UNIVERSIDADE DE SÃO PAULO
CENTRO DE ENERGIA NUCLEAR NA AGRICULTURA

LEANDRO NASCIMENTO LEMOS

# Integrative and *in silico* modeling of multi-omics data of Archaea and Bacteria phyla in Amazon soils

Piracicaba
2019

LEANDRO NASCIMENTO LEMOS

# Integrative and *in silico* modeling of multi-omics data of Archaea and Bacteria phyla in Amazon soils

Tese apresentada ao Centro de Energia Nuclear na Agricultura da Universidade de São Paulo para obtenção do título de Doutor em Ciências

Área de Concentração: Biologia na Agricultura e no Ambiente

Orientador: Prof. Dra. Tsai Siu Mui

Piracicaba

2019

Dados Internacionais de Catalogação na Publicação (CIP)

**Seção Técnica de Biblioteca - CENA/USP**

Lemos, Leandro Nascimento

Modelagem integrativa e *in silico* de dados multi-ômicos de filos de Archaea e Bacteria em solos amazônicos / Integrative and *in silico* modeling of multi-omics data of Archaea and Bacteria phyla in Amazon soils / Leandro Nascimento Lemos; orientadora Tsai Siu Mui. - - Piracicaba, 2019.

111 p. : il.

Tese (Doutorado – Programa de Pós-Graduação em Ciências. Área de Concentração: Biologia na Agricultura e no Ambiente) – Centro de Energia Nuclear na Agricultura da Universidade de São Paulo.

1. Bioinformática 2. Ecologia microbiana 3. Microbiologia do solo 4. Microbioma 5. Solo tropical - Amazônia I. Título.

CDU 579.26 + 575.112

**Elaborada por:**
Marilia Ribeiro Garcia Henyei
CRB-8/3631
Resolução CFB Nº 184 de 29 de setembro de 2017

***For my mother and my father.***

# Acknowledgements

I am grateful to my **mother Maria** and **father Arize**, who have provided me through moral and emotional support in my life. I am also grateful to my other family members and friends who have supported me along the way.

I would like to express my sincere gratitude to my advisor **Profa. Tsai Siu Mui** for the continuous support of my Ph.D study and related research, for her patience, motivation and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would like to express my gratitude to my internship advisor, **Profa. Christa Schleper**, for the opportunity, support and guidance during the scientific internship. I also thank all members of the Archaea Biology and Ecogenomics Division for continuous discussions on the microbial genomes, in particular Melina Kerou and Lokeshwaran Manoharan.

I thank **Profa. Karoline Faust** for critical reading of the Study 5 of this thesis.

I thank **Prof. Victor Pylro and Dr. Lucas William Mendes** for critical reading of the Study 3, 4 and 5 of this thesis.

To FAPESP (São Paulo Research Foundation), for the doctoral scholarship (2016/18215-1), the BEPE scholarship (2017/24037-1) and funding for the experimental work (2014/50320-4).

To CNPq (National Council for Scientific and Technological Development), for the doctoral scholarship (161931/2015-4).

" (…) get a large typewriter
and as the footsteps go up and down
outside your window

hit that thing
hit it hard

make it a heavyweight fight

make it the bull when he first charges in

and remember the old dogs
who fought so well:
Hemingway, Celine, Dostoevsky, Hamsun.

If you think they didn't go crazy
in tiny rooms
just like you're doing now.
(...)

(How to be a great writer - Charles Bukowski [1920 – 1994])

## ABSTRACT

LEMOS, L. N**. Integrative and *in silico* modeling of multi-omics data of Archaea and Bacteria phyla in Amazon soils**. 2019. 111 p. Tese (Doutorado em Ciências) - Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, 2019.

A large scale of multi-omics datasets, that have been generated in various microbiome projects, such as the "Dimensions US-BIOTA-Sao Paulo: Collaborative Research: Integrating Dimensions of Microbial Biodiversity across Land Use Change in Tropical Forests (FAPESP 2014/50320-4)", require a quantitative interpretation of the interactions among the three dimensions (phylogenetic, genetic and functional) of microbial diversity. This thesis was based on the need to use an integrated computational approach to investigate the role of Bacteria and Archaea in Amazon soils, using massive sequencing technologies, bioinformatics and reconstruction of genomes from metagenomes (MAGs) to understand not only the diversity, but also the evolution, metabolism and biogeography distribution. Firstly, in the Chapter 1 we introduced an overview of the main topics covered in this thesis. Then, we outlined the main approaches applied for microbiome studies based on high-throughput sequencing technologies and we introduced the most commonly used strategies for bioinformatics analyses and data integration. In the third chapter, the application of an integrative approach allowed us to discover that the nitrogen-related traits associated with nitrification in Archaea (e.g. ammonia oxidation) metabolism seems to be a derived character and emerged late in the diversification of the Thaumarchaeota (Archaea) group. Furthermore, after analyzing more than 27,000 public environmental samples, we discovered that the non-ammonia oxidizing clade has habitat-specific subgroups (e.g. Group 1.1c is more specific for soils and non-saline sediments). We also described the first Thaumarchaeota genome from tropical soils. Additionally, in the fourth chapter we discovered that the small-sized genome was a trait of the new CPR/Patescibacteria (Bacteria) phyla in cattle-pasture of Amazon soils. We also expanded the range of environments within the radiation of this new bacterial group appears and highlight the importance of MAGs methods for the expansion of reference databases. Lastly, chapter five explored the effects of forest-to-pasture conversion and an increase in soil moisture levels on Archaea composition in Amazon soils. Our results indicated that the community alterations caused by the higher soil moisture levels are most pronounced in the pasture, where communities were more sensitive, enhancing the potential of methanogenesis, while forest may act as buffers during the rainy season and harbors more stable communities. This thesis highlights the importance of the use of advanced bioinformatics tools and integrated computational approaches for a better understanding of the evolutionary processes, metabolic pathways and environmental distribution of complex soil microbiome members.

Keywords: Microbiome. Microbial Ecology. Soil microbiology. Bioinformatics. Amazon tropical soils.

## RESUMO

LEMOS, L. N. **Modelagem integrativa e *in silico* de dados multi-ômicos de filos de Archaea e Bacteria em solos amazônicos**. 2019. 111 p. Tese (Doutorado em Ciências) - Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, 2019.

