www.sciencemag.org/cgi/content/full/1123360/DC1

Supporting Online Material for

Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA

Hendrik N. Poinar,* Carsten Schwarz, Ji Qi, Beth Shapiro, Ross D. E. MacPhee,
Bernard Buigues, Alexei Tikhonov, Daniel H. Huson, Lynn P. Tomsho, Alexander Auch,
Markus Rampp, Webb Miller, Stephan C. Schuster*

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

Published 20 December 2005 on *Science* Express
DOI: 10.1126/science.1123360

This PDF file includes:

Supplementary online material (SOM)

CONTRIBUTIONS

H. Poinar, R.D.E. Macphee, A. Tichonov, B. Buiges, collected and sampled the mammoth remains in Siberia. C. Schwarz performed screening, large scale DNA extraction and quantitation. L.P. Tomsho and S.C. Schuster conducted sequencing analysis. W. Miller, S.C. Schuster, B. Shapiro, H. Poinar performed genomic comparisons and data analyses. D.H. Huson, A. Auch, Ji Qi, M. Rampp, S.C. Schuster carried out metagenome taxon identification, H. Poinar and S.C. Schuster designed experiments. B. Shapiro, R.D.E. MacPhee, W. Miller, S.C. Schuster and H. Poinar wrote the manuscript.

Materials and Methods

Sample preparation

Mammoth Sampling and clean room work

Specimen 2005/915 was removed from the permafrost in the summer of 1999 and taken to the ice museum in Khatanga which maintains a constant temperature of –15ºC. The mandible was sampled using a large 2cm x 5cm coring drill bit. To ensure to the greatest extent possible that minimum contamination was introduced to the samples, all drill bits were sterilized in bleach, UV irradiated and individually wrapped in sterile bags prior to being taken to Siberia. Once drilled, bone cores were placed into gamma sterilized 50ml Falcon tubes, which were sealed and brought back to a clean room dedicated for the extraction of DNA from fossil samples, at McMaster University. The samples were placed into –80ºC freezers until they were processed. All samples were processed as is typical for ancient DNA work, thus all work was conducted in a clean room, using sterilized materials, which were either bleached or UV irradiated. All buffers were made with double distilled water which was subsequently UV irradiated, thus any contamination that did arise in our sequence analysis is most likely to originate from the sample itself and not post permafrost or laboratory contaminants.

DNA Extraction

In total 1 g of mammoth bone (2005/915, mandible) was chopped into smaller pieces using a hammer and chisel. Eleven pieces of ~90 mg (and one blank) were each incubated with 1.5 ml of 0.5 M EDTA (EMD Chemicals Inc.) pH 8 overnight (ON) at 22°C on a rotary wheel. The next day samples were spun for 5 minutes at 16.000 * g, the supernatants were removed and stored for separate purification and each pellets was then incubated in 1.5 ml of the following digestion buffer: (10 mM Tris (EMD Chemicals Inc.) @ pH 8.0, 0.5 % Sarcosyl (Sigma), 250 µg/ml Proteinase K (Fisher Biotech), 5 mM $CaCl_2$ (EMD Chemicals Inc.), 50 mM DTT (EMD Chemicals Inc.), 1 % PVP (EMD Chemicals Inc.), 2.5 mM PTB (Prime Organics, Inc.) ON at 55°C on a rotary wheel. The supernatants (EDTA digest) were transferred to new tubes and subsequently extracted with 0.5 ml phenol/chloroform/isoamylalcohol (25/24/1, Fisher Biotech) and 0.5 ml chloroform (Fluka). The resulting aqueous phases were spun through microcon ultrafiltration units (30k, Millipore) and each washed 3 times with 300µl 0.1 x TE (10 mM Tris pH 8 and 0.1 mM EDTA, centrifugation each time to dryness of membrane @ 16,000 * g). DNA was finally recovered by adding 100µl of 0.1 x TE to each filter unit, 5 minutes mixing at 1000 rpm (Eppendorf thermomixer), followed by an up-side-down spin into collection tubes. All eleven DNA solutions were pooled.

The next day the pellet digests were spun for 5 minute at 16.000 * g and supernatants were extracted and spun through microcon filter units as done the previous day. All 11 DNA solutions were pooled as well. The decalcification pool and the digest pool were combined and concentrated using a single microcon filter unit (30k) to obtain a final volume of ~100µl which was used for library construction.

