

Terpene synthase gene amplicons from subseafloor sediment

Ruth L. Schmidt^a, Tatsuhiko Hoshino^b, Yuki Morono^b, Julien Tremblay^c, Dana Ulanova^{d,e,#}

^aCentre Armand-Frappier Santé Biotechnologie, Institut national de la recherche scientifique, Laval H7V 1B7, Québec, Canada.

^bGeomicrobiology Group, Kochi Institute for Core Sample Research, Institute for Extra-cutting-edge Science and Technology Avant-garde Research, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Monobe B200, Nankoku, Kochi 783-8502, Japan

^cEnergy, Mining and Environment Research Centre, National Research Council Canada, 6100 Royalmount avenue, Montreal, Canada, H4P2R2.

^dDepartment of Marine Resource Science, Faculty of Agriculture and Marine Science, Kochi University, Monobe B200, Nankoku, Kochi 783-8502, Japan

^eCenter for Advanced Marine Core Research, Kochi University, Monobe B200, Nankoku, Kochi 783-8502

#Address correspondence to

Dana Ulanova, ulanova@kochi-u.ac.jp

We sequenced putative terpene synthase (TS) gene fragments. These fragments were amplified using subseafloor sediment sample collected off Shimokita Peninsula, Japan, and primers targeting bacterial geosmin and 2-methylisoborneol synthases. Amplicons were sequenced using Illumina platform. This technical note describes details of raw read processing and OTU taxonomic annotation.

Processing of raw reads obtained in next-generation sequencing of PCR amplicons

The obtained sequences were processed in the AmpliconTagger v1.3.0 pipeline (1). In brief, raw reads were scanned for sequencing adapters and PhiX spike-in sequences. Remaining paired-end reads were filtered for quality so that sequences that had 1 N or more; had average quality scores lower than 25; or had more than 60 nucleotides with a Phred quality score lower than 15 were removed. The remaining paired-end reads were assembled into their full amplicon sequences using their overlapping common parts using FLASH v1.2.11 (2). Assembled sequences were de-replicated at 100% identity and clustered at 99% identity using VSEARCH v2.7.1 (3). Clusters having less than 10 reads were discarded. Remaining clusters were clustered again at 97% identity to generate Operational Taxonomic Units (OTUs). These OTUs were filtered for chimeras using VSEARCH's implementation of UCHIME *de novo*. Custom geosmin and 2-MIB reference databases were used. Briefly, for each data type (geosmin and 2-MIB), final OTUs were blasted against the NCBI nt database. Hits with an e-value < 1e-20, alignment length >= 100 and alignment percentage >= 60 were kept to build the RDP training set. RDP training sets were generated as previously described (<https://github.com/jtremblay/RDP-training-sets>). The RDP classifier (4) was then used with this training set to assign a taxonomic lineage to each OTU. The RDP classifier gave

a score (0 to 1) to each taxonomic depth of each OTUs. Each taxonomic depth having a score ≥ 0.5 was kept to reconstruct the final lineage. Taxonomic lineages assigned to bacterial or archaeal lineages were combined with the OTUs abundance matrix obtained above to generate a raw OTU table. The sequence-specific primer sequences were removed using MEGA v.7.0.26 (5).

The obtained OTUs were annotated using blastX algorithm (6) and NCBI GenBank non-redundant protein sequence database (nr, accessed October – November 2020). Non-TS OTUs were excluded.

References

1. Tremblay J, Yergeau E. 2019. Systematic processing of ribosomal RNA gene amplicon sequencing data. *Gigascience* 8:1–14.
2. Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27:2957–2963.
3. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: A versatile open source tool for metagenomics. *PeerJ* 2016:e2584.
4. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl*

Environ Microbiol 73:5261–5267.

5. Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33:1870–1874.
6. Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. Nat Genet 1993 33 3:266–272.