GSA Data Repository 2016199

# Fossil DNA persistence and decay in marine sediment over hundred-thousand-year to million-year time scales

Kirkpatrick et al.

## METHODS:

### Sediment Sampling

The sediment was recovered as 9.5 m cores. Stainless-steel blades, which were cleaned and then sterilized with ethanol, were used to cut sediment into 1.5 m sections. Sampling for DNA was conducted on the drillship catwalk so as to minimize exposure time and potential contamination from the shipboard environment. We sampled with pre-sterilized (autoclaved) cut-off syringes that we pushed into a freshly cut section faces immediately after unwrapping the autoclaved foil cover. We wore fresh nitrile gloves to handle samples, and subsequent to removal from the sections the syringes were immediately put in fresh Whirlpak bags and sealed before freezing at -80 °C.

### DNA extraction and amplification

To extract DNA we used a flame-sterilized spatula to first exposed sediment off the frozen syringe, and then added 0.25 g of underlying frozen sediment to a 2-mL bead tube with 0.3 g of 0.1 mm autoclaved zirconia/silica beads (Biospec Products, Bartlesville, OK, USA) pre-

16 aliquoted. We also ran kit blanks with beads and all the solutions but no sediment. To each
17 sample tube, we added 500 μL of MoBio Bead Solution, 250 μL of phenol-chloroform-isoamyl
18 alcohol (pH 8.0; ThermoFisher Scientific, Waltham, MA, USA), and 60 μL of PowerLyzer®
19 solution C1 before beadbeating for 90 s on a Biospec Mini-Beadbeater-96. We then spun the
20 tubes for 10 min at 16,873 g before removing the supernatant and proceeding with standard kit
21 directions. We conducted all open-tube steps in a Envirco Laminar Flow Work Station
22 (Albuquerque, NM, USA), except the phenol steps, for which we used a fume hood. At the end
23 of the Mobio protocol, we combined multiple 0.25 g extracts (up to 6) from the same depth as
24 necessary and eluted DNA with 80 μL of kit solution C6. We cleaned extracts with Agencourt
25 AMPure XP beads (Beckman Coulter, Inc., Brea, CA, USA) per manufacturer's directions, using
26 Mobio kit solution C6 to resuspend. 2.5 μL aliquots were set aside for DNA quantification with a
27 Qubit® 2.0 Fluorometer using the dsDNA HS Assay Kit (ThermoFisher Scientific). We
28 calculated our detection limit (1.01 ng DNA / g sediment) based on the relative fluorescent units
29 (RFUs) of three times the standard deviation of Mobio solution C6 (Tris-Cl). We measured a
30 subset of both shallow and deep sediment samples in triplicate to determine sample measurement
31 error (±0.4 ng / g sediment). We then stored extracts at -80° C until proceeding further.
32 In order to create amplicons for sequencing, we conducted PCR with partial Nextera
33 adapters attached to primers targeting the v4v5 hypervariable region of bacterial 16S rDNA

(Huse et al., 2010). Our sequences ranged in length from 410 – 430 bp. The primers were 518F (5'- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAGCAGCYGCGGTAAN-3') and a 8:1:1 mix of three 926R primers (5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCGTCAATTCNTTTRAGT–3', 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCGTCAATTTCTTTGAGT-3', 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCGTCTATTCCTTTGANT-3'). We amplified DNA with the addition of 1× BSA in our PCR. We used either 0.5 ng extracted DNA as template, or the maximum volume possible for the deepest, lowest biomass samples (<0.5 ng below 80 m CCSF) and blanks. Reaction volume was 25 μL. We used thermal cycling conditions of 94° C for 9:30, then 30 cycles of 94° for 30 s, 57° for 45 s, and 72° for 30s, with a final 72° extension for 5:00. After another cleanup with Agencourt AMPure XP beads, where we pooled replicate reactions, we submitted the DNA samples and blanks for sequencing to the University of Rhode Island (URI) Next Generation Sequencing facility (http://web.uri.edu/gsc/next-generation-sequencing/). Illumina (San Diego, CA, USA) MiSeq sequencing was conducted using reagent kit v3, 2 × 300 bp read length paired-end chemistry, with phiX added. 15.5% of reads aligned to phiX.