Quantitative PCR

The number of amplifiable mt-DNA fragments were determined by quantitative PCR (qPCR) on an Mx3000P qPCR system (Stratagene) using one forward primer, CytB_F111, and six different reverse primers, CytB_R171, -R 241, -R 371, -R 572, -R 763 and -R 1010 amplifying fragments of 84 to 921bp in length of the cytochrome B gene, see table S1. These primers have been designed to amplify both African and Asian elephant (*Loxodonta africana* and *Elephas maximus* respectively) and match perfectly to two cytochrome B sequences of *Mammuthus primigenius* (accession numbers D50842 and D83047) while excluding human. Standard template DNA was generated with primers CytB_F111 and CytB_R1010 on *Elephas maximus* whole genomic DNA. Standard curves were generated for each of the six primer combinations by amplifying a serial dilution with the following copy numbers: 50, 500, 5,000 and 50,000 generated from a purified PCR product of known concentration (measured via UV spectrophotometry). All six primer combinations were tested to a sensitivity of ~10 copies or less under the reaction conditions specified below. To test the sensitivity of the primers and to see whether contaminating DNA in higher concentrations might yield non specific amplification products or potentially interfere with the quantitation of low mammoth DNA in our samples, we spiked the 10 copy standard as well as the no template controls (NTC) with 4 ng of human whole genomic DNA (ca. 650 nuclear copies or ~ half a million mtDNA copies). In all cases the primers did not amplify cytB from human DNA (spiked NTCs) and quantitation of the 10 copies of elephant DNA spiked with half a million human mtDNA copies, consistently yielded ca. 10 amplifiable molecules. Efficiencies of the six assays were between 87 and 100 % as calculated from the slope of the standard curve, R^2-values between 0.996 and 1.000. To check for inhibition in our extracts we quantitated our extract straight as well as 1:10 dilution. No inhibition was detected in our extracts. Each 20 µl reaction contained the following: 1x PCR Buffer II (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems), 250 µM dNTPS (each, Amersham), 0.75 mg/ml BSA (Sigma), 250 nM of each primer (CytB_F111 and the reverse primer, IDT), ref dye (1:500 dil., Stratagene), 0.167 x SYBR Green (Sigma), 1 unit AmpliTaq Gold (Applied Biosystems), and 5µl of DNA template (standard, extract or water). The temperature profile for the reaction included an initial activation of the enzyme at 95°C for 7 min, followed by 45 cycles of the following 95°C for 30 sec, 60°C for 30 sec and 72°C for 90 sec and a final extension at 72°C for 3 min. The dissociation curves were generated using the following thermal profile: 95°C for 1 min, 55°C for 30 sec and 95°C for 30 sec. Optical data for the amplification was acquired following each extension step, for the dissociation curve during the 55°C to 95°C ramp. The cycling conditions were the same for all six primer pair combinations.

Library construction and DNA sequencing

The Mammoth DNA library was constructed, as previously described (Margulies *et al.* 2005), by shearing our DNA extract into fragments which were blunt-ended and phosphorylated by enzymatic polishing using T4 DNA polymerase, T4 polynucleotide kinase, and *Klenow* DNA polymerase. The polished DNA fragments were then subjected to adapter ligation followed by isolation of the single-stranded template DNA (sstDNA). The quality and quantity of the sstDNA library was assessed using the Agilent 2100 Bioanalyzer. The sstDNA library fragment was captured onto a single DNA capture bead and clonally amplified within individual emulsion droplets. The emulsions were disrupted using isopropanol, the beads without an amplified sstDNA fragment were removed, and the beads with an amplified sstDNA fragment were recovered for sequencing. The recovered sstDNA beads were packed onto a 70x75mm PicoTiterPlate™ and loaded onto the GS 20 Sequencing System (454 Life Sciences, Branford, CT) as previously described. The mammoth bone metagenome sequence has been assigned NCBI Trace Archive SID 131303.