A enorme quantidade de dados multi-ômicos gerados em projetos de microbiomas, tais como o projeto "Dimensões US-BIOTA - São Paulo: pesquisa colaborativa: integrando as dimensões da biodiversidade microbiana ao longo de áreas de alteração do uso da terra em florestas tropicais (FAPESP 2014/50320-4)", requer uma interpretação quantitativa das interações entre às três dimensões (filogenética, genética e funcional) da diversidade microbiana. O objetivo geral desta tese foi aplicar uma abordagem computacional integrativa para investigar o papel ecológico de Archaea e Bacteria de solos amazônicos, usando tecnologias de sequenciamento massivo de DNA, bioinformática e reconstrução de genomas a partir de metagenomas (MAGs), para entender não apenas a diversidade ecológica, mas também a evolução, o metabolismo e a distribuição biogeográfica. No primeiro capítulo, apresentamos uma visão geral dos principais tópicos abordados nesta tese. Depois, discutimos as principais abordagens comumente usadas em análises de bioinformática e na integração de dados gerados em estudos de microbiomas. No terceiro capítulo, a aplicação de uma abordagem integrativa nos permitiu descobrir que os traços funcionais associados a nitrificação, em Archaea (e.g., oxidação da amônia), pode ser um caráter derivado, que emergiu depois da diversificação das Thaumarchaeota (Archaea). Além do mais, analisando mais de 27.000 amostras ambientais, que estão depositadas em bancos de dados públicos, descobrimos que o grupo de Thaumarchaeota não-oxidadoras de amônia tem uma especificidade por determinados habitats (e.g., Group 1.1c é mais específico de solos e sedimentos não-salinos). Neste estudo, descrevemos pela primeira vez um genoma de Thaumarchaeota de solos tropicais. Adicionalmente, no capítulo quatro os resultados indicaram que os membros do filo CPR/Patescibacteria (Bacteria), identificados em solos de pastagem da Amazônia, apresentam genomas pequenos. Estes resultados também expandem a diversidade de ambientes dentro da distribuição deste novo grupo bacteriano. O terceiro e quarto capítulo destacam a importância dos métodos de reconstrução de genomas a partir de dados metagenômicos para a expansão dos bancos de dados de genomas de referência. Por último, no Capítulo 5, exploramos o efeito da conversão de floresta-em-pastagem e o aumento dos níveis de umidade do solo na composição de Archaea em solos amazônicos. Nossos resultados indicam que as alterações causadas pelos maiores níveis de umidade do solo são mais pronunciadas em pastagens, onde as comunidades foram mais sensíveis, aumentando o potencial de metanogênese, enquanto a floresta pode atuar como "*buffer*" durante a estação chuvosa e abrigar comunidades mais estáveis. Esta tese destaca a importância do uso de ferramentas avançadas de bioinformática e de abordagens computacionais integradas para uma melhor compreensão dos processos evolutivos, vias metabólicas e distribuição ambiental de microrganismos de microbiomas complexos.

Palavras-chave: Microbioma. Ecologia Microbiana. Microbiologia de Solo. Bioinformática. Solos tropicais amazônicos.

## SUMMARY

# 1. INTRODUCTION

Microorganisms are widely distributed in all sorts of environments all over the planet and are associated with most biogeochemical cycles, influencing greenhouse gas emissions and nutrient cycling (OFFRE; SPANG; SCHLEPER, 2013). Soil, for example, is the most diverse and complex habitat on Earth and is dominated by Archaea, Bacteria, and Fungi (FIERER, 2017), generating a complex network with a variety of communication and cooperation strategies (DANIEL, 2005). At the global scale, understand the soil microbiome is important to develop new models for both C and N-cycling dynamics (PAJARES; BOHANNAN, 2015), enabling us to mitigate the greenhouse gas emissions (LAMMEL et al., 2015).

The Amazon rainforest is a great reservoir of biodiversity, hosting 25% of all known terrestrial animal and plant species and is responsible for regulating biogeochemical cycles (Wilson et al., 2016) with effects on the climate (MALHI et al., 2008). The microbial communities of Amazon soils are important to maintain the functional equilibrium in native forests (MENDES et al., 2015). In addition, the Amazon floodplain forests (e.g., varzea forests) also show a high diversity of microorganisms that are influencing in the environmental functions. For example, the methane produced by the organic matter degradation in the Amazon floodplain forests represents 5% of the total methane emission of the world (DEVOL et al., 1990). Further, previous studies demonstrated that land-use change (e.g. forest-to-pasture conversion) can alter the abundance, composition, and diversity of specific bacterial taxa detected in these soils, such as Acidobacteria (NAVARRETE et al., 2015), Verrucomicrobia (RANJAN et al., 2015), and some groups of Fungi (MUELLER et al., 2014).

Up to date, only a small number of studies have investigated the diversity of Archaea in Amazon soils (HAMAOUI et al., 2016; NAVARRETE et al., 2011; TUPINAMBÁ et al., 2016). With the improvement of culture-independent sequencing methods (SPANG et al., 2015), it has been possible to investigate the complex role of Archaea in nutrient cycling and linking that information with the genomic and metabolic content (EVANS et al., 2015). These findings have direct implications on our view of the tree of life, expanding our knowledge about Archaea diversity and evolution.

At present, Archaea are divided into five major clades: TACK (Protearchaeota), Asgard, Korarchaeota, Euryarchaeota e DPANN (SPANG; CACERES; ETTENA, 2017). The clade TACK presents Archaea associated with the methane and nitrogen cycle, such as Thaumarchaeota and Bathyarchaeota. The phylum Thaumarchaeota has been identified as a keystone group in aerobic ammonia-oxidizing processes (nitrification) (PESTER; SCHLEPER; WAGNER, 2011), and includes *Nitrosopumilus* and *Nitrosospharea* genera (PESTER et al., 2012). Regarding Thaumarchaeota functional features, studies performed in controlled conditions (WEBER et al., 2015) have reconstructed near-complete genomes from metagenomics dataset (culture-independent methods) (BEAM et al., 2014; LIN et al., 2015) and have demonstrated the potential to additional metabolic features, which are not associated with the ammonia oxidation. Dragon and Beowulf, the first two non-ammonia oxidizers Thaumarchaeota, were described by Beam et al. (2014). According to Beam and collaborators, these microorganisms are chemoorganotrophs and capable of performing the oxidation of sulfide to sulfate, or fermentation and litho/organotrophic oxygen or nitrate reduction (BEAM et al., 2014). The third non-ammonia oxidizer (Fn1 Thaumarchaeota) was described by Lin et al. (2015) and has the potential to degrade long-chain of fatty acids (LCFA) via β-oxidation. The description of new non-ammonia oxidizing Thaumarchaeota can open a new world to explore the functions of this high-diversity phylogenetic group.