**Sequence analysis**

We trimmed and merged fastq sequence data from Illumina's Basespace® using CLC Workbench version 6.0 (CLC Bio). We used a quality score corresponding to a Phred score of 15 as our trim cutoff, determined empirically to yield the greatest number of successfully merged pairs with CLC, with >100 bp of overlap. Either lower or higher Phred scores resulted in fewer merged pairs. After exporting merged reads as a fasta file, we used the Mothur MiSeq pipeline (Kozich et al., 2013; Schloss et al., 2009). Our assessment of total reads per sample (Supplementary Table 1) is based on the sequences remaining after quality control and prior to taxonomic assignment. We used SILVA v119 to align sequences. Our protocol deviates from the standard operating procedure (SOP; http://www.mothur.org/wiki/MiSeq_SOP) in that rather than removing reads with a taxonomic assignment of chloroplast, before clustering we removed everything except those reads with a taxonomic assignment of chloroplast. No remaining sequences overlapped between sediment samples and the kit blank. We did not remove singletons, because i) we're not reporting on OTU richness or diversity, but rather taxonomy ii) it does not change the identification of the most common OTUs and iii) to the extent the removing singletons could remove "real" environmental data (unique sequences), it would inflate the apparent importance of the "dominant" OTUs (Supplementary Figure 2) -- to be conservative we have avoided this. Overall, at U1339 singleton OTUs represented 3% of cpDNA reads and

9% at U1343. Sample information and read counts (total and chloroplast) are in Supplementary Table 1.

To build our phylogenetic tree of major photosynthetic eukaryote lineages, we aligned the top 10 most abundant Operational Taxonomic Units (OTUs; defined at the 99% cutoff) with database sequences using ClustalX (Jeanmougin et al., 1998) and SeaView (Gouy et al., 2010). We used a 99% cutoff because we are considering variation within a single bacterial clade (chloroplasts); this may inflate OTU numbers, but we believe this is the conservative approach because i) to the extent it creates "artificial" OTUs it will actually downplay the importance of our "dominant" OTUs, and ii) we are not analyzing OTU-based metrics of diversity. We aligned these fragments with the analogous stretch of cpDNA from database sequences representing both major and minor siliceous microfossil taxa reported from these sites via microscopy (Takahashi et al., 2011), as well as outgroups and potential floral contaminants from our office and campus spaces. This alignment was manually curated, though few problems were found. We then used the alignment to calculate phylogeny two ways: i) using MrBayes 3.2 (Ronquist *et al.*, 2012), with GTR weighting, a gamma distributed rate variation, and $10^6$ generations; and ii) a maximum likelihood tree with 1000 bootstraps using phyML (Guindon and Gascuel, 2003). We did not note major variations between the two different methods; both trees are shown here for comparison (Figure 3a, b).

**Detection of "rare" sequences**

Many sequence types in next-generation datasets such as these are found at frequencies that appear extremely low (e.g. <0.1% of reads), but are still represented by dozens or hundreds of repeated sequences. Consequently, whether or not we would expect to detect these "rare" sequences is not intuitive, so we calculated a probabilistic model based on random re-sampling (i.e., assuming an infinitely large pool of sequences). This is a first-order test of detectability, but does not take in to account amplification biases.