Characterization of Mammoth nuclear DNA libraries

We aligned the sequencing reads with current (as of November 2005) assemblies of the genome sequences of African elephant (*Loxodonta africana*), human, and dog (*Canis familiaris*), downloaded from http://genome.ucsc.edu/. Alignments were computed by the program blastz (Schwartz et al., 2003), with parameters chosen to identify only high-identity matches. Specifically, alignments were scored by +1 for a match, -3 for a mismatch, and -3-k for a gap of length k. Alignments were retained only if they contained a gap-free segment of score at least 30. Thus, for alignments covering 50 nucleotides, the minimum identity was 90%. To permit alignments with elephant to be computed in a reasonable time despite the presence of large numbers of unrecognized (i.e., by the RepeatMasker program) elephantid-specific interspersed repeats, we ran blastz with the parameter M=20, so that a sequencing read was ignored once it had been aligned to 20 elephant positions. Because blastz avoids "masked" regions on its initial steps, special care was needed to identify reads contained in old interspersed repeats; our approach was to extract masked segments of the elephant assembly, unmask them, and align to our reads, setting M=2. The 14 sequence reads that may be human contamination (based on high-identity matches to human but not to elephant or dog) are given in Table S4. To determine substitution patterns between mammoth and elephant (Table S2), we used the subset of reads that aligned to only one position (excluding repetitive elements) in the elephant assembly. We believe that those alignments have the highest likelihood of pairing orthologous regions, and hence of accurately measuring the rate and pattern of evolutionary substitutions and DNA damage.

To confirm our expectations that at most 5% of the mammoth or human sequence would align to the other species at 90% identity, we performed the following computational experiment. We extracted 95bp intervals (i.e., the average size of our reads) randomly from the elephant sequence and removed the masking for interspersed repeats. These artificial reads were run through our pipeline that aligns reads to the human genome at high stringency. On average, 4.9% of the reads aligned.

Characterization of Mammoth mitochondrial DNA libraries

To determine whether the distribution of reads that aligned with high stringency to the mammoth mitochondrial genome was random with respect to genome position, we calculated the length of 209 fragments that would result from cutting the circular genome at the 5’ base beginning each read. An empirical distribution was then generated under the null hypothesis of random sampling by simulating 100 million instances of randomly cutting a circular fragment of the same length 209 times, and by comparing the real distribution to the resulting distribution of fragment lengths.

Characterization of non-mammalian DNA libraries

Metagenomic analysis

302,692 reads from the sequenced library were blasted against the non-redundant and environmental database using compute services at www.migenas.org. The result of blasting all fragments against the nr and env_nr and databases is summarized in a so-called fragment hit file, which contains a single line for each fragment specifying the

NCBI taxon ids and bit scores for each match to a sequence in one of the data bases. Species names were extracted from blast output and converted into taxonIDs (November, 2005 (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/)) using the software GenomeTaxonomyBrowser (Huson, Auch, Qi, and Schuster, in preparation). This program is initialized using the current NCBI taxonomy, which contains approx. 280.000 taxa. In this taxonomy, we summarize the number of times each taxon is matched to one of the fragments. A fragment that matches more than one species is summarized under the lowest common taxon, so, for example, a fragment that matches both human and mouse will be counted under Euarchontoglires. The program draws the tree that links all hit taxa and summarizes the number of hits for each taxon. The statistics for all hit species were calculated and displayed as summarizing graphics or a table. We used versions of the BLAST software v2.2.10 and the database releases from November 01, 2005.

References for supplementary online materials

S. Schwartz et al., Genome Research 12 (2003).

M. Margulies *et al.*, *Nature* **437**, 376 (2005).

Supplementary online tables

Table S1

Primer	Sequence	Product size (bp)
CytB_F111	5’-AGGAGCATGCCTAATTACCCA-3’	-
CytB_R171	5’-GATGAAAATGCAGTTATTGTGTCA-3’	84
CytB_R241	5’-TGCTCCGTTTGAGTGTAGTTG-3’	151
CytB_R371	5’-CCTATGAAGGCGGTGGCTA-3’	279
CytB_R572	5’-AAGGAAGGTTAGGTGTACTCCTGCTAGTG-3’	490
CytB_R763	5’-TAAGTGGATCAGCTGGTATGTAGTT-3’	677
CytB_R1010	5’-TCTACTGGTTGACTGCCAATTC-3’	921

Table S1, qPCR primers used on mammoth DNA extracts

Table S2

(a)

	Elephant				
Mammoth		A	C	G	T
	A	1,311,474	3,142	9,530	3,650
	C	2,230	842,070	2,139	7,990
	G	8,252	2,254	849,260	2,786
	T	2,262	15,297	2,724	1,266,890