Archaea are also associated with carbon cycling and methane production (methanogenesis). Classical studies regarding microbial metabolism demonstrated that only the Euryarchaeota phylum is linked with methanogenesis (GRIBALDO; BROCHIER-ARMANET, 2006). However, Evans and collaborators (2015), using genome-centric approaches, discovered a new phylum (Bathyarchaeota) with the potential capacity to produce methane. Since then, new genomes of Bathyarchaeota have been recovered and their genomic content related to ecological traits (DOMBROWSKI et al., 2017). The role of Bathyarchaeota in tropical soils is totally unknown.

On the other hand, some new Bacteria groups, such as the phyla Candidate Phyla Radiation (CPR)/Patescibacteria, was described recently and represented 15% of the total fraction of the Bacteria domain (BROWN et al., 2015; PARKS et al., 2018). To date, there is only one species cultivated from this taxonomic group (HE et al., 2014), and all other members remain uncultivated (CASTELLE et al., 2018). Although some species have the potential to carbon degradation (DANCZAK et al., 2017) and nitrogen metabolism (CASTELLE et al., 2018), the ecology and functional metabolism of CPR/Patescibacteria needs to be investigated. Furthermore, the CPR/Patescibacteria group can be an excellent

bacterial phylum to explore genomic traits associated with the genome size and adaptations in tropical soils. In fact, Konstantinidis and Tiedje (2004) suggested that the ecological triumph of the soil microorganisms are the large genome size, which could be explained by the availability of diverse but scarce of resources. The large genome size could be an adaptation to survive in a non-stable environment (DINI-ANDREOTE et al., 2012; KONSTANTINIDIS; TIEDJE, 2004), while more stable environments could select microorganisms with small genomes and less non-redundant functions (MORRIS; LENSKI; ZINSER, 2012), such as parasitic (*e.g. Mycoplasma*).

The use of new molecular biology techniques based on DNA/RNA sequencing has revolutionized the soil microbiome studies (ROESCH et al., 2007; WOODCROFT et al., 2018). Furthermore, with the development of new software and data analysis (EDGAR, 2013; WU et al., 2014) have been possible to unlock the soil black box and discovery new microbial taxa and functions associated them (KROEGER et al., 2018; WOODCROFT et al., 2018). These techniques are useful to link microbial diversity to the functioning of microbial communities and ecosystems (KRAUSE et al., 2014) or biodiversity patterns to biochemistry (NELSON; MARTINY; MARTINY, 2016). However, the large amount of data that are generated in these studies makes bioinformatics analysis one of the main challenges to extract biological information and testing hypothesis from microbiome datasets (LEMOS et al., 2017). Conceptually, the soil microbiome analysis can be performed in two individual and/or complementary ways: (I) metagenomic (DNA) and/or metatranscriptomic (RNA) analysis to study the functional traits and link soil microbial taxa to soil processes (FIERER et al., 2012), and (II) 16S rRNA (metaxonomics) to study the phylogenetic structure of the microbial community based on amplicon sequencing.

First of all, to quantify gene and categorical functions the availability of the reference gene catalogs is necessary, once the short metagenomic reads are mapped to the catalog to profile the taxa and gene content of each sample (QIN et al., 2010; SUNAGAWA, et al., 2015). The reference gene catalogs can be used to rapid and multi-omic profiling of the metagenomic samples (LI et al., 2014) and allow discovery of functional signatures (FORSLUND et al., 2015). However, the cultivability of soil microorganisms is difficult, once we do not have all information to simulate the perfect conditions for microbial growth. To solve this problem, it has been proposed the reconstruction of microbial genomes from metagenomic data, which is divided into three main steps (WOODCROFT et al., 2018; SORENSEN et al., 2019). The first step is (I) the sample collection and wet-lab protocols (*e.g.* DNA extraction and shotgun sequencing). After the sequencing of the short reads, they

are assembled into larger contigs (II), and to cluster the final contig datasets into individual populations using compositional proprieties (*e.g.* GC content and coverage), the genome binning is used (WU; SIMMONS; SINGER, 2016). To check the quality of each recovered individual genome, some metrics can be applied (*e.g.* completeness and contamination). Each individual genome is annotated using functional genome annotation that is dependent on the sequencing similarity to other known gene or protein to assess the potential function. Additionally, a manual curation can be used to improve the annotation. Lastly, the microbial genome can be submitted to a database (*e.g.* NCBI or IMG) to storage and public access (III).

The phylogenetic diversity of the soil microbial community can also be accessed by the use of amplicon sequencing (*e.g.* amplification and sequencing of the 16S rRNA gene) (ROESCH et al., 2007). An import assumption in this type of analysis is the concept of Operational Taxonomic Units (OTUs) (SOKAL, 1963), or Amplicon Sequencing Variant (ASV), which was recently proposed to replace OTUs in the microbial diversity studies (CALLAHAN et al., 2017). OTU or ASV are the biological units used to estimate the richness and diversity of a microbial community and the robustness of their identification depends on multiple bioinformatics steps, that include raw data filtering, chimera identification, and the removal of non-biological sequences (LEMOS et al., 2017). The application of the ASVs solves the problems generated by the clustering of sequencing reads (CALLAHAN; MCMURDIE; HOLMES, 2017). The main limitation of the use of OTUs is the arbitrary dissimilarity threshold to definition of OTUs (*e.g.* 95% or 97% of similarity), while the use of ASV can control, model, and correct Illumina sequencing errors and distinguish sequence variants at one nucleotide of difference (CALLAHAN et al., 2016). Some computational pipelines were published using this concept, but with differences in the way to correct the sequencing errors. For example, DADA2 generates a trained parametric error model to correct and collapse the sequence errors into ASV (CALLAHAN et al., 2016). On the other hand, UNOISE3 applies a one-pass clustering strategy with two parameters pre-defined by the author to generate ‘zero-radius ASV’ (EDGAR, 2016). One limitation of these computational strategies is that they are only applied to Illumina sequencing reads and are not recommend to Ion Torrent, PacBio or other sequencing technologies.

