A census-based estimate of Earth's bacterial and archaeal diversity - Additional Methods -

Stilianos Louca[1,2,3,4,*], Florent Mazel[3,5], Michael Doebeli[3,4,6] & Laura Wegener Parfrey[3,4,5]

[1]*Department of Biology, University of Oregon, USA*
[2]*Institute of Ecology and Evolution, University of Oregon, USA*
[3]*Biodiversity Research Centre, University of British Columbia, Canada*
[4]*Department of Zoology, University of British Columbia, Canada*
[5]*Department of Botany, University of British Columbia, Canada*
[6]*Department of Mathematics, University of British Columbia, Canada*
[*]Corresponding author

This document describes additional Materials and Methods for the Global Prokaryotic Census (GPC), not included in the original paper [1]. The material below was not needed for the original paper, but may be useful for interpreting some of the additional components and analyses included in the GPC.

Construction of GPC phylogenetic trees

To create a draft phylogenetic tree from the GPC OTUs (97% similarity), we proceeded as follows. Representative sequences of OTUs were aligned using the QIIME script `parallel_align_seqs_pynast.py` [2] and using a subset of the SILVA alignments (obtained by clustering SILVA at 90% similarity) as a template. A total of 694,841 OTUs could be successfully aligned, and failed OTUs were omitted from tree building. Nucleotide positions in the alignments with $> 95\%$ gaps, as well as the top 5% most entropic positions, were removed from the alignments. After nucleotide filtering, any sequences with fewer than 100 non-gap positions were omitted, yielding a total of 639,347 OTUs for tree building. The taxonomic identities of OTUs at the domain, phylum, class and order level (where available) were used to define constraints for `FastTree` [3], by constraining all OTUs within a taxon to be on a single side of a bifurcation. A total of 813 constraints were defined. Using the alignments and the taxonomic constraints, we generated a phylogenetic tree with `FastTree` v2.1.10 (options "`-spr 4 -gamma -no2nd -constraintWeight 1000`"). The tree was rerooted so that Bacteria and Archaea split at the root. Tips not placed consistently with this splitting (e.g., OTUs identified as Archaea but placed on the Bacterial branch), were omitted from the tree. The final tree, comprising 605,762 tips, is provided in Newick format. A similar approach was used for constructing the tree at 99% clustering similarity.

References and Notes

[1] Louca S, Mazel F, Doebeli M, Parfrey WL. A deep census of Earth's Bacteria and Archaea. PLOS Biology. In review; .

[2] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Meth. 2010; 7(5):335–336.

[3] Price MN, Dehal PS, Arkin AP. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009; 26(7):1641–1650. doi:https://doi.org/10.1093/molbev/msp077.