(b)

	Human				
Mammoth		A	C	G	T
	A	89,584	996	3,541	1,015
	C	1,044	58,569	857	4,027
	G	4,060	856	59,484	1,165
	T	1,022	3,849	916	87,981

(c)

	Dog				
Mammoth		A	C	G	T
	A	81,915	948	3,321	966
	C	829	53,967	787	3,276
	G	3,440	799	54,432	921
	T	882	3,622	886	80,723

(d)

	Elephant mitochondria				
Mammoth mitochondria		A	C	G	T
	A	4,922	4	156	12
	C	6	3,085	3	125
	G	143	5	2,907	3
	T	8	196	8	4,836

Table S2. Base substitutions of uniquely aligned sequences between (a) mammoth and elephant, (b) mammoth and human, (c) mammoth and dog and (d) mitochondrial sequences between mammoth and elephant.

Table S3

Organisms	# hits included in analysis	% total reads
Total # Aligned reads	302,692	100%
Bacteria	17,425	5.76%
Proteo-bacteria	5,282	1.75%
Bacteroidetes	497	0.16%
Chlorobi	248	0.08%
Firmicutes	940	0.31%
Actinobacteria	2,740	0.91%
Archaea	736	0.24%
Euryarchaeota	615	0.20%
Crenarchaeota	42	0.01%
Eukaryota other than **Gnathostomata** (Jawed Vertebrates)	12,563	4.15%
Rhabditida	277	0.09%
Fungi	806	0.27%
Saccharomycetaceae (Ashbya)	119	0.04%
Trichocomaceae (Aspergyllus)	108	0.04%
Sordariomycetes (Neurospora / Magnaporte)	213	0.07%
Entamoeba	64	0.02%
Dictyosteliida	127	0.04%
Viridiplantae	751	0.25%
Brassicales	170	0.06%
Oryza	420	0.14%
Virus	278	0.09%
dsDNA virus	193	0.06%
retro-transcribing virus	20	0.01%
ssRNA virus	46	0.02%
Environmental sequences	42,816	14.15%
Unidentified sequences	55,830	18.44%

Table S3. Metagenomic analysis of species distribution in the permafrost preserved mammoth sample. All reads other than Gnathostomata were included in the analysis.

Table S4

```
>063873_1601_3574 length=100 hits hg17.chrX:134,471,259-134,471,358
CTGGAATACAAGCTCCTGCCATATTAATAAGCCCCGATAGAACATTTGACAAGATGTTAT
CCGGTTGCGGTTGGCTATTTTCCATCCTCTCTTCATGGAC

>081303_1615_1491 length=53 hits hg17.chr5:148,772,875-148,772,928
TACTATCACACATGGAATGACCTCAACTGTCCCTCTGTCCAAACCAGGGAACC

>165492_0392_2187 length=70 hits hg17.chr10:125,505,014-125,505,084
AAAATATTAGGATCTGAGACGGTTCACAGAAACACATTTACTAAAAACTGAATTTAATTT
TATAGGCACC

>202582_1192_0130 length=98 hits hg17.chr3:99,168,901-99,168,999
TGAAGTTTAATCTCGCCCATCAAGACAGAGACTTTTTCAAGGAATTCTGGGAGCAGAAGC
CCCTTCTTATTCAGAGAGATGACCTGCACTGGCACATA

>254782_0575_2079 length=70 hits hg17.chr1:56,328,803-56,328,872
GTTTGTTTGGAAACGTGACTTGCAGGCTTTGAACTGAACCCAAAGGAATGTTTCAGGTTT
GCTCTTGGCC

>278120_0324_1268 length=73 hits hg17.chr7:69,224,032-69,224,103
GGCTGTCTTCGTCTGTGACATGTGTGCACACGTAGACATTCTTGCATGTGCTTTTCTGTC
AGCCCAGATTGAC

>278457_0941_3956 length=97 hits hg17.chrX:9,040,316-9,040,413
CTTTCTTTTTCTGCAATAAAGCATGCCTGGCTGTTCTCTCATGGATAGAACTGCCTCTGT
CCTATTGCTACTTACTCTTTGAATATTATGACAAAAG

>050627_3479_2221 length=70 hits hg17.chr1:31,898,995-31,899,056
AGCCGCTCCGGCCGCTGCCCGCAGCGCGCGCGGGGCCGAGTGACGGCCGAGGCGGGACGC
GGCCGTCACC

>067373_3011_1125 length=97 hits hg17.chr9:66,022,795-66,022,891
TTCAGAACACTAAAAGAAGATAACATTGTGAACAAATCTCCCAGGGAATCCACTCGAGAG
ATGAGGCACACTGACCACAGATGAGTGAGGTCCTCAA

>087508_3821_1885 length=62 hits hg17.chr5:70,777,129-70,777,187
GATTCTGCATTGAAGTTCTAATGACTAACTGACATTCTGCACTGCAGCAAGTGACTGTTA
TG

>104579_3181_1327 length=86 hits hg17.chr10:102,437,797-102,437,882
ATTTTCTGGCTAGTGAATCAAGTGGAGGGAGCTAATTACATGAAGATCTGAACAAAAATA
ACTCCTAATTTTCAAGGATAATGGAA

>114209_3716_2133 length=100 hits hg17.chr3:96,734,548-96,734,647
GTTAGGAAGTGGAGAAATACATGAGCAAGATCATTGTTATAACAGAGACAGATGTTACAA
TCTGTTTCTTCTTTGGTCATTCACTATCAAGATACTTAAG

>140766_3704_2132 length=85 hits hg17.chr6:41,714,802-41,714,885
ACACACACACACAAACACACACACATACACTCCGCGGTGTCTGTCCGTCTGGGATTTGTG
TCTCAACTGTTTCTGCCCAGTGTGC

>238278_3494_0147 length=110 hits hg17.chr12:130,027,817-130,027,918
CGAGGTGGAGCTTCCCTCCTGTCTCACTGGTGGACCCTCTCGGTGGGAAACCGCGGGGAT
GAACCGATCTCTAAGACTAAGGTTCACAGGCACCACAGGGGCTCTCACAG
```

