environmental microbiology reports

Environmental Microbiology Reports (2020) **12**(6), 651–655 doi:10.1111/1758-2229.12880

Brief Report

Metagenome assembled-genomes reveal similar functional profiles of CPR/Patescibacteria phyla in soils

Leandro Nascimento Lemos, [1*]
Lokeshwaran Manoharan, [2,3]
Lucas William Mendes, [1]
Andressa Monteiro Venturini, [1]
Victor Satler Pylro [4*] **and Siu Mui Tsai** [1]

[1]*Cell and Molecular Biology Laboratory, Center for Nuclear Energy in Agriculture CENA, University of Sao Paulo USP, Piracicaba, Brazil.*
[2]*Division of Archaea Biology and Ecogenomics, Department of Ecogenomics and Systems Biology, University of Vienna, Vienna, Austria.*
[3]*National Bioinformatics Infrastructure Sweden (NBIS), Lund University, Lund, Sweden.*
[4]*Microbial Ecology and Bioinformatics Laboratory, Department of Biology, Federal University of Lavras (UFLA), Lavras, Brazil.*

Abstract

Soil microbiome is one of the most heterogeneous biological systems. State-of-the-art molecular approaches such as those based on single-amplified genomes (SAGs) and metagenome assembled-genomes (MAGs) are now improving our capacity for disentailing soil microbial-mediated processes. Here, we analysed publicly available datasets of soil microbial genomes and MAG's reconstructed from the Amazon's tropical soil (primary forest and pasture) and active layer of permafrost, aiming to evaluate their genome size. Our results suggest that the Candidate Phyla Radiation (CPR)/Patescibacteria phyla have genomes with an average size fourfold smaller than the mean identified in the RefSoil database, which lacks any representative of this phylum. Also, by analysing the potential metabolism of 888 soil microbial genomes, we show that CPR/Patescibacteria representatives share similar functional profiles, but different from other microbial phyla and are frequently neglected in the soil microbial surveys. Finally, we argue that the use of MAGs may be a better choice over SAGs to expand the soil microbial databases, like RefSoil.

Received 12 March, 2020; accepted 14 August, 2020. *For correspondence. E-mail llemos@usp.br; E-mail victor.pylro@ufla.br; Tel./Fax +55(31) 3829-5176.

Introduction

Many hypotheses may explain the complexity and high diversity of soil microbiomes. In the genomic context, Raes *et al*. (2007) argued that each habitat selects a specific range of microbial genome sizes, regarding the environment stability (Angly *et al*., 2009). More stable environments select microorganisms, such as parasites (e.g., *Mycoplasma pneumoniae*) (Himmelreich *et al*., 1996) and symbionts (McCutcheon, 2010), with small genomes and less non-redundant functions (Morris *et al*., 2012). On the other hand, unstable and complex environments, like soil, favour microorganisms with larger genomes and accessory genes, with greater metabolic versatility and the ability to survive and acclimate in a changing-environment with diverse but limited resources (Konstantinidis and Tiedje, 2005; Dini-Andreote *et al*., 2012).

The vast majority of soil microorganisms have not yet been cultivated, given our limitation to simulate all required conditions for microbial growth. As a consequence, several soil microbial functions remain unknown, resulting in a break in the link between the microbial taxonomy and soil processes. The recently proposed group of bacteria, Candidate Phyla Radiation (CPR)/Patescibacteria (Brown *et al*., 2015), has only one representative already cultivated, the strain TM7x (He *et al*., 2015). However, they may encompass from 15% (Brown *et al*., 2015) to nearly 50% (Hug *et al*., 2016) of the diversity in the Bacteria domain. The lack of biosynthetic capabilities (Brown *et al*., 2015) and potential for co-metabolism interdependencies (He *et al*., 2015; Lemos *et al*., 2019) is a common genomic trait shared by all CPR/Patescibacteria members, as a consequence of the small size of their genomes (usually <1.5 Mbp). These biological traits could prove to be the major challenge for cultivating and uncovering the functional potential of these organisms.

