the difficulty in removing the contaminant sequences using conventional "bleaching" approaches, are due to the samples' curation histories (see Supplemental Material).

To estimate the fraction of our reads that consist of thylacine nuclear DNA, we performed an experiment designed to cope with

**Figure 1.** Dorsal and ventral images of the skin of USNM 125345, young adult male, which died in 1905 at the National Zoo; thylacine specimen 1 sequenced in our study (courtesy National Museum of Natural History Photographic Services).

**Figure 2.** Thylacine specimen 2 sequenced in our study, from the mammal collection of the Swedish Museum of Natural History (Stockholm), NRM 566599, which died in 1893 at the London Zoo. (Photo: Staffan Waerndt, ©Swedish Museum of Natural History.)

the lack of a complete genome sequence from a phylogenetically close relative and with the restricted lengths of the sequenced thylacine fragments. We extracted annotated protein-coding regions from the published genome of the South American opossum (*Monodelphis domestica)* (Mikkelsen et al. 2007), obtaining 31,473,851 bp from a set of 185,575 nonoverlapping exons. Using protein-coding regions rather than all available genome sequence is advantageous because coding regions have fewer insertions and deletions and are easier to align reliably over large evolutionary distances (~138 million years [Myr], i.e., 69 Myr in each lineage) (Nilsson et al. 2004). We took the first 100 bp of 140,490 thylacine reads having at least 100 bp, generated 888,157 nonoverlapping 100-bp intervals from our numbat reads, and generated 1,000,000 nonoverlapping 100-bp fragments from tammar wallaby (*Macropus eugenii*) reads in the NCBI Trace Archive. These sets of 100-bp sequences were aligned to the opossum coding regions under identical conditions. The fraction of reads from the ethanol-preserved thylacine 2 that aligned to coding regions (0.40%) was roughly one-third of the fraction for numbat and wallaby. Assuming that the numbat and wallaby data consists entirely of endogenous DNA, we therefore estimate that 25%–40% of that sample consists of thylacine nuclear DNA. The dry thylacine sample had a somewhat lower fraction that aligned to coding regions, 0.27%.

DNA damage levels (measured as cytosine-to-uracil deaminations, leading to observed cytosine→thymine miscoding lesions) are higher in the thylacine samples than in some published mammoth hair samples (Gilbert et al. 2007a). In the dry-preserved specimen, 0.51% of the C residues in the living animal's mitochondrial genome were replaced by T; for the ethanol-preserved specimen the rate was 0.60%. By comparison, in one woolly mammoth mitochondrial genome (Miller et al. 2008), the rate was 0.14%. This is most likely explained by either the probable higher thermal age (Smith et al. 2003; Gilbert et al. 2007a) of the thylacine samples in comparison to the mammoth sample, the effect of the preservation/storage process (tanning process or long-term storage in dilute ethanol), or a combination thereof (see Supplemental Material). On the other hand, the damage rate is less than that for a mammoth bone sample (Gilbert et al. 2007b), confirming the utility of hair shafts as an excellent reservoir for ancient DNA.

Although our sequence data from the ethanol-preserved specimen include ~20 million bp from the thylacine nuclear genome, we were unable to PCR-amplify nuclear sequences any longer than 150 bp from that sample. This finding agrees with previous observations that thylacine DNA samples are more degraded than those from certain other extinct species and underscores the high sensitivity of the next-generation sequencing approach.

## Phylogeny of the thylacine

Previous analyses using *cytb*, 12S rRNA, and the nuclear gene for protamine P1 have supported the sister-group relationship of Thylacinidae and Dasyuridae, with Myrmecobiidae (numbat) basal within the order Dasyuromorphia (Krajewski et al. 1997, 2000). However, these early studies were based on small matrices (<2.9 kb), and failed to resolve the position of the thylacine with robust statistical support.