This thesis was based on the need to use an integrated computational approach to understand the ecology and evolution of Archaea and Bacterial in Amazon soils, using massive sequencing technologies, reconstruction of genomes from metagenomes and microbial phylogenetic marker genes (*e.g.* metataxonomics) to understand not only the diversity, but also the evolutionary history, potential metabolism and biogeography

information. Firstly, in Chapter 2, we outlined the main approaches applied for microbiome studies based on high-throughput sequencing technologies and we introduced the most commonly used strategies for bioinformatics analyses and data integration. Thus, in Chapter 3, the application of an integrative computational approach allowed us to discover that the nitrogen-related traits associated with nitrification in Archaea (ammonia oxidation) metabolism seems to be a derived feature and emerged late in the diversification of the Thaumarchaeota group. Furthermore, after analyzing more than 27,000 public environmental samples, we discovered that the non-ammonia oxidizing clade has habitat-specific subgroups (*e.g.*, Group 1.1c is more specific of soils and non-saline sediments). In Chapter 4, we discovered that the small-sized genome is a trait of the new CPR/Patescibacteria phyla in thawing permafrost and cattle-pasture soils, and we also expanded the range of environments within the radiation of this new bacterial group. Our data also highlight the importance of binning methods for the expansion of RefSoil database. Additionally, Chapter 5 was complementary to understand the forest-to-pasture conversion and the increase of soil moisture levels impacts on the Archaea community composition in amazon soils.

### 1.1. Hypothesis

The Study 1 (Chapter 3) presented in this thesis sought to test the hypothesis that the non-ammonia oxidizing Thaumarchaeota shows a heterotrophic metabolism and could be evolved from thermal to moderate temperature habitats (*e.g.* soils and sediments), and then originated the ammonia-oxidizing Thaumarchaeota. The Study 2 (Chapter 4) tested the hypothesis that complex microbiomes, such as those present in soil ecosystems, favor microorganisms with larger genomes and accessory genes, due their greater metabolic versatility, which allow them to survive and acclimate in a changing-environment with diverse but limited resources. In addition, the Study 3 (Chapter 5) tested the hypothesis that the forest-to-pasture conversion and the increase of soil moisture levels modify the Archaea community composition.

## 1.2. Objectives

### 1.2.1. General objective

The general objective of this thesis was to use an integrative computational approach to investigate the evolution, potential metabolic pathways, and biogeographical distribution of Archaea and CPR/Patescibacteria (Bacteria). The integrative computational approach was described with details on Chapter 3 and complemented on Chapter 4 and Chapter 5.

### 1.2.2. Specific objectives

To achieve the general objective of this thesis, the following specific objectives were considered:

I. To infer the evolutionary process, potential metabolism, and biogeographic distribution of the Thaumarchaeota (Archaea) group, by applying an integrative computational approach based on bioinformatics methods, such as metagenome assembly and binning, and the use of public databases (*e.g.* Earth Microbiome Project).

2. To explore the genome size features of soil microorganisms, by applying an integrated meta-analysis of Amazon soil metagenomes, and public available genomes and metagenomes.

3. To determine how the Archaea community composition responses to long-term (land-use change) and short-term (soil moisture level alterations) disturbance in Amazon soils.

## 1.3. Structure of the thesis

This thesis comprises five chapters and introduction and four chapters presented in scientific manuscript format written in English language. The supplementary materials indicated in each chapter are available in the Appendix section.

**References**

BEAM, J. P. et al. Niche specialization of novel Thaumarchaeota to oxic and hypoxic acidic geothermal springs of Yellowstone National Park. **The ISME Journal**, London, v. 8, n. 4, p. 938–951, 2014.

BROWN, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. **Nature**, London, v. 523, n. 7559, p. 208-211, 2015.

CALLAHAN, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. **Nature Methods**, London, v. 13, n. 7, p. 581–583, 2016.

CALLAHAN, B. J.; MCMURDIE, P. J.; HOLMES, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. **The ISME Journal**, London, v. 11, n. 12, p. 2639–2643, 2017.

CASTELLE, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. **Nature Reviews Microbiology**, London, v. 16, n. 10, p. 629–645, 2018.

DANCZAK, R. E. et al. Members of the Candidate Phyla Radiation are functionally differentiated by carbon- and nitrogen-cycling capabilities. **Microbiome**, London, v. 5, n. 1, p. 112, 2017.

DANIEL, R. The metagenomics of soil. **Nature Reviews Microbiology**, London, v. 3, n. 6, p. 470–478, 2005.

DEVOL, A. H. et al. Seasonal dynamics in methane emissions from the Amazon River floodplain to the troposphere. **Journal of Geophysical Research**, Washington, DC, v. 95, n. D10, p. 16417-16426, 1990.

DINI-ANDREOTE, F. et al. Bacterial genomes: habitat specificity and uncharted organisms. **Microbial Ecology**, Heidelberg, v. 64, n. 1, p. 1–7, 2012.

DOMBROWSKI, N. et al. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. **Microbiome**, London, v. 5, n. 1, p. 106, 23 2017.

EDGAR, R. UPARSE: highly accurate OTU sequences from microbial amplicon reads. **Nature Methods**, London, v. 10, p. 996-998, 2013.

EDGAR, R. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv, New York, 081257, 2016. doi: 10.1101/081257.

EVANS, P. N. et al. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. **Science,** New York, v. 350, n. 6259, p. 434–438, 2015.

FIERER, N. et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. **Proceedings of the National Academy of Science of the USA**, Washington, DC, v. 109, n. 52, p. 21390–21395, 2012.

FIERER, N. Embracing the unknown: disentangling the complexities of the soil microbiome. **Nature Reviews. Microbiology**, London, v. 15, n. 10, p. 579–590, 2017.