Here, we applied an integrated meta-analysis of publicly available genomes and metagenomes, aiming to explore the genome size features and potential functional

profiles of soil microorganisms using two data sets: (i) the RefSoil database (Choi *et al*., 2017) and (ii) a metagenome-assembled genomes (MAGs) dataset from a permafrost thaw gradient (Woodcroft *et al*., 2018), which is the most complete natural soil environment database to date. To contrast the MAGs with small genome size described by Woodcroft *et al*. (2018) (permafrost), we also used six MAGs from tropical Amazon soils, including the description of two new CPR/Patescibacteria (Supporting Information SI) and four CPR/Patescibacteria MAG's already described by Kroeger *et al*. (2018). We estimated the functional profile of 888 soil microbial genomes (including those from RefSoil, the most complete and curated soil microbial genomes database, and all CPR/Patescibacteria MAG's above described), to better understand the correlation between the genome size and the profile of potential functions performed by soil microorganisms, using the clusters of orthologous groups (COGs) genome annotation and a set of multivariate statistics (Supporting Information SI). We also used the PhenDB software (Feldbauer *et al*., 2015) to test whether the soil CPR/Patescibacteria were potential symbionts. PhenDB uses machine learning and statistics predictions to infer phenotypic traits, based on a manual knowledge curation from the scientific literature and COG profiles (Feldbauer *et al*., 2015).

Results and discussion

Our analysis revealed that the average size of the microbial genomes available in the RefSoil was 4.5 ± 1.0 Mbp (Fig. 1A), but it lacks CPR/Patescibacteria representatives. Similarly, almost all MAGs retrieved from the thawing permafrost data set had their average genome size close to those observed in the RefSoil (Fig. 1B), but the CPR/Patescibacteria genomes available had an average genome size of 0.9 ± 0.2 Mbp, which is fourfold smaller than the mean identified in the RefSoil and in the thawing permafrost databases. We also found the same pattern when checking the genome size of the CPR/Patescibacteria from the tropical Amazon soils (Supporting Information Tables S1–S3; and Krueger *et al*. (2018)). The presence of small-sized bacterial genomes (*e.g*., CPR/Patescibacteria) may (i) confront the