Table S4. The 14 examples of potential human contamination, giving location (within 3% error) in the May 2004 human genome assembly.

Supplementary online figures

Figure S1

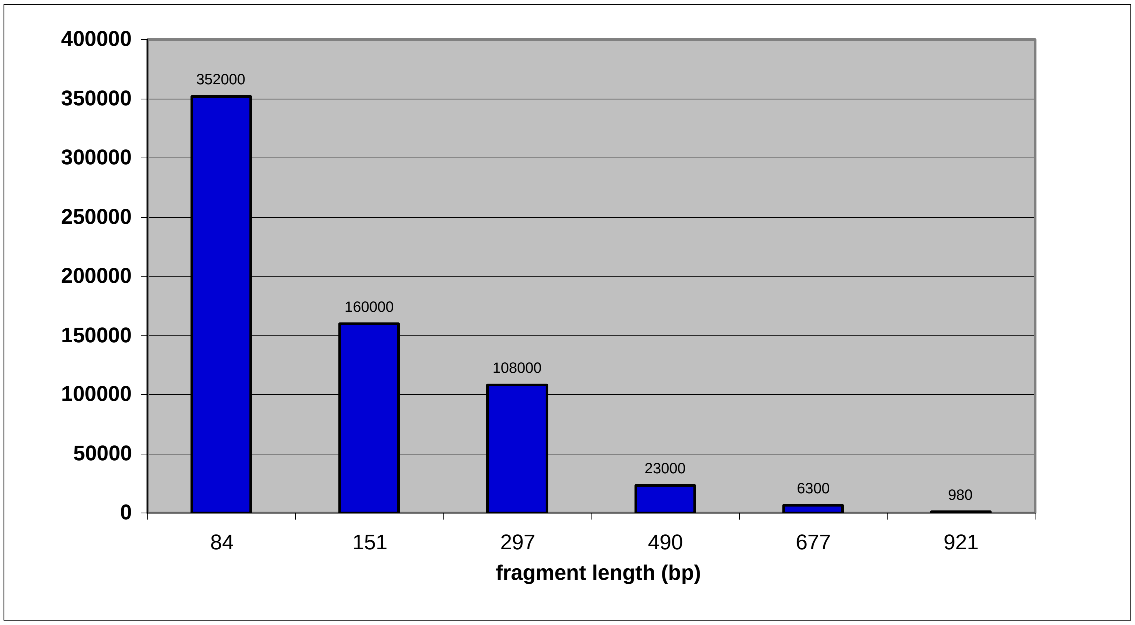

Figure S1. Number of mtDNA copies of varying fragment lengths from sample 2005/915.