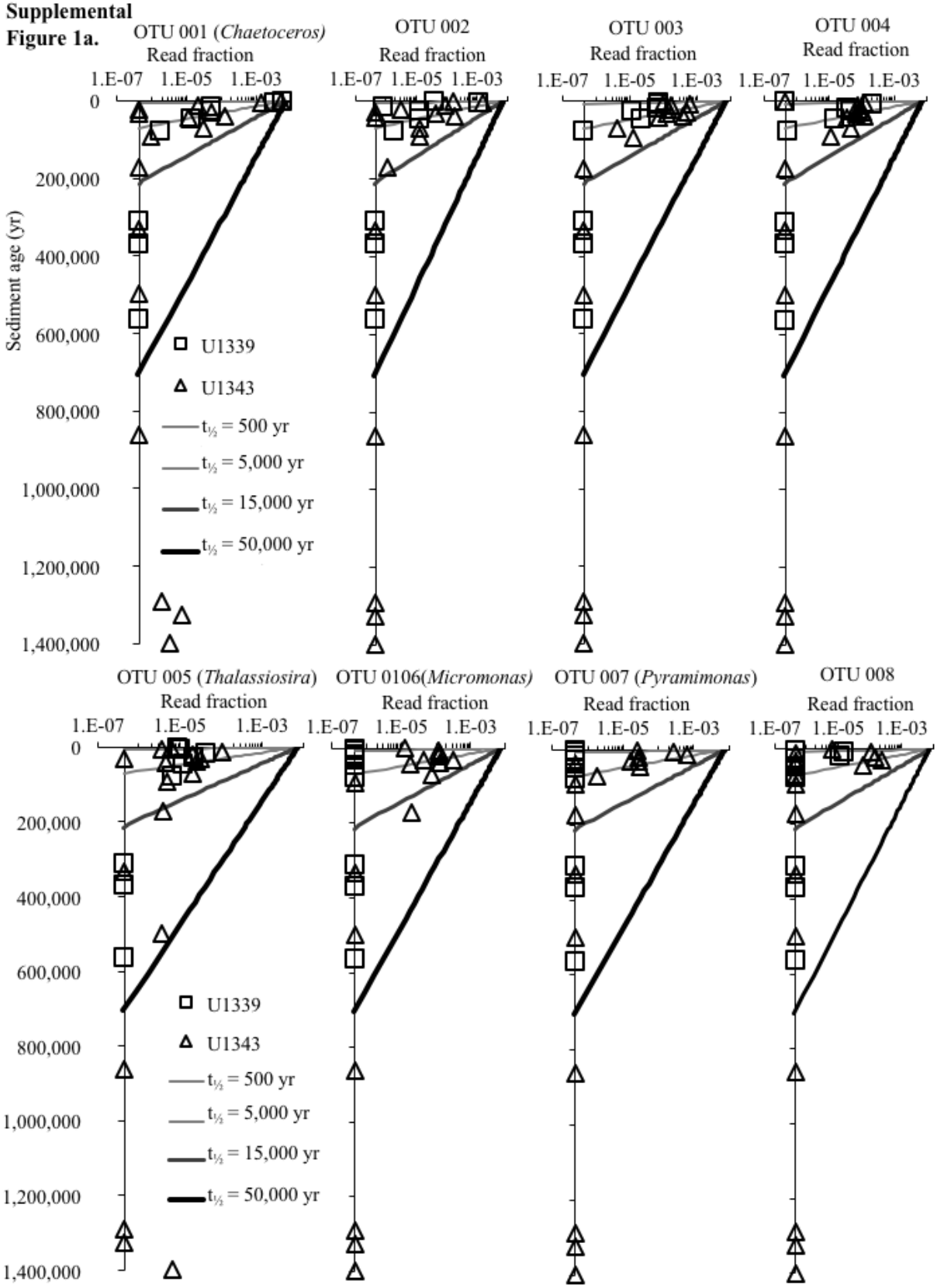
If a given taxa, such as OTU XYZ, has a frequency of 1% in the data, then picking one read would give a 0.01 chance of detecting that sequence – or a 0.99 chance of missing that sequence. Picking two reads would give 0.99 × 0.99 = 0.9801 chance of not detecting that sequencing. In other words,

$$p_{taxa} = (1\text{-}f_{taxa})^{N}$$

Where $p_{taxa}$ is the probability of non-detection for a given taxa or sequence type; $f_{taxa}$ is the frequency of that sequence type in the dataset; and N is the number of sequences analyzed. The minimum number of chloroplast reads at any depth for U1339 was 57, or 0.026% of reads for that sample, corresponding to a probability of "missing" this group of sequences at $1.7 \times 10^{-25}$. The minimum number of chloroplast reads at any depth for U1343 was 13, or 0.006% of