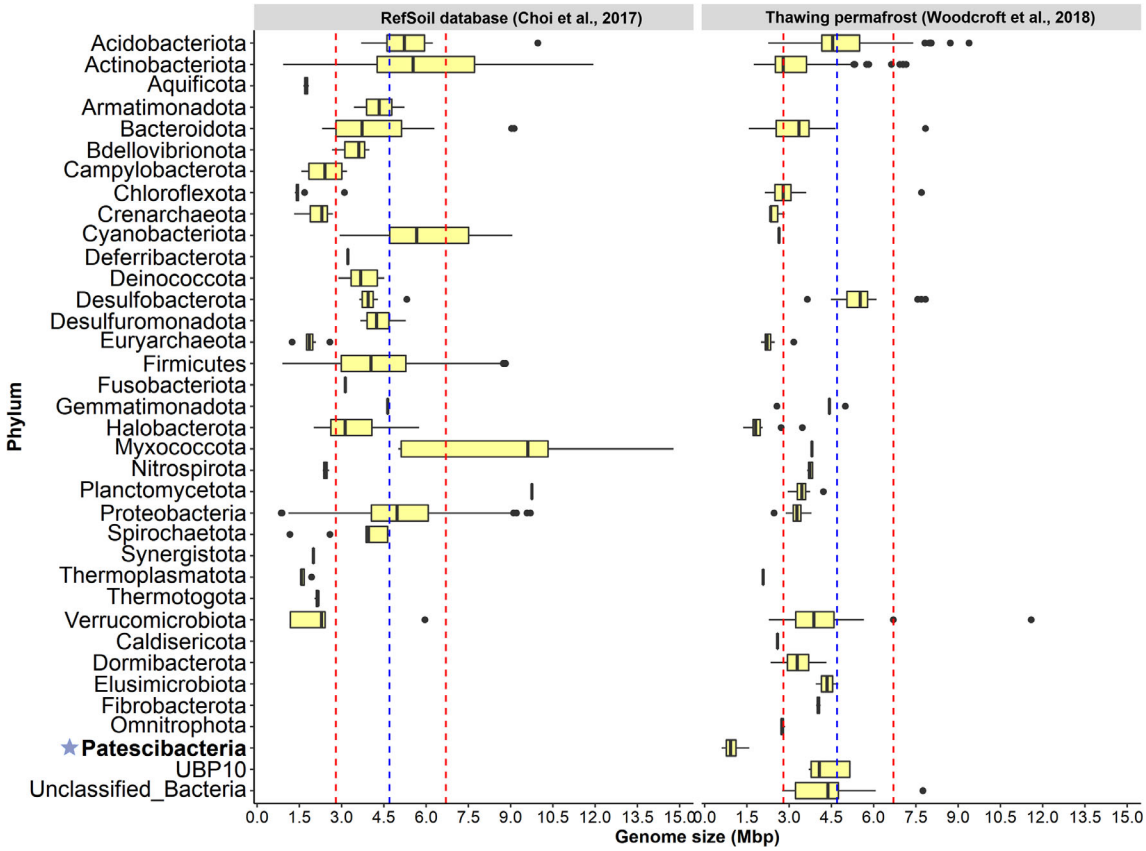

Fig. 1. Soil genome size distributions. (A) RefSoil database and (B) Metagenome-assembled genomes (MAGs) from thawing permafrost metagenomes. The Box Plot indicates data distribution through their quartiles. Blue and red dashed lines indicate the average and standard deviation of the genome size estimation from the RefSoil database. Extended lines from the boxplot indicate the variability outside the upper and lower quartiles, and the single dots represent out points.