The hierarchical relationships among members of Dasyuromorphia were resolved by reconstructing a phylogeny of Marsupialia with the maximum likelihood and Bayesian algorithms using a 14.1-kb mitochondrial genome matrix that included 24 diverse marsupial taxa (Fig. 3A). Analyses using the reduced-bias coding scheme (RB) of Phillips et al. (2006), inclusion of all sites in protein-coding genes + RNA genes (with and without a Bayesian mixed model), found strong support for inclusion of the Thylacinidae within, and occupying the basal position of, Dasyuromorphia (100% bootstrap support [BS], 1.0 posterior probability [PP]), with Myrmecobiidae as the sister clade to Dasyuridae (87%–95% BS, 0.99 PP). Previous studies of a nuclear-only, a RB-mitochondrial, and a combined nuclear + mtDNA data set reported an association

**Table 1.** Some properties of the sequenced thylacine and numbat samples

| Specimen | No. reads[a] | Length[b] | %mtDNA[c] | mt cov[d] | %nuDNA[e] | nu/mt[f] | %human[g] |
|---|---|---|---|---|---|---|---|
| Thylacine 1 | 104,956 | 87.5 | 12.1 | 66.8 | 20 | 1.7 | 8.9 |
| Thylacine 2 | 1,037,369 | 67.0 | 1.1 | 49.7 | 30 | 27.3 | 4.3 |
| Numbat hair | 129,585 | 95.6 | 29.3 | 206.5 | 70 | 2.4 | 0 |
| Numbat liver | 537,424 | 219.3 | 0.5 | 41.2 | 99 | 180 | 0 |

[a]Number of reads produced by the Roche GS-FLX.
[b]Average number of bases per read.
[c]Percentage of reads found to come from the mitochondrial genome.
[d]Average number of reads containing a mitochondrial position.
[e]Estimated percentage of the reads coming from the nuclear genome (see text).
[f]Estimated ratio of nuclear to mitochondrial reads.
[g]Estimated percentage of the reads that are contaminant human DNA.

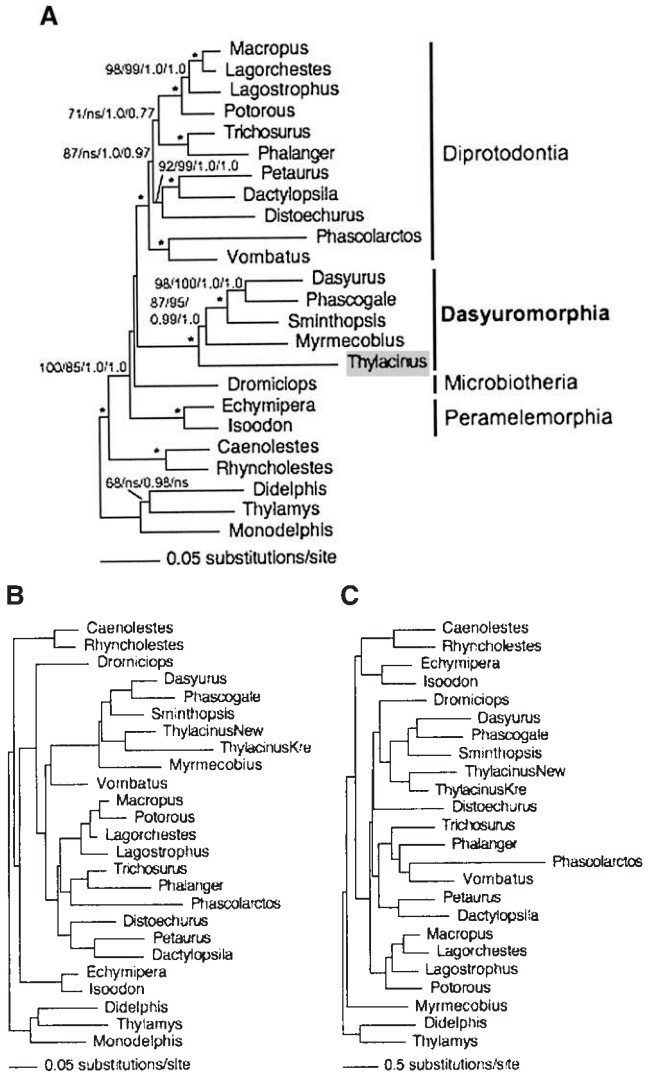


**Figure 3.** (*A*) A maximum likelihood phylogeny of marsupials based on the reduced-bias mitochondrial DNA matrix that places Thylacinidae inside Dasyuromorphia. Didelphimorphia was used as the outgroup, based on results of published nuclear and mitochondrial studies (Amrine-Madsen et al. 2003; Nilsson et al. 2004; Phillips et al. 2006). Bootstrap support values followed by Bayesian posterior probabilities (BPP) are shown *above* respective nodes. The first bootstrap value is derived from the reduced-bias matrix used to generate the tree, and the second bootstrap value is based on the unmodified sequence of the coding mitochondrial genes. The first and second BPP values correspond to the RY-coded and non-RY-coded (mixed model) analyses. Nodes with 100% bootstrap values and 1.0 posterior probability are marked with asterisks. Branch lengths were estimated under the general-time-reversible + gamma + invariants model of sequence evolution. (ns) Not significant, with bootstrap support below 50. (*B,C*) Maximum likelihood phylogeny derived from (*B*) 12S rRNA and (*C*) *cytb* sequences. Note the large divergence between the sequences from this study (labeled ThylacinusNew) and those previously published (ThylacinusKre) (Krajewski et al. 1997, 2000).