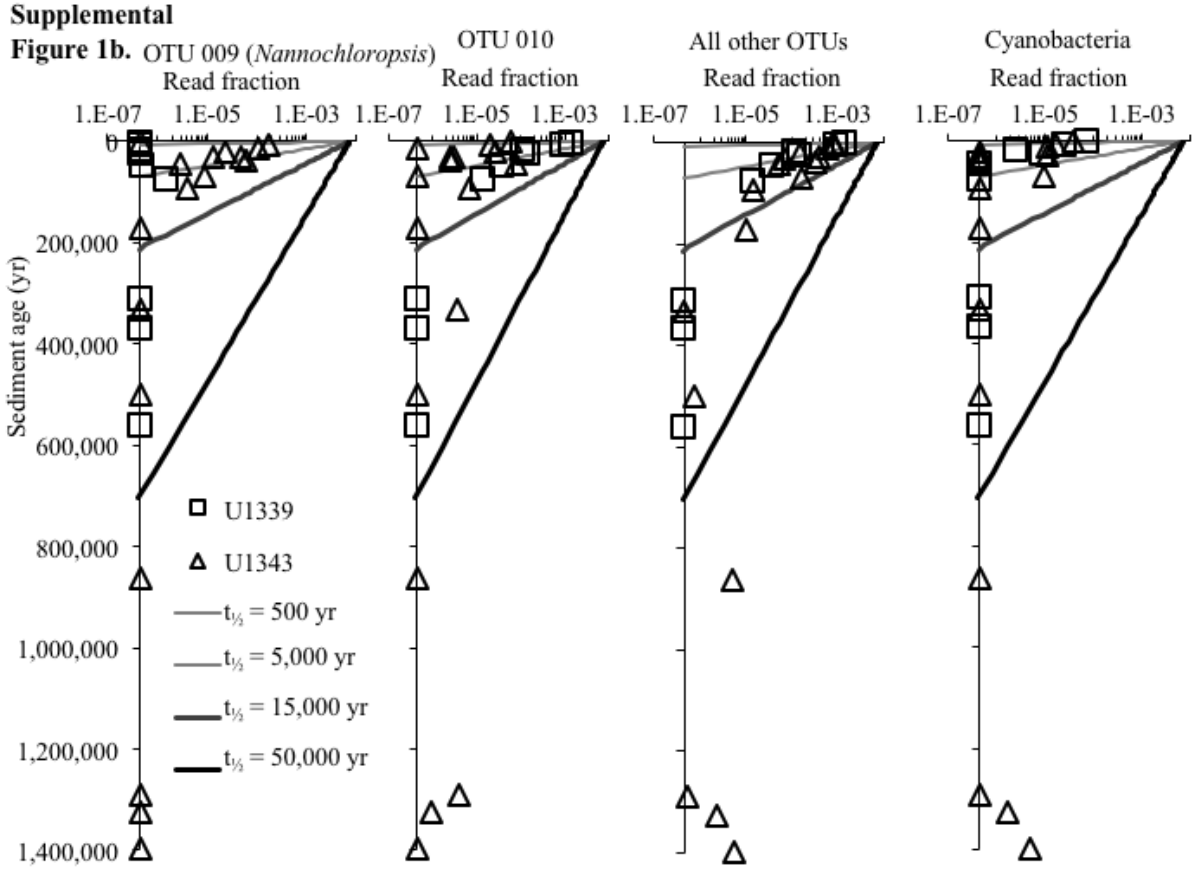
reads for that sample, corresponding to a probability of “missing” this group of sequences at 2.3 $\times 10^{-6}$.

