nature geoscience

Supplementary information

https://doi.org/10.1038/s41561-026-02054-6

Molecular evidence for a relict marine community in an Antarctic Dry Valleys subglacial brine-fed system

In the format provided by the authors and unedited

Metatranscriptomics

Total RNA was extracted from frozen, RNAlater preserved sample using the NucleoBond RNA Soil kit (Macherey-Nagel, Düren, Germany); cleanup and DNase treatment by RNA Clean & Concentrator (Zymo Research, Irvine, CA, USA). Then, total RNA was subjected to amplification and cDNA synthesis using the Ovation RNA-Seq System V2 (TECAN, Redwood City, CA, USA). Double stranded cDNA was fragmented using Covaris E210 system with the target size of 300bp. Libraries were constructed by Ovation Ultralow System V2 (TECAN) following the manufacturer's protocol. Ampure XP beads (Beckman Coulter, Brea, CA, USA) were used for final library purification. Library quality was analyzed on a 2200 TapeStation System with Agilent High Sensitivity DNA 1000 ScreenTape System (Agilent Technologies, Santa Clara, CA, USA). Resulting libraries were subjected to paired-end Illumina sequencing (NovaSeq 6000, PE150; Illumina, San Diego, CA, USA) to an average depth of 39.8 million raw reads per library.

MetaT assemblies were generated using a modified RNAseq Annotation Pipeline [1]. Raw reads were adapter- and quality-trimmed using fastp v0.23.2 trimmomatic option [2, 3]. Read quality was assessed by FastQC v0.12.1 [https://www.bioinformatics.babraham.ac.uk/projects/fastqc]. rRNA-matched trimmed were further filtered out using BBduk (https://jgi.doe.gov/data-and-tools/software-tools/bbtools/; rDNA databases PR2 v4.12.0, RFAM v14.1, SILVA v138). Transcriptomes were assembled using MEGAHIT v1.2.9 [4] and SPAdes v3.15.5 (rnaSPAdes mode) [5], then combined at 95% sequence identity using linclust of mmseqs2 release 14 (--min-seq-id 0.95 -c 0.8 --cov-mode 1 --cluster-mode 2) [6]. Open reading frames (ORFs) were determined and translated by GeneMarkS v5.1 [7] followed by annotation against common protein domain databases Pfam v37.0, PANTHER v18.0, TIGRFAM v15.0, KEGG v29-10-2024, and EggNOG v5 using a combination of tools (diamond v2.1.9, eggNOG-mapper v2.1.10, Interproscan v5.69-101.0, kofamscan v1.3.0 [8-12]. Lineage Probability Index (LPI) was calculated from top 100 diamond blastp hits by dividing the sum of probabilities of each taxonomic term by the normalization factor corresponding to its taxonomic level in the lineage and choosing the term with the highest index [13]. ORF read counts were obtained by Bowtie2 v2.5.0 (--local mode, otherwise as default) [14]. Normalization was performed by the TMM method of edgeR v4.6.3 [15]. Statistical

significance of differential expression was assessed by Wilcoxon rank sum test with Benjamini-Hochberg adjustment for multiple testing (*p*-adj level 0.05).

Haplotype maps were generated using V4 ASVs and publicly available reference data. Briefly, for each ASV, 20 best hits were searched using blastn [16] in PR2 database v5.0.0 [17] max e-value 1e-5. MAFFT v7.490 [18] default parameters was employed to align the ASVs with these references. Phylogenies were inferred using the HKY85 matrix in PhyML v3.3.20180621 [19] with no bootstrapping to remove chimeric and divergent sequences. Endemic ASVs were then defined based on an abundance threshold (≥50% reads in Blood Falls samples); ASVs not meeting this threshold were considered cosmopolitan (see popart-preparation.py). PopART v1.7 [20] was used to assess haplotype networks using Median Joining with epsilon set to 0. The input files contained log-transformed abundance for each ASV to enable weighted haplotype maps.

1. R Bertrand, E. M. *et al*. Phytoplankton-bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. *Proc Natl Acad Sci U S A* **112**, 9938–9943 (2015).

2. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinforma Oxf Engl 2018; 34: i884–i890.

3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30: 2114–2120.

4. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676 (2015).

5. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012; 19: 455-77.

6. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35, 1026–1028 (2017).

7. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 2001; 29(12):2607-18.

8. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinforma Oxf Engl 2020; 36: 2251–2252.

9. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods 2015; 12: 59–60.

10. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 2021; 38: 5825–5829.

11. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019; 47: D309–D314.

12. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinforma Oxf Engl 2014; 30: 1236–1240.

13. Delaye L, Vargas C, Latorre A, Moya A. Inferring Horizontal Gene Transfer with DarkHorse, Phylomizer, and ETE Toolkits. Methods Mol Biol Clifton NJ 2020; 2075: 355–369.

14. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012).

15. Robinson, MD, McCarthy, DJ, Smyth, GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-140.

16. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. J Mol Biol 215, 403–410 (1990).

17. Guillou, L. et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res 41, D597-604 (2013).

18. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30, 772–780 (2013).

19. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003 Oct;52(5):696-704.

20. Leigh, JW, Bryant D (2015). PopART: Full-feature software for haplotype network construction. Methods Ecol Evol 6(9):1110–1116.