between Dasyuromorphia and Peramelemorphia (Amrine-Madsen et al. 2003; Phillips et al. 2006), albeit with weak bootstrap support. Our analyses instead prefer Dasyuromorphia with Diprotodontia (Fig. 3A). However, when we limited our matrix to the same taxa sampled by Phillips et al. (2006) and Amrine-Madsen et al. (2003) we also recovered their Dasyuromorphia + Peramelemorphia clade, suggesting that resolution of Australasian marsupial phylogeny is sensitive to taxon sampling.

## Comparison to published sequences

GenBank contains two putative sequences from the thylacine mitochondrial genome, a 950-bp sequence of the 12S rRNA gene (GenBank accession U87405) and a 1146-bp sequence of the *cytb* gene (M99452) (Krajewski et al. 1997, 2000). These sequences differ from the corresponding parts of our two assemblies (which are identical at these genes) by 11.2% and 15.6%, respectively, at the nucleotide level, though they group together with our sequences in phylogenies based on each gene (Fig. 3B,C). Our attempts to produce a multifold-coverage mitochondrial genome from the female specimen that was studied in Krajewski et al. (1997) failed. However, when we compared the *cytb* amino acid translations from the published Krajewski sequence to our assembly of the direct offspring of the Krajewski specimen (individual USNM 125345), which in theory should be equivalent, we found that they differed at 33 positions (i.e., they are only 91% identical). We analyzed the differences between the two sequences in some detail (see Supplemental Material); our best guess is that the previous sequence is a so-called "numt"(nuclear-inserted mitochondrial sequence) (Song et al. 2008), i.e., a copy of mitochondrial DNA located in the nuclear genome. In addition we verified all nucleotides of our *cytb* assemblies using PCR amplification and Sanger sequencing. Interestingly, an earlier publication (Thomas et al. 1989) reported a 103-bp 12S fragment and a 116-bp *cytb* fragment that are consistent with our assembly; they differ from the GenBank entries in seven positions, while agreeing with ours in all but one position. Moreover, as we have argued elsewhere (Gilbert et al. 2007a), it is essentially impossible that the sequencing method used here would produce data from a numt.

## Metagenomics

All reads from the two thylacine samples were compared against the nonredundant NCBI protein database using BLASTX (translated DNA to protein comparison). A comparative metagenomic species profile, determined using the version beta2.15 of the MEGAN software (Huson et al. 2007), is depicted in Figure 4. A total of 15,625 reads (23%) and 223,661 reads (22%) could be assigned to taxa for samples 1 and 2, respectively. As the metagenome of sample 2 was sequenced to a 10-fold greater depth than sample 1, much of our discussion will therefore refer to sample 2. The majority of the assigned reads are of bacterial origin (154,670), with a small proportion coming from archaeal (401) and viral (375) sources. As the specimen was only poorly preserved postmortem by the London Zoo, its remains apparently underwent internal putrefaction. This is evident from a significant amount of DNA found from *Enterococcus faecalis* and other bacteria associated with post-mortem degradation.

Intriguingly, four out of the ten most abundant bacterial species detected in the hair of thylacine specimen 2 are marine microorganisms (*Psychrobacter* sp. PRwf-1qq, *Psychrobacter cryohalolentis*, *Oceanobacillus iheyensis*, marine actinobacterium PHSC201C1; see Supplemental Table S5), in contrast to the ones detected in numbat hair (Supplemental Fig. S5A,B). Furthermore, other members of the hair biome are microorganisms, such as *Shewanella frigidimarina*, *Shewanella baltica*, *Marinobacteria* sp., or *Psychrobacter arcticus* (Supplemental Fig. S4B), which also derive from a marine environment.