**Supplemental Figure 1a.**




106

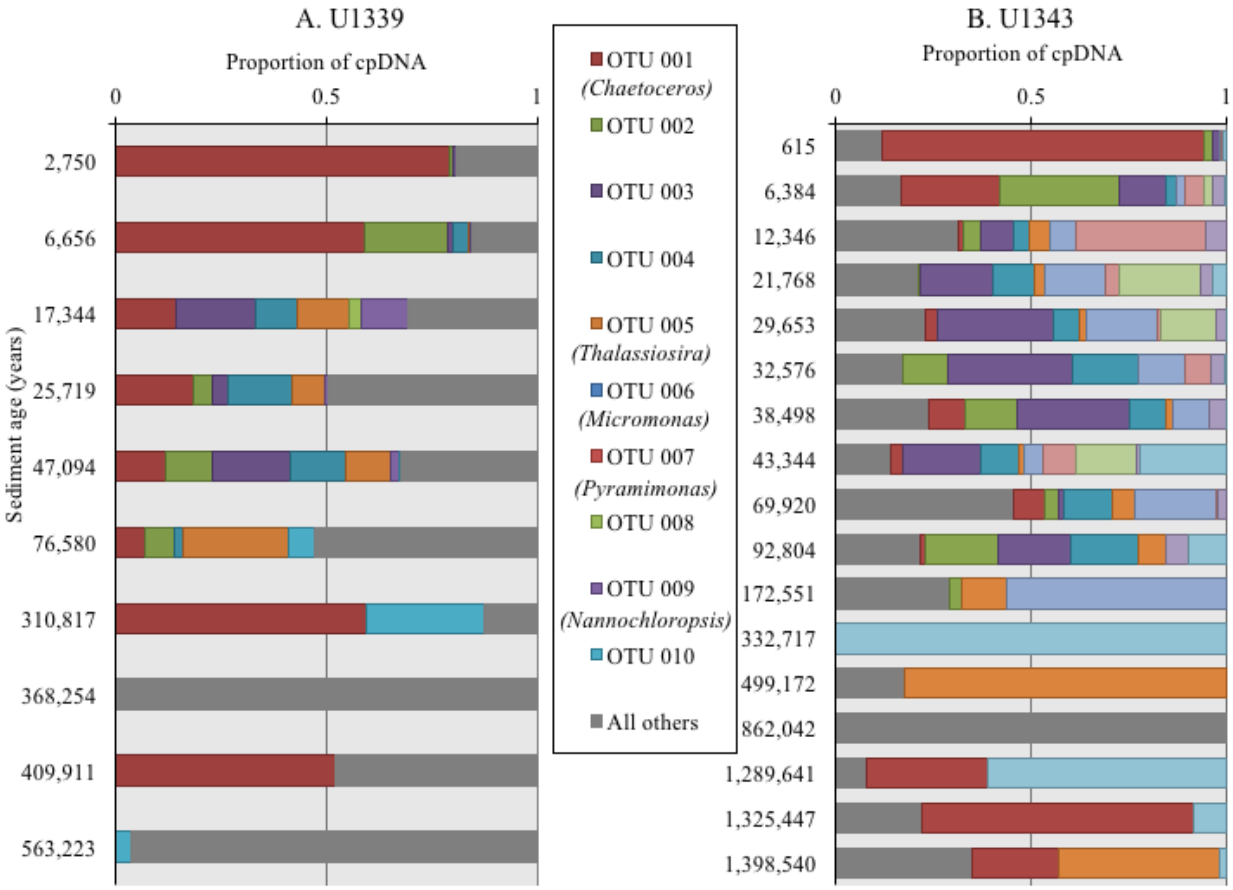
**Supplemental**
**Figure 1b.**




107

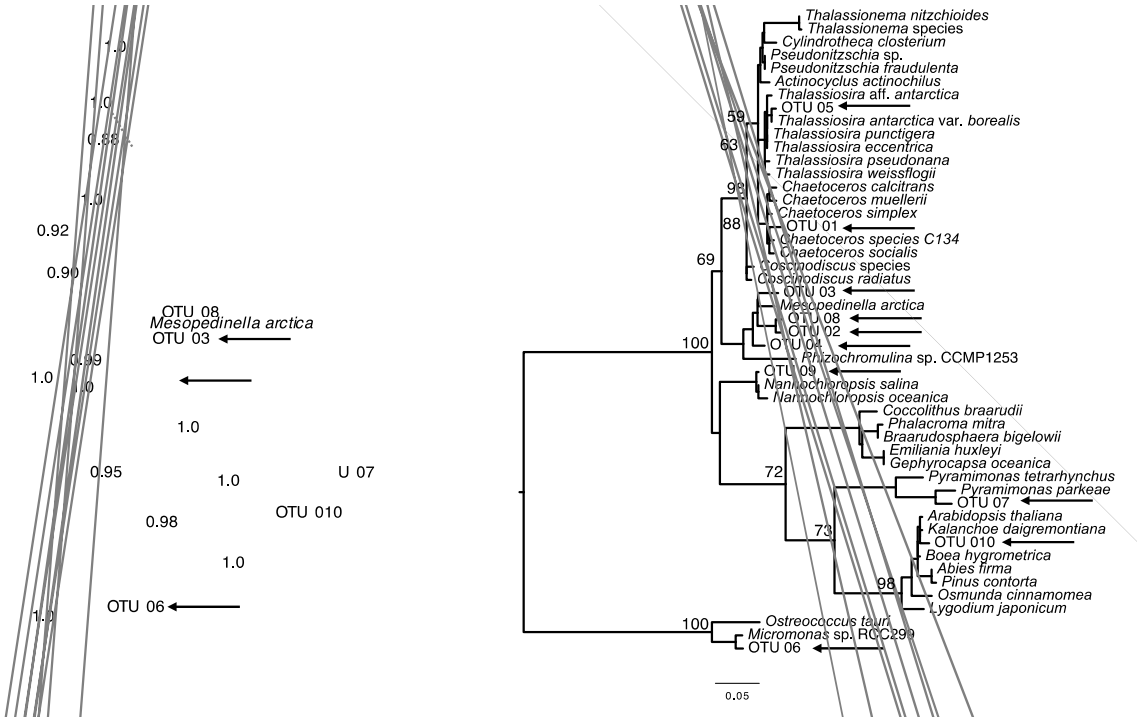
Supplemental Figure DR1. Profiles versus depth of individual chloroplast OTUs (Operational Taxonomic Units), as a fraction of total reads. The first 10 panels are the 10 most abundant OTUs (Operational Taxonomic Unit) overall (Figure 1a, 1b). The second to last panel includes the fraction of reads due to every remaining read outside of the top 10, including free-living cyanobacteria (Figure 1b). The last panel is for free-living cyanobacteria only (Figure 1b). Genus-level taxonomic assignments are given in parentheses when known. X-axes are logarithmic. The y-axis intercept represents the lowest value for which random re-sampling would still detect a given sequence. Zero values are plotted on the y-axis. Half-life decay curves are shown for comparison.

**Supplemental Figure 2.**




Supplemental Figure DR2. The breakdown, versus sediment age, of the 10 most commonly sequenced cpDNA OTUs (Operational Taxonomic Units). Each depth is normalized to 1, regardless of the fraction of chloroplast DNA seen at that depth. Genus-level taxonomic assignments are given in parentheses when known (Figure 2). Left: Site U1339. Right: Site U1343. Legend is the same for both.

Supplemental Figure DR3.