Figure S2

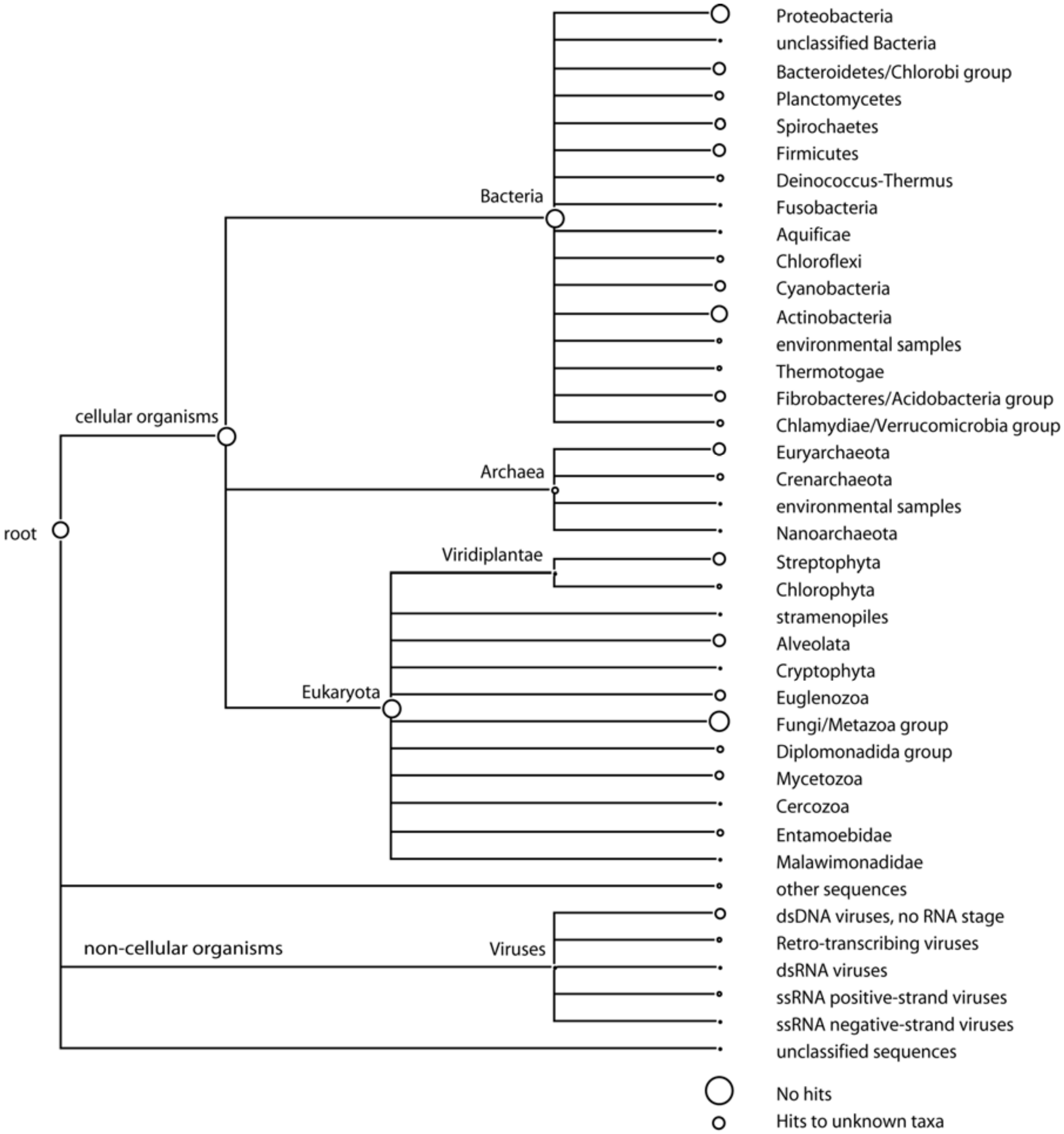

Figure S2. Metagenome taxon identification using GenomeTaxonomyBrowser
The sequences from 302,692 reads were searched against the sequence non-redundant and environmental databases (nr and env_nt). Statistics of identified species are also in Table 3.