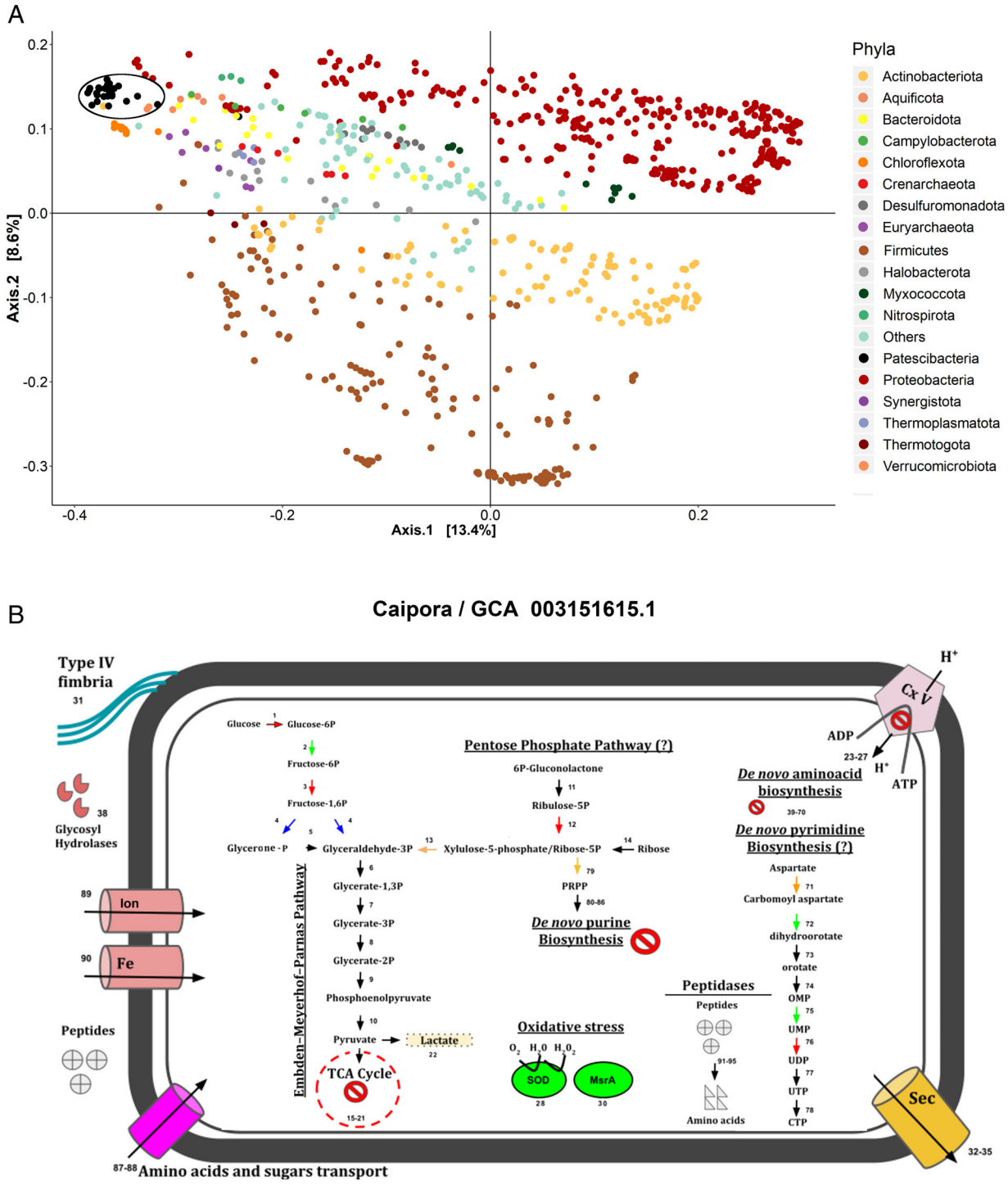

Fig. 2. Soil Patescibacteria/CPR functional genome profile.
A. Principal coordinates analysis (PcoA) based on the Jaccard similarity index (presence/absence of COG – clusters of orthologous groups) of soil microbial genomes and the CPR/Patescibacteria phyla functional profiles. Genomes from the same phyla have the same colour. The black-arrowed circle indicates the CPR/Patescibacteria cluster.
B. Schematic illustrating major putative microbial functions of the key pathways of soil CPR/Patescibacteria (Caipora MAG from Amazon cattle pasture and GCA_003151615.1 from thawing permafrost soils). The main pathways are illustrated by an underlined name and the question mark (?) symbol indicates incomplete and unclear pathways. Black arrows indicate proteins present in both genomes. Green and orange arrows indicate proteins present only in Caipora or GCA_003151615.1 MAGs, respectively. Red 'no entry' signs indicate missing pathways. Blue arrows indicate proteins present in other CPR/Patescibacteria genomes from Amazon or permafrost soils, which were used to complement detailed in this schematic illustration. A detailed list of genes encoded by soil Patescibacteria/CPR can be found in the Supplementary Table S1. TCA, tricarboxylic acid cycle; PRPP, Phosphoribosyl pyrophosphate; SOD, superoxide dismutase; MsrA, methionine sulfoxide reductase; Cx V, Electron Transport – Cytochrome; Sec, Protein.

hypothesis that large genomes favour soil microorganisms and (ii) highlight the importance of using complementary approaches to expand the RefSoil database, because some phyla with peculiar genomic traits, such as CPR/Patescibacteria, are 'hard-to-cultivate' and has been absent in the RefSoil database.

We evaluated the potential metabolism of 888 soil microbial genomes, and we observed that members of the CPR/Patescibacteria phylum have similar functional profiles among them, but different from most other soil bacteria (Fig. 2A). We identified that all soil CPR/Patescibacteria studied here lack the main metabolic pathways to generate energy, which have been identified in the most of heterotrophic soil bacteria, such as tricarboxylic acid cycle (TCA) pathway and electron transport chain to generate ATP. They likely ferment organic compounds via the Embden–Meyerhof–Parnas – glycolysis pathway (EMP) (Supporting Information Table S4), generating lactate as a final product (Fig. 2B). Furthermore, we identified some genes (*e.g.*, Transketolase) that could link EMP and pentose phosphate pathway (PPP), which supplies glyceraldehyde-3-phosphate for sugar degradation, as described to CPR/Patescibacteria (formerly known as OD1 phyla) (Wrighton *et al*., 2012). All metabolic predictions described here were based on sequence similarity of predicted proteins, since experimental validations would require cultivation of CPR/Patescibacteria. Some key-genes from central carbon metabolism (*e.g.*, phosphofructokinase) could be divergent or they have alternative proteins that still remain unknown (Castelle *et al*., 2018).