Supplemental Figure DR3. Comparison of phylogenetic trees. Left: Tree constructed with Bayesian inference (MrBayes 3.2; Ronquist *et al*., 2012). Clade confidence values are shown, and arrows indicate sequences from this studies. Right: Maximum likelihood tree constructed with PhyML (Guindon and Gascuel, 2003) using 1000 boostraps. Boostrap support values greater than 50 are shown for the major branch points.

133 Supplemental Table DR1. Sample data. Age model is from Takahashi et al. (2011).

134

| **Sample site** | **Sample depth (m CCSF)** | **Sediment age (yr)** | **Cleaned dataset size (number of reads)** | **Reads identified as chloroplast** |
|---|---|---|---|---|
| U1343 | 0.16 | 615 | 162,247 | 1,043 (0.64%) |
| U1343 | 1.66 | 6,384 | 236,990 | 1,516 (0.64%) |
| U1343 | 3.21 | 12,346 | 190,551 | 434 (0.23%) |
| U1343 | 5.66 | 21,768 | 183,113 | 248 (0.14%) |
| U1343 | 7.71 | 29,653 | 264,598 | 548 (0.21%) |
| U1343 | 8.47 | 32,576 | 150,567 | 147 (0.10%) |
| U1343 | 10.01 | 38,498 | 216,245 | 559 (0.26%) |
| U1343 | 11.27 | 43,344 | 234,277 | 250 (0.11%) |
| U1343 | 18.18 | 69,920 | 324,963 | 415 (0.13%) |
| U1343 | 24.13 | 92,804 | 196,478 | 70 (0.04%) |
| U1343 | 44.87 | 172,551 | 322,319 | 62 (0.02%) |
| U1343 | 86.51 | 332,717 | 231,588 | 13 (0.01%) |
| U1343 | 129.79 | 499,172 | 257,233 | 34 (0.01%) |
| U1343 | 224.14 | 862,042 | 197,960 | 27 (0.01%) |
| U1343 | 335.32 | 1,289,641 | 241,626 | 154 (0.06%) |
| U1343 | 344.63 | 1,325,447 | 288,110 | 250 (0.09%) |
| U1343 | 363.64 | 1,398,540 | 250,859 | 295 (0.12%) |
| U1339 | 0.88 | 2,750 | 215,839 | 1,586 (0.73%) |

| | | | | |
|---|---|---|---|---|
| U1339 | 2.13 | 6,656 | 162,240 | 1,139 (0.70%) |
| U1339 | 5.55 | 17,344 | 299,394 | 912 (0.30%) |
| U1339 | 8.23 | 25,719 | 244,312 | 604 (0.25%) |
| U1339 | 15.07 | 47,094 | 261,778 | 236 (0.09%) |
| U1339 | 32.12 | 76,580 | 268,696 | 100 (0.04%) |
| U1339 | 88.36 | 310,817 | 275,530 | 133 (0.05%) |
| U1339 | 104.87 | 368,254 | 252,962 | 118 (0.05%) |
| U1339 | 217.58 | 563,223 | 217,460 | 57 (0.03%) |

135

SUPPLEMENTAL REFERENCES

Gouy, M., Guindon, S., and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27, 221–4. doi:10.1093/molbev/msp259.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704. doi:10.1080/10635150390235520.

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–98. doi:10.1111/j.1462-2920.2010.02193.x.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G., and Gibson, T. J. (1998). Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* 23, 403–5.

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* 79, 5112–5120. doi:10.1128/AEM.01043-13.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012, MrBayes 3.2: efficient Bayesian

phylogenetic inference and model choice across a large model space. Systematic Biology, v. 61, p. 539–542, doi: 10.1093/sysbio/sys029.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. a., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi:10.1128/AEM.01541-09.

Takahashi, K., Ravelo, A. C., Alvarev Zarikian, C. A., and Scientists, E. 323 (2011). *Proc. IODP, 323*. Tokyo: Integrated Ocean Drilling Program Management International, Inc. doi:10.2204/iodp.proc.323.2011.