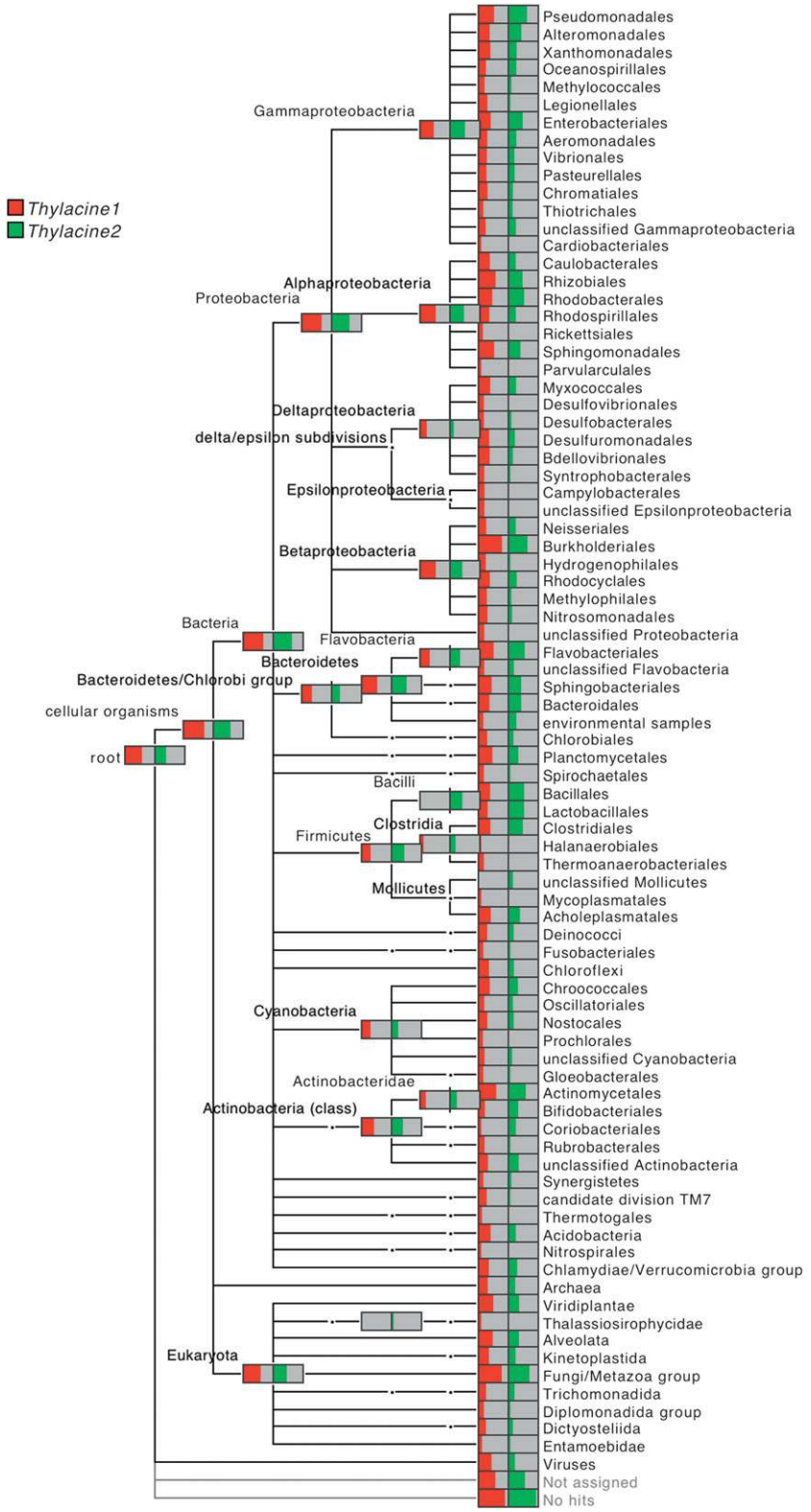


**Figure 4.** Comparative metagenomics study of the two biomes detected in the hair of thylacine specimens 1 and 2.

A comparison between the hair metagenomes from the two thylacines reveals striking differences in the most abundant taxa (Supplemental Table S5). The majority of the marine biota is absent from specimen 1, with the exception of *Polaromonas* sp. JS666. This individual was born outside of its native Tasmania during passage to the United States. It therefore could only acquire microbial communities initially through mother–offspring contact or, later, through feeding and the zoo environment. Its taxa are mostly soil-associated organisms, such as Rhizobiaceae and predatory δ-proteobacteria (*Bdellovibrio* sp.) (Supplemental Fig. S4A). Alternatively, the differences in hair biomes could be explained through the differing preservation methods applied to the specimens. While specimen 2 very likely was put in ethanol as a whole, thereby giving rise to internal putrefaction processes, specimen 1 was handled by a professional taxidermist at the USNM. The extent to which the process of taxidermy added to the thylacine hair biome is unclear, though many taxa seen at high frequency are absent from the ethanol-stored carcass. These are *Acidovorax* sp., *Pseudomonas* sp., *Sphingomonas* sp., *Burkolderia* sp., and *Flavobacteria* sp. One of the prominent microbial species overall, observed at similarly high rates in both data sets, is the aerotolerant, anaerobic gram positive bacterium *Propionibacterium acne*s (Brüggemann et al. 2004), a commensal that is known to degrade components of the skin and also has been reported in the gastrointestinal tract of several mammalian species.

A three-way comparison of the biomes of thylacine hair (specimen 2), numbat hair, and numbat liver also reveals viral loads (Supplemental Fig. S6). A significant number of reads with high similarity to known viruses were detected in the thylacine hair sample (Supplemental Fig. S6A), in contrast to none in numbat hair (Supplemental Fig. S6B); the numbat sample revealed hits only to bacteriophages, which are likely integral to the recorded environmental bacterial organisms. A small number of viral hits were reported for the numbat liver sample (Supplemental Fig. S6C), which otherwise was free of microbial traces. However, some of these taxon identifications might result from noncoding sequences of the thylacine and human genomes, since mammalian genomes are known to contain large numbers of evolutionarily old endo-viruses. Even when

accounting for this possibility, the very high viral load found only in the thylacine hair sample is likely to have originated from the body fluids that were released from the carcass into the preservation liquid and penetrated the hair material during the storage in the aqueous ethanol mixture. The unexpected large presence of viral DNA on the outside of the animal could therefore still indicate a poor health status at the time of this individual's death in the London Zoo in 1893.