Overall, the soil CPR/Patescibacteria genomes described here also lacks genes required for the *de novo* biosynthesis of nucleotides, amino acids, and cofactors (Supporting Information Tables 2, 4 and Fig. 2B). However, transporter proteins may contribute to amino acids and sugars uptake for maintaining the essential metabolism. Peptidases were also identified and have potential to degrade small peptides into amino acids inside the cells. The presence of oxidative stress genes indicates a potential oxygen tolerance to survive under aerobic conditions. This metabolic dependence and reduced genome sizes are signatures of CPR/Patescibacteria (Castelle *et al*., 2018), and its potential cometabolism interdependencies with other microorganisms via a symbiotic lifestyle have been previously discussed (Lemos *et al*., 2019). This ecological trait may support the maintenance of small-genome bacteria in soil, contrasting the hypothesis that complex ecosystems select microorganisms with larger genomes and accessory genes (Raes *et al*., 2007).

Here, predictions based on machine learning algorithms indicated that the soil CPR/Patescibacteria genomes have a potential symbiotic lifestyle (Supporting Information Table S4), which may support their ecological success in complex environments, like soil. Furthermore, the metabolic limitations we found corroborates an episymbiotic lifestyle (Castelle *et al*., 2018; Lemos *et al*., 2019) or parasitism, as already described in the interaction between the obligate epibiont TM7x (CPR/Patescibacteria) and *Actinomyces odontolyticus* strain (XH001) in the oral environment, where TM7x kills its host (He *et al*., 2015). On the other hand, the soil bacterium '*Candidatus* Udaeobacter copiosus' (Verrucomicrobia) was recently described as being free-living and carrying a 2.81 Mbp genome (Brewer *et al*., 2017). Metabolic predictions indicated that C. *U. copiosus* can keep a reduced genome by acquiring costly amino acids and vitamins from the environment (Brewer *et al*., 2017). A few numbers of genomes shorter than 1.5 Mbp are available on RefSoil (Supporting Information Table S3), and all of them exhibit a parasitic lifestyle (*e.g.*, *Neorickettsia* and *Tropheryma*). These findings reinforce the idea that soil CPR/Patescibacteria could also be associated with a symbiotic/parasitic lifestyle, which may support their ecological success in this environment, besides suggesting their role in the soil microbial community assemblage and structure.

Choi *et al*. (2017) recommended the use of single-cell methods for expanding the RefSoil database. However, the single-cell genomes they presented did not offer the quality recommended by the minimum information about a single amplified genome (MISAG) standards (*e.g.*, high-quality draft genomes: >90% of completeness and <5% of contamination) (Bowers *et al*., 2017) (Supporting Information Table S5). We argue that binning methods may be a better choice over single-cell approaches for expanding the RefSoil database, leading to a more complete and informative soil microbial reference database.

Concluding remarks

In summary, the small-sized genome is a peculiar trait of the soil CPR/Patescibacteria phyla members. Here, we highlight the wide range of terrestrial environments within the radiation of this bacteria group, opening new opportunities for understanding their ecological role in soils. We showed that even under distinct climate conditions (tropical soils and permafrost), CPR/Patescibacteria show similar functional profiles, and lack essential biosynthetic capacities (e.g. *de novo* amino acids and nucleotide biosynthesis). Further studies are required to elucidate the ecology of CPR/Patescibacteria, such as the design of new 16S rRNA gene primers to comprehensively measure the abundance, diversity, and distribution of CPR/Patescibacteria in soils worldwide. Additionally, metatranscriptomics and/or RNA-SIP based methods may increase our understanding of their metabolic functions under distinct environmental conditions.