FORSLUND, K. et al. Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. **Nature**, London, v. 528, n. 7581, p. 262–266, 2015.

GRIBALDO, S.; BROCHIER-ARMANET, C. The origin and evolution of Archaea: a state of the art. **Philosophical Transactions of the Royal Society B: Biological Sciences**, London, v. 361, n. 1470, p. 1007-1022.

HAMAOUI, G. S. et al. Land-use change drives abundance and community structure alterations of thaumarchaeal ammonia oxidizers in tropical rainforest soils in Rondônia, Brazil. **Applied Soil Ecology**, Amsterdam, v. 107, p. 48–56, 2016.

HE, Y. et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. **Microbiome**, London, v. 3, p. 20, 2015. doi: 10.1186/s40168-015-0081-x.

JANSSON, J. K.; HOFMOCKEL, K. S. The soil microbiome—from metagenomics to metaphenomics. **Current Opinion in Microbiology**, London, v. 43, p. 162–168, 2018a.

KONSTANTINIDIS, K. T.; TIEDJE, J. M. Towards a Genome-Based Taxonomy for Prokaryotes. **Journal of Bacteriology**, Washington, DC, v. 187, n. 18, p. 6258–6264, 2005.

KRAUSE, S. et al. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. **Frontiers in Microbiology**, Lausanne, v. 5, p. 251, 2014. doi: 10.3389/fmicb.2014.00251.

KROEGER, M. E. et al. New biological insights into how deforestation in Amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. **Frontiers in Microbiology**, Lausanne, v. 9, p. 1635, 2018. doi: 10.3389/fmicb.2018.01635.

LAMMEL, D. R. et al. Specific microbial gene abundances and soil parameters contribute to C, N, and greenhouse gas process rates after land use change in Southern Amazonian soils. **Frontiers in Microbiology**, Lausanne, v. 6, p. 1057, 2015. doi: 10.3389/fmicb.2015.01057.

LEMOS, L. N. et al. Bioinformatics for microbiome research: concepts, strategies, and advances. In: PYLRO, V.; ROESCH, L. **The Brazilian microbiome**. New York: Springer, 2017. p. 111-123.

LI, J. et al. An integrated catalog of reference genes in the human gut microbiome. **Nature Biotechnology**, London, v. 32, n. 8, p. 834–841, 2014.

LIN, X. et al. Metabolic potential of fatty acid oxidation and anaerobic respiration by abundant members of Thaumarchaeota and Thermoplasmata in deep anoxic peat. **The ISME Journal**, London, v. 9, n. 12, p. 2740–2744, 2015.

MALHI, Y. et al. Climate change, deforestation, and the fate of the Amazon. **Science**, New York, v. 319, n. 5860, p. 169–172, 2008.

MENDES, L. W. et al. Soil-borne microbiome: linking diversity to function. **Microbial Ecology**, Amsterdam, v. 70, n. 1, p. 255–265, 2015.

MORRIS, J. J.; LENSKI, R. E.; ZINSER, E. R. The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. **MBio**, Washington, DC, v. 3, n. 2, p. e00036-12, 2012.

MUELLER, P. et al. Global-change effects on early-stage decomposition processes in tidal wetlands – implications from a global survey using standardized litter. **Biogeosciences**, Göttingen, v. 15, n. 10, p. 3189–3202, 2018.

NAVARRETE, A. A. et al. Land-use systems affect Archaeal community structure and functional diversity in western Amazon soils. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 35, n. 5, p. 1527–1540, 2011.

NAVARRETE, A. A. et al. Differential response of Acidobacteria subgroups to forest-to-pasture conversion and their biogeographic patterns in the Western Brazilian Amazon. **Frontiers in Microbiology**, Lausanne, v. 6, p. 1443, 2015. doi: 10.3389/fmicb.2015.01443.

NELSON, M. B.; MARTINY, A. C.; MARTINY, J. B. H. Global biogeography of microbial nitrogen-cycling traits in soil. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 113, n. 29, p. 8033–8040, 2016.

OFFRE, P.; SPANG, A.; SCHLEPER, C. Archaea in biogeochemical cycles. **Annual Review of Microbiology**, Palo Alto, v. 67, p. 437–457, 2013.

PAJARES, S.; BOHANNAN, B. J. M. Ecology of nitrogen fixing, nitrifying, and denitrifying microorganisms in Tropical Forest soils. **Frontiers in Microbiology**, Lausanne, v. 7, p. 1045, 2016. doi: 10.3389/fmicb.2016.01045.

PARKS, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. **Nature Biotechnology**, London, v. 36, n. 10, p. 996–1004, 2018.

PESTER, M. et al. amoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions. **Environmental Microbiology**, Oxford, v. 14, n. 2, p. 525–539, 2012.

PESTER, M.; SCHLEPER, C.; WAGNER, M. The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. **Current Opinion in Microbiology**, London, v. 14, n. 3, p. 300–306, 2011a.

QIN, J. et al. A human gut microbial gene catalogue established by metagenomic sequencing. **Nature**, London, v. 464, n. 7285, p. 59–65, 2010.

RANJAN, K. et al. Forest-to-pasture conversion increases the diversity of the phylum Verrucomicrobia in Amazon rainforest soils. **Frontiers in Microbiology**, Lausanne, v. 6, p. 779, 2015. doi: 10.3389/fmicb.2015.00779.

ROESCH, L. F. W. et al. Pyrosequencing enumerates and contrasts soil microbial diversity. **The ISME Journal**, London, v. 1, n. 4, p. 283–290, 2007.

SOKAL, R. R. The Principles and Practice of Numerical Taxonomy. **Taxon**, New York, v. 12, n. 5, p. 190–199, 1963.

SORENSEN, J. W. et al. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. **Nature Microbiology**, London, v. 4, n. 1, p. 55–61, 2019.

SPANG, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. **Nature**, London, v. 521, n. 7551, p. 173–179, 2015.

SPANG, A.; CACERES, E. F.; ETTEMA, T. J. G. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. **Science**, New York, v. 357, n. 6351, p. eaaf3883, 2017.