### Prospects for a complete thylacine genome

The thylacine constitutes a much greater challenge for molecular characterization than many other species that became extinct within the last 10,000 yr. Despite efforts by many groups, only three short sequences have previously been deposited in GenBank (12S, *cytb*, and the protamine *P1* gene) (Krajewski et al. 1997, 2000). Even the generation of those two mitochondrial sequences proved to be challenging and error prone (>10%). Thus, it comes as a pleasant surprise that our study not only was able to sequence almost the entire mitochondrial genome (minus the highly repetitive hypervariable regions) with high redundancy, but also ~300,000 nuclear-genome reads from the ethanol-preserved specimen 2 (~20 Mb) and ~20,000 nuclear reads from specimen 1.

This bodes well for the feasibility of a complete nuclear genome of the thylacine using the hair/next-generation sequencing approach. While a draft version of the woolly mammoth (*Mammuthus primigenis*) nuclear genome has recently been produced (Miller et al. 2008), much of the success of that project has been attributed to the fact that mammoth specimens, in general, have a cold storage history. In contrast, the thermal age (Smith et al. 2003; Gilbert et al. 2007a) of all thylacine specimens is high, as no efforts were made a century ago to preserve tissues other than tanned or ethanol stored. However, the observed average read length for sample 1 of 87.5 bp, with a substantial number of reads at the maximum read length of the current Roche FLX instruments (~250 bp), suggests that up to 50 Mb of sequence could be generated in a single run. Assuming a genome size similar to that of *Monodelphis domestica* (3.6 Gb), less than $1,000,000(US) in reagent cost would allow for the generation of a data set in excess of onefold coverage using the sequencing technology employed for this study. Given the interest in the species and the projected decrease in next-generation sequencing costs, the goal of a complete thylacine genome seems feasible.

### Conclusions

This study validates the use of ancient DNA extracted from hair in combination with whole-genome shotgun sequencing even for samples that have a temperate storage history and for cases where no sequence data are available from a closely related species. The redundant sequence coverage yields a high-fidelity assembly, while avoiding PCR-based artifacts, which can include the presence of nuclear-inserted mitochondrial sequences (numts) (Gilbert et al. 2007a; Song et al. 2008). In contrast, the traditional PCR-based approach is laborious at best and can fail utterly, as it has with the thylacine, particularly when no closely related sequence is available. Finally, our method produces a full mitochondrial genome, which supports analyses that are more robust than those possible with short mitochondrial fragments. Availability of an accurate thylacine mitochondrial sequence should facilitate further studies of this appealing species, such as an analysis of the temporal changes in mitochondrial diversity leading up to extinction.