Acknowledgements

This study was financed in part by the São Paulo Research Foundation (FAPESP processes 2014/50320-4, 2016/18215-1, 2017/24037-1, 2015/13546-7 and 2017/09643-2), National Council for Scientific and Technological Development (CNPq grant 140032/2015-0, 161931/2015-4 and 311008/2016-0), and the Coordination for the Improvement of Higher Education Personnel - Brasil (CAPES) - Finance Code 001. We acknowledge René Rachou Institute - Fiocruz Minas, Bioinformatics Platform for computational logistic and support. We thank all members of the Archaea Biology and Ecogenomics Division for continuous discussions on the microbial genomes, in particular Christa Schleper and Melina Kerou. This work was also supported by the Brazilian Microbiome Project (http://www.brmicrobiome.org).

Ethics statement

This article does not contain any studies with human participants or animals performed by any of the authors.

Data availability

The metagenome-assembled genomes from amazon soil dataset are publicly available in the DDBJ/EMBL/Genbank databases under accession number WARW00000000 and WARV00000000.

References

Angly, F.E., Willner, D., Prieto-Davó, A., Edwards, R.A., Schmieder, R., Vega-Thurber, R., *et al*. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* **5**: e1000593.

Bowers, R.M., *et al*. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**: 725–731.

Brewer, T.E., Handley, K.M., Carini, P., Gilbert, J.A., and Fierer, N. (2017) Genome reduction in an abundant and ubiquitous soil bacterium "Candidatus Udaeobacter copiosus". *Nat Microbiol* **2**: 16198.

Brown, C., *et al*. (2015) Unusual biology across a group comprising more than 15% of domain bacteria. *Nature* **523**: 208–211.

Castelle, C.J., Brown, C.T., Anantharaman, K., Probst, A.J., Huang, R.H., and Banfield, J.F. (2018) Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol* **16**: 629–645.

Choi, J., Yang, F., Stepanauskas, R., Cardenas, E., Garoutte, A., Williams, R., *et al*. (2017) Strategies to improve reference databases for soil microbiomes. *ISME J* **11**: 829–834.

Dini-Andreote, F., Andreote, F.D., Araújo, W.L., Trevors, J.T., and van Elsas, J.D. (2012) Bacterial genomes: habitat specificity and uncharted organisms. *Microb Ecol* **64**: 1–7.

Feldbauer, R., Schultz, F., Horn, M., and Rattei, T. (2015) Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics* **16**: S1.

He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.Y., *et al*. (2015) Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc Natl Acad Sci* **112**: 244–249.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C., and Herrmann, R. (1996) Complete sequence analysis of the genome of the bacterium mycoplasma pneumoniae. *Nucleic Acids Res* **24**: 4420–4449.

Hug, L.A., *et al*. (2016) A new view of the tree of life. *Nat Microbiol* **1**: 16048.

Konstantinidis, K.T., and Tiedje, J.M. (2005) Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**: 6258–6264.

Kroeger, M., Delmont, T.O., Eren, A.M., Meyer, K.M., Guo, J., and Khan, K. (2018) New biological insights into how deforestation in Amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. *Front Microbiol* **9**:1635. https://doi.org/10.3389/fmicb.2018.01635.

Lemos, L.N., Medeiros, J.D., Dini-Andreote, F., Fernandes, G.R., Varani, A.M., Oliveira, G., and Pylro, V.S. (2019) Genomic signatures and co-occurrence patterns of the ultra-small Saccharimonadia (phylum CPR/Patescibacteria) suggest a symbiotic lifestyle. *Mol Ecol* **28**: 4259–4271.

McCutcheon, J.P. (2010) The bacterial essence of tiny symbiont genomes. *Curr Opin Microbiol* **13**: 73–78.

Morris, J.J., *et al*. (2012) The black queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**: e00036–e00012.

Raes, J., Korbel, J.O., Lercher, M.J., von Mering, C., and Bork, P. (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* **8**: R10.

Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoeset, N.C., *et al*. (2012). Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science*. **337**(6102):1661–1665.

Woodcroft, B.J., Singleton, C.M., Boyd, J.A., Evans, P.N., Emerson, J.B., Zayed, A.A.F., *et al*. (2018) Genome-centric view of carbon processing in thawing permafrost. *Nature* **560**: 49–54.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1. Supporting information
Appendix S2. Tables.