SUNAGAWA, S. et al. Ocean plankton. Structure and function of the global ocean microbiome. **Science,** New York, v. 348, n. 6237, p. 1261359, 2015.

TUPINAMBÁ, D. D. et al. Archaeal Community Changes Associated with Cultivation of Amazon Forest Soil with Oil Palm. **Archaea,** Vancouver, v. 2016, p. 3762159, 2016.

WEBER, E. B. et al. Ammonia oxidation is not required for growth of Group 1.1c soil Thaumarchaeota. **FEMS Microbiology Ecology**, Amsterdam, v. 91, n. 3, 2015. doi: 10.1093/femsec/fiv001.

WILSON, C. et al. Contribution of regional sources to atmospheric methane over the Amazon Basin in 2010 and 2011. **Global Biogeochemical Cycles**, Singapore, v. 30, n. 3, p. 400–420, 2016.

WOODCROFT, B. J. et al. Genome-centric view of carbon processing in thawing permafrost. **Nature**, London, v. 560, n. 7716, p. 49–54, 2018.

WU, Y. W.; SIMMONS, B. A.; SINGER, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. **Bioinformatics**, Oxford, v. 32, n. 4, p. 605-607, 2016.

# 2. BIOINFORMATICS FOR MICROBIOME RESEARCH: CONCEPTS, STRATEGIES, AND ADVANCES[1]


## ABSTRACT

Advances in next-generation sequencing technologies allow comparative analyses of the diversity and abundance of whole microbial communities, and of important ecosystem functional genes, at far greater depths than ever before. However, the current major challenge for the use of this immense amount of genetic information is undoubtedly how to convert the information into rational biological conclusions. As an attempt to solve this issue, we now rely on a set of complex computational/statistical analyses, the use of which, however, could be a drawback for most researchers in the biological sciences. In this chapter, we outline the main approaches applied for microbiome studies based on high-throughput sequencing technologies and we introduce the most commonly used strategies for data handling, sequence clustering, taxonomic and functional assignment, and microbial community comparisons.


Keywords: Amplicon sequencing; metagenomics; computational methods; softwares.

## 2.1. Introduction

Current scientific and technological advances have revolutionized the way that we usually studied microbiological resources (CARDENAS; TIEDJE, 2008). Since the introduction of next-generation sequencing (NGS) about 15 years ago, scientists have generated an unprecedented amount of genomic information, which has been cataloged in multiple biological databases (CHEN et al., 2010; QUAST et al., 2013; COLE et al., 2014; KEEGAN et al., 2016; PAEZ-ESPINO et al., 2017). However, these improvements in DNA sequencing methodologies arrived before we had the ability to comprehensively analyze the

[1] This chapter was published as part of a book chapter on September 1, 2017 in the "*The Brazilian Microbiome: Current Status and Perspectives*" book.
LEMOS, L. N. et al. Bioinformatics for microbiome research: concepts, strategies, and advances. In: PYLRO, V.; ROESCH, L. (Eds.) **The Brazilian microbiome:** current status and perspectives. Cham: Springer International Publishing, 2017. p. 111–123.

huge amount of data that was generated, and this makes bioinformatics one of the main bottlenecks in microbiome studies.

Studies that gather genomic information from single microbial populations, or even single-cell genomic studies, are useful for separating closely related strains, finding small genomic changes by comparative genomics, and disentangling the “microbial dark matter” as well (see more in (RINKE et al., 2013)). These kinds of studies rely heavily on genomic annotation, which reveals information regarding a microbe’s complete metabolic potential, indicating what makes this organism different from others. Therefore, precise annotation of the genome and standardization of the nomenclature of each identified gene (the term “high-quality annotation” is used in the literature) is of fundamental importance. Comparisons between genomes may provide evidence of the biological processes involved in differentiation and genomic evolution, as well as revealing important aspects of the genotype and phenotype relationship.

Besides the strategies used for analyzing single populations or cells, there are also other approaches focused on profiling entire microbial communities. With the possibility of obtaining millions (or billions) of microbial sequences from complex samples (e.g., environmental and host-associated samples), these approaches are now widely used by researchers. The computational analysis of these big datasets is now allowing us to reveal the microbial taxonomic structure in each sample - through data analyses of microbial phylogenetic marker genes, e.g., rRNA 16S (metataxonomics)-and their potential functional traits, by shotgun metagenomic (DNA) and/or metatranscriptomic (RNA) analyses. In fact, the generation of data for the target sequencing of phylogenetic markers, metagenomics, and metatranscriptomics is now reasonably well established and several DNA sequencing platforms based on different technologies are currently available (GOODWIN et al., 2016). However, considerable computational effort is required for the processing of NGS sequencing data and this sudden reliance on computing has been problematic for most researchers in the biological sciences. Without programming skills or expertise in computer science, researchers who rely on computational approaches are troubled by issues such as software installation and efficient software combinations, the determination of parameters, and the manipulation of large data files. Thus, to enable the systematic processing of large volumes of sequence data, including the structured storage of sampled data and metadata and the standardization of data analyses, there are fundamental requirements both for computers with a scalable structure and for well-trained bioinformaticians.

In this chapter, we intend to broaden readers' views of the main bioinformatics strategies for studying microbes using high-throughput sequencing technologies. This outline includes the most commonly used approaches for data handling, sequence clustering, taxonomic and functional assignment, and microbial community comparison.

## 2.2. Strategies

### *2.2.1. Functional Profile Based on Metagenomic and Metatranscriptomic Data: 'What can/do they do?'*

#### Gene Prediction and Functional Gene Annotation

Finding the encoding genes in metagenomic DNA sequences is the first step in predicting protein function. This is a big challenge in bioinformatics, because the prediction needs to be performed on short fragmented reads (incomplete genes). Many softwares, such as Ophelia (HOFF et al., 2009), FragGeneScan (RHO et al., 2010), MetaGeneMark (ZHU et al., 2010), and Glimmer-MG (KELLEY et al., 2012) have been developed to annotate short metagenomic reads. For example, Ophelia (HOFF et al., 2008) uses fragment length-specific models for gene prediction, while FragGeneScan (RHO et al., 2010) also combines sequencing error information and codon usage in a probabilistic model. This information improves accuracy in the prediction of coding sequences.