Nuclear genome sequence data from museum specimens will let the results of evolutionary events be observed directly, rather than studies being limited to inferences based solely on samples from living organisms. In particular, they will permit evaluation of genetic factors that affect or are affected by the extinction process, providing a valuable tool for future preservation efforts in terms of assessing extinction risk in extant endangered species.

A systematic study is needed to determine factors affecting the utility of museum specimens for molecular analysis. In our case, did analysis of one Smithsonian sample fail and another prove very successful (see Supplemental Material) because of differences in exposure to light over a long period? Experiments using comparatively expendable museum specimens with varying documented histories of preservation, storage, etc., are needed to inform museum personnel which sampling risks are more likely to pay off than others, and also how best to preserve both existing and newly acquired specimens for future molecular analysis. One component of such studies will be to determine which associations with microbial species result from particular preservation and/or storage histories, and which reflect the microbiome of the living animal. These studies, together with an inventory of global museum holdings, will set the stage for coordinated and appropriate use of museum specimens to deepen our understanding of evolution and the extinction process.

## Methods

### Specimens

Thylacine sample 1 was from USNM 125345, a taxidermically preserved adult male that died at the U.S. National Zoo in 1905. Thylacine sample 2 was from NRM 566599 in the mammal collection of the Swedish Museum of Natural History (Stockholm), an adult female that died in the London Zoo in 1893 and has been stored in an ethanol solution. The hair and liver numbat samples are from the same individual, DEC Donnelly District specimen number M108, killed by an automobile in the Greater Kingston National Park, Western Australia. The sex of the animal was not recorded. The Supplemental Material contains further information on the specimens.

### Sample preparation and sequencing

Library preparation for the thylacine and numbat hair samples were performed as previously described for the woolly mammoth (Gilbert et al. 2007a, 2008), and all fragment sizes were retained. For the numbat liver sample, a gel extraction of fragments was performed, followed by library preparation as previously described (Miller et al. 2008). Results from these earlier papers, together with data reported here for the two numbat samples (Table 1), indicate that essentially no human contamination is added to the sample by our protocols.

### Assembling the thylacine and numbat mitochondrial sequences

Sequence reads aligning to the mitochondrial genome of the northern quoll (*Dasyurus hallucatus*; GenBank Accession AY795973.1) were filtered and trimmed to require at least 70% alignment coverage and 60% identity, then assembled. Subsequently, all reads were aligned to the assembled contigs at higher stringency (80% coverage and 95% identity) and the hits were reassembled. The thylacine mitochondrial genomes have GenBank

accession numbers FJ515780 and FJ515781, while the numbat mitochondrial genome is FJ515782.

### Measuring human contamination

We aligned the reads to the March 2006 assembly of the human nuclear genome (chromosomes 1–22, X and Y) and to the human mitochondrial genome (GenBank entry NC_001807), and required that they align with at least 90% coverage and 97% identity.

### Mitochondrial genome phylogenetic analysis

We compiled a DNA alignment of 24 marsupial mitochondrial genomes. These included mitochondrial genomes generated during this study (thylacine sample 2 and the numbat) and genomes downloaded from GenBank. The control region was excluded because of ambiguous alignment across marsupials. All protein-coding genes (*NADH6* was excluded following Phillips et al. [2006]), ribosomal RNA genes, and tRNA genes were aligned using CLUSTALX (Thompson et al. 1997); then the alignments were visually examined and adjusted to minimize the number of indels. Ambiguous alignment regions were identified and 502 nt sites were excluded, yielding a 10,785-bp sequence matrix.

The compositional bias in the mitochondrial genomes of marsupials leads to a loss of phylogenetic signal, causing errors in tree reconstruction (Phillips et al. 2006). We therefore used the "reduced-bias" approach of Phillips et al. to reduce sources of compositional bias. This scheme RY-codes mtRNAs (both stems and loops), and Y-codes the first positions and excludes the third positions of protein-coding genes. We also used the same model of sequence evolution (GTR + I + $\Gamma_4$) as Phillips et al. (2006). A non-RY-coded data set was also subjected to an ML search. For both ML searches the starting parameters were estimated during the initial reconstruction of the ML tree with a full heuristic search using a neighbor-joining (NJ) starting tree and tree-bisection-reconnection (TBR) branch swapping in PAUP* (Swofford 2003). Model parameters were then re-estimated from the ML tree and a new ML search was performed using the new parameters. These steps were repeated until the model parameters stabilized, and then these parameters were used to reconstruct the final ML phylogeny on which an ML bootstrap evaluation was performed (100 heuristic replicates with TBR). Trees were also reconstructed using a Bayesian algorithm implemented in MrBayes version 3.1.2 (Ronquist and Huelsenbeck 2003), using both a RY-coded and a non-RY-coded (all sites included) mixed model approach. Two independent runs were performed with four independent chains, sampled every one-thousandth generation for 5 million generations. The first 1 million generations were discarded as burn-in. We determined convergence between the two runs when the average standard deviation of split frequencies was less than 0.01.

### Estimating the fraction of thylacine nuclear DNA

We extracted putative protein-coding intervals from the nuclear genomic sequence of the South American opossum (*Monodelphis domestica*), based on an assembly (monDom4) and Ensembl gene annotation (dated 2006-07-31) that we downloaded from the UCSC Genome Browser at http://genome.ucsc.edu/. We aligned these intervals to sets of 100-bp fragments taken from the initial portions of sufficiently long thylacine and numbat reads, and from Sanger reads of the tammar wallaby (*Macropus eugenii*) downloaded from ftp://ftp.ncbi.nih.gov/pub/TraceDB/macropus_eugenii/.

These fragments were aligned to the opossum coding regions using the BLASTZ program (Schwartz et al. 2003), scoring matches at +1 and mismatches at –2, and penalizing a gap of length $n$ by $-4 - n$. The fractions of aligning 100-bp fragments were 0.27% for thylacine 1, 0.40% for thylacine 2, 1.24% for numbat liver, and 1.10% for wallaby, which we summarize by saying that roughly three times more of the fragments aligned for numbat and wallaby than for the better of the thylacine samples. Estimating the ratio of nuclear to mitochondrial reads followed readily; see Supplemental Material for details.

## Acknowledgments

We thank Don Wilson, Linda Gordon, Michael Carleton, Helen Kafka, Paul Rhymer, Lauren Helgen, and Christen Wemmer for assistance at the Smithsonian Institution. We also thank Dr. Michael Bunce and Dr. Peter Spencer (Murdoch University), and Peter Mortensen, Olavi Grönwall, and Erik Åhlander (Swedish Museum of Natural History) for assistance with sample collection and data about the thylacine specimens, together with Adrian Wayne (Department of Environment and Conservation, Western Australia) for providing the numbat tissues. We are grateful to the National Museum of Natural History's Photographic Services for the images in Figure 1, and to Cathy Riemer for suggestions about the presentation. Sequencing-by-synthesis costs were covered in part by generous funding from Penn State University. This work was supported by the National Human Genome Research Institute, grant HG002238 (to W.M.), and by the Gordon and Betty Moore Foundation (to J.Q. and S.C.S.). This project was funded, in part, under a grant from the Pennsylvania Department of Health using Tobacco Settlement Funds appropriated by the legislature.

## References

Amrine-Madsen, H., Scally, M., Westerman, M., Stanhope, M.J., Krajewski, C., and Springer, M.S. 2003. Nuclear gene sequences provide evidence for the monophyly of australidelphian marsupials. *Mol. Phylogenet. Evol.* **28:** 186–196.

Brüggemann, H., Henne, A., Hoster, F., Liesegang, H., Wiezer, A., Strittmatter, A., Hujer, S., Dürre, P., and Gottschalk, G. 2004. The complete genome sequence of *Propionibacterium acnes*, a commensal of human skin. *Science* **305:** 671–673.

Gilbert, M.T.P., Tomsho, L.P., Rendulic, S., Packard, M., Drautz, D.I., Sher, A., Tikhonov, A., Dalén, L., Kuznetsova, T., Kosintsev, P., et al. 2007a. Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. *Science* **317:** 1927–1930.

Gilbert, M.T., Binladen, J., Miller, W., Wiuf, C., Willerslev, E., Poinar, H., Carlson, J.E., Leebens-Mack, J.H., and Schuster, S.C. 2007b. Recharacterization of ancient DNA miscoding lesions: Insights in the era of sequencing-by-synthesis. *Nucleic Acids Res.* **35:** 1–10.

Gilbert, M.T.P., Drautz, D.I., Lesk, A.M., Ho, S.Y.W., Qi, J., Ratan, A., Hsu, C.-H., Sher, A., Dalén, L., Götherström, A., et al. 2008. Intraspecific phylogenetic analysis of Siberian woolly mammoths using complete mitochondrial genomes. *Proc. Natl. Acad. Sci.* **105:** 8327–8332.

Huson, D.H., Auch, A.F., Qi, J., and Schuster, S.C. 2007. MEGAN analysis of metagenomic data. *Genome Res.* **17:** 377–386.

Johnson, C.N. and Wroe, S. 2003. Causes of extinction of vertebrates during the Holocene of mainland Australia: Arrival of the dingo, or human impact? *Holocene* **13:** 941–948.

Krajewski, C., Driskell, A.C., Baverstock, P.R., and Braun, M.J. 1992. Phylogenetic relationships of the thylacine (Mammalia: Thylacinidae) among dasyuroid marsupials: Evidence from cytochrome b DNA sequences. *Proc. Biol. Sci.* **250:** 19–27.

Krajewski, C., Buckley, L., and Westerman, M. 1997. DNA phylogeny of the marsupial wolf resolved. *Proc. Biol. Sci.* **264:** 911–917.

Krajewski, C., Blacket, M.J., and Westerman, M. 2000. DNA sequence analysis of familial relationships among dasyuromorphian marsupials. *J. Mamm. Evol.* **7:** 95–108.

Linch, C.A., Whiting, D.A., and Holland, M.M. 2001. Human hair histogenesis for the mitochondrial DNA forensic scientist. *J. Forensic Sci.* **46:** 844–853.

Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., and Heger, A. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447:** 167–177.

Miller, W., Drautz, D.I., Ratan, A., Pusey, B., Qi, J., Lesk, A.M., Tomsho, L.P., Packard, M.D., Zhao, F., Sher, A., et al. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456:** 387–390.

Nilsson, M.A., Arnason, U., Spencer, P.B., and Janke, A. 2004. Marsupial relationships and a timeline for marsupial radiation in South Gondwana. *Gene* **340:** 189–196.

Paddle, R. 2000. *The last Tasmanian tiger: The history and extinction of the thylacine*. Cambridge University Press, Cambridge, U.K.

Pask, A.J., Behringer, R.R., and Renfree, M.B. 2008. Resurrection of DNA function in vivo from an extinct genome. *PLoS One* **3:** e2240. doi: 10.1371/journal.pone.0002240.

Phillips, M.J., McLenachan, P.A., Down, C., Gibb, G.C., and Penny, D. 2006. Combined mitochondrial and nuclear DNA sequences resolve the interrelations of the major Australasian marsupial radiations. *Syst. Biol.* **55:** 122–137.

Poinar, H.N., Schwarz, C., Qi, J., Shapiro, B., Macphee, R.D., Buigues, B., Tikhonov, A., Huson, D.H., Tomsho, L.P., Auch, A., et al. 2006. Metagenomics to paleogenomics: Large-scale sequencing of mammoth DNA. *Science* **311:** 392–394.

Ronquist, F. and Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572–1574.

Sarich, V., Lowenstein, J.M., and Richardson, B.J. 1982. *Carnivorous marsupials* (ed. M. Archer ), pp. 707–709. Royal Zoological Society of New South Wales, New South Wales, Australia.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003. Human-mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.

Sinclair, W.J. 1906. Mammalia of the Santa Cruz beds: Marsupialia. *Reports of the Princeton University expedition to Patagonia 1896–1899*, Vol. IV, pp. 333–460. Princeton University Press, Princeton, N.J.

Smith, C.I., Chamberlain, A.T., Riley, M.S., Stringer, C., and Collins, M.J. 2003. The thermal history of human fossils and the likelihood of successful DNA amplification. *J. Hum. Evol.* **45:** 203–217.

Song, H., Buhay, J.E., Whiting, M.F., and Crandall, K.A. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci.* **105:** 13486–13491.

Swofford, D.L. 2003. *PAUP*: Phylogenetic analysis using parsimony (*and other methods)*. Sinauer Associates, Sunderland, MA.

Thomas, R.H., Schaffner, W., Wilson, A.C., and Pääbo, S. 1989. DNA phylogeny of the extinct marsupial wolf. *Nature* **340:** 465–467.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. 1997. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25:** 4876–4882.

*Received June 25, 2008; accepted in revised form November 18, 2008.*