Methods based only on homology, such as BLASTx (ALTSCHUL et al., 1990) and DIAMOND (BUCHFINK; XIE; HUSON, 2015), do not use ab initio gene prediction. Homology-based methods allow searching for similar sequences in protein databases e.g., non-redundant database (nr/National Center for Biology Information [NCBI]). The similarity search is slower than the direct comparison of ab initio predicted sequences, because the sequences must be translated into the six reading frames. Currently, DIAMOND is an alternative for annotating metagenomic reads, because of its speed in annotating millions of sequences in a short time (BUCHFINK; XIE; HUSON, 2015).

**Assigning Taxonomy**

Assigning the taxonomy of short metagenomic reads may be done in two ways: (i) by approaches based on comparative analysis with all genome regions, including conserved housekeeping genes and highly variable genes; or (ii) by approaches based only on similarity to conserved housekeeping genes. The first option is used in most softwares, including homology-based methods such as MEGAN (HUSON et al., 2017). MEGAN uses an output BLAST score (best hits) search for taxonomic prediction from the lowest common ancestor. In this case, all metagenomic reads are aligned against a protein database of all microbial genomes deposited in the NCBI, for example. The limitation of this algorithm is the low speed of the BLAST search, which uses millions of reads. However, other softwares have been developed to align metagenomic reads against databases; for instance, DIAMOND (BUCKFINK; XIE; HUSON, 2015), Kraken (WOOD; SALZBERG, 2014), and Centrifuge (KIM et al., 2016). There are also taxonomic prediction methods based on comparisons of each read against a clade-specific gene marker catalog, such as that in MetaPhlAn (SEGATA et al., 2012).

**Genome Assembly from Metagenomic Data**

Currently, metagenomes are analyzed by two main approaches: gene-centric and genome-centric. Gene-centric approaches are based on unassembled individual genomes and individual genes are predicted from short fragmented reads (PROSSER, 2015; BRULC et al., 2009). On the other hand, genome-centric approaches consider individual microbial populations reconstructed by total metagenome assembly (ALBERTSEN et al., 2013; LEMOS et al., 2017).

Strategies based on gene-centric analysis are limited by the length of short metagenomic reads. Although specific software exists for gene prediction based on short sequences, assembling short reads into contiguous sequences (contigs) is more powerful. Currently, softwares such as MetaVelvet (NAMIKI et al., 2012) and metaSPAdes (BANKEVICH et al., 2012) subdivide short reads in graphs per k-mer lengths (*De Brujin* graphs). There are several methods for assembling short reads; however, here we focus only on *De Brujin* methods, because they are the most commonly used metagenomic assemblers. MetaVelvet divides the graph into sub-graphs and each sub-graph represents an individual genome (NAMIKI et al., 2012).

Post-assembly analysis may enable improved gene prediction and functional annotation. Because contigs are longer than the usual short reads, they can be used for the reconstruction of near, partial, or complete microbial genomes of uncultivable bacteria (ALBERTSEN et al., 2013). This approach is known as “binning”. The main idea of binning is the clustering of assembled contigs into individual populations according to the compositional content of sequences, such as guanine-cytosine (GC) content, tetra-nucleotide frequency, and sequence coverage (WU et al., 2014). Some softwares available for binning and reconstructing individual microbial genomes from metagenomic data are MaxBin (WU et al., 2014), GroopM (IMELFORT et al., 2014), and MetaBAT (KANG et al., 2015).

### 2.2.2. *Taxonomic profile based on 16S Amplicon Data: ‘Who is there?’*

#### Picking Operational Taxonomic Units

Taxonomic identification is an important step in microbial community analyses. The robustness of these analyses depends on a series of initial processing steps, including raw data filtering, chimera identification, and the removal of spurious non-biological sequences (SCHLOSS; WESTCOTT, 2011). An important concept used in microbial community analysis is the grouping of sequences into operational taxonomic units (OTUs). This concept was applied for the first time in botanical research by Sokal (1963), but with the advances in molecular methods, this concept began to be used by microbiologists (MCCAIG; GLOVER; PROSSER, 1999). Multiple DNA sequences are clustered into an individual OTU by an arbitrary level of sequence identity (for example, 97% identity roughly representing genus and 95% identity representing family) (SCHLOSS; HANDELSMAN, 2005). The great advantage of grouping sequences into OTUs is the reduction of computational needs, once the number of sequences is reduced by picking a representative sequence from a pool of sequences in an OTU. Although this concept is widely applied and accepted by the scientific community, its application is questionable, because the similarity cutoffs applied to partial 16S gene sequences have no biological meaning and different biological entities present different identity levels. However, the lack of a better approach to deal with this issue justify its current use.

The strategy of picking OTUs has been applied since the beginning of microbial community analysis and may be used with three different options for OTU picking: closed reference-based (BLAST (ALTSCHUL et al., 1990), UCLUST (EDGAR, 2010), USEARCH (EDGAR, 2010)), open-reference-based (UCLUST, USEARCH), and de novo (CD-HIT (Fu et al., 2012), Mothur (SCHLOSS et al., 2009), prefix/suffix, trie, UCLUST, USEARCH) (NAVAS-MOLINA et al., 2013). The closed-reference strategy is based on comparative identity between amplicon sequences and a reference database (e.g., Greengenes (DESANTIS et al., 2006). The open-reference strategy is also based on alignment against a reference database; however, sequences that do not cluster with the reference are subsequently clustered by the de novo approach. The de novo approach is used for clustering amplicon sequences by pairwise comparison, without the need for a reference database. These algorithms are implemented in different softwares and they have been evaluated by numerous benchmarking studies (SCHLOSS; WESTCOTT, 2011; WESTCOTT; SCHLOSS, 2015; BONDER et al., 2012). The softwares most widely used to cluster biological sequences are UCLUST (EDGAR, 2010) (which is applied as a default method in the QIIME pipeline for all OTU picking approaches), Mothur (SCHLOSS et al., 2009) (picking OTUs by a de novo approach, based only on genetic distance methods), and UPARSE (EDGAR, 2013) (which uses USEARCH to pick OTUs by a de novo approach). However, none of these softwares or algorithms is free of bias, so the researcher must evaluate which algorithm or software is best for their dataset. For example, the QIIME pipeline keeps a large fraction of chimeric OTUs, inflating microbial diversity estimates (EDGAR, 2013). On the other hand, UPARSE (EDGAR, 2013) might discard true OTUs because of its highly stringent default filtering parameters, thus making a false-negative type of error (KOPYLOVA et al., 2016). Genetic-distance methods implemented in Mothur, such as the average neighbor algorithm, seem to be the most robust approach (SCHLOSS, 2016), but these methods require great computational power, which might prevent the analysis of very large datasets in ordinary desktop computers. A common problem of open-reference strategies is the creation of unstable OTUs, where the cluster that a sequence is assigned to is affected by the number of sequences in the dataset (He et al., 2015). Close-reference approaches generate stable OTUs; however, a considerable disadvantage of such approaches is the unavailability of complete public datasets if the approach excludes any OTUs that are not defined in a pre-existing reference dataset. The choice of the best algorithm to use depends on the biological and ecological question and the throughput of data.

## Assigning Taxonomy

Several methods have been developed aiming to predict microbial taxonomy based on partial sequences of the 16S rRNA gene. The most widely used is the naïve Bayesian classifier implemented in the Ribosomal Database Project (RDP) (WANG et al., 2007). With this method, sequences of 400 bp in length can be classified at genus level, and the method also uses bootstrap confidence scores to support the taxonomic assignment (WANG et al., 2007). Other methods available are implemented in QIIME (CAPORASO et al., 2010) and Mothur (SCHLOSS et al., 2009). QIIME default classification uses only similarity among sequences to infer taxonomy (KUCZYNSKI et al., 2012). Mothur uses k-mer counting and the Wang naïve Bayesian classifier, similarly to the RDP method (SCHLOSS et al., 2009).

Few studies have been conducted to compare the performance of the taxonomic prediction algorithms used in microbial diversity studies. Bokulich and colleagues (BOKULICH et al., 2015) have demonstrated that the RDP classifier and Mothur provide the same results for taxonomy prediction, although the RDP classifier has the advantage of discovering novel taxa. The RDP (COLE et al., 2009), Greengenes (DESANTIS et al., 2006), and SILVA (PRUESSE et al., 2007) are the main databases used for taxonomy assignment. The RDP database covers 27 phyla (RDP Release 11), including those that are uncultivable (e.g., Bathyarchaeota archaea). Greengenes had its last update in 2013, with the implementation of the tax2-tree tool to transfer taxonomy to a phylogenetic tree (MCDONALD et al., 2012), but this database does not contain any new recently described phyla (HUG et al., 2016). SILVA is the most complete database, covering all phyla in its last update (RELEASE 132).

## Measuring Alpha and Beta Diversity

Several tools are available to measure the alpha and beta diversity of an ecological community. These include statistical packages (e.g., Vegan (OKSANEN et al., 2016)) that are implemented in general pipelines, such as QIIME and Mothur. Alpha diversity is the local diversity of a single sample and beta diversity is the diversity among different samples (LEMOS et al., 2011). Specific methods are available for determining each type of diversity (alpha or beta). Alpha diversity indexes, such as the Shannon diversity index (SHANNON, 1948) and the Simpson diversity index (SIMPSON, 1949), measure the species richness and evenness of the community structure. On the other hand, beta diversity indexes are applied for

direct comparisons of the abundance profile or presence/absence of OTUs using distance metrics, either by counting methods (e.g., Bray-Curtis (BRAY; CURTIS, 1957)) or by phylogenetic reconstruction methods (e.g., UniFrac (LUZOPONE; KNIGHT, 2005)). The advantage of using phylogenetic approaches for comparisons of microbial communities is the possibility of using low sequence coverage. However, the use of methods based on absolute counting needs high sequence coverage to improve accuracy (LEMOS et al., 2011).

**References**

ALBERTSEN, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. **Nature Biotechnology**, London, v. 31, n. 6, p. 533–538, 2013.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, Amsterdam, v. 215, n. 3, p. 403–410, 1990.

BANKEVICH, A. et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. **Journal of Computational Biology**, Nova Rochelle, v. 19, n. 5, p. 455–477, 2012.

BOKULICH, N. A. et al. A standardized, extensible framework for optimizing classification improves marker-gene taxonomic assignments. **PeerJ PrePrints**, San Diego, v. 2, e934, 2015. doi: 10.7287/peerj.preprints.934v2.

BONDER, M. J. et al. Comparing clustering and pre-processing in taxonomy analysis. **Bioinformatics**, Oxford, v. 28, n. 22, p. 2891–2897, 2012.

BRAY, J. R.; CURTIS, J. T. An Ordination of the Upland Forest Communities of Southern Wisconsin. **Ecological Monographs**, Washington, DC, v. 27, n. 4, p. 325–349, 1957.

BRULC, J. M. et al. Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 106, n. 6, p. 1948–1953, 2009.

BUCHFINK, B.; XIE, C.; HUSON, D. H. Fast and sensitive protein alignment using DIAMOND. **Nature Methods**, London, v. 12, n. 1, p. 59–60, 2015.

CARDENAS, E.; TIEDJE, J. M. New tools for discovering and characterizing microbial diversity. **Current Opinion in Biotechnology**, London, v. 19, p. 544-549, 2008.

CAPORASO, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. **Nature Methods**, London, v. 7, n. 5, p. 335–336, 2010.

CHEN, T. et al. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. **Database: The Journal of**

**Biological Databases and Curation**, Oxford, v. 2010, baq013, 2010. doi: 10.1093/database/baq013.

COLE, J. R. et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. **Nucleic Acids Research**, Oxford, v. 42, p. D633-642, 2014. Database issue.

DESANTIS, T. Z. et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. **Applied and Environmental Microbiology**, Washington, DC, v. 72, n. 7, p. 5069–5072, 2006.

EDGAR, R. C. Search and clustering orders of magnitude faster than BLAST. **Bioinformatics**, Oxford, v. 26, n. 19, p. 2460–2461, 2010.

EDGAR, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. **Nature Methods**, London, v. 10, n. 10, p. 996–998, 2013.

FU, L. et al. CD-HIT: accelerated for clustering the next-generation sequencing data. **Bioinformatics**, London, v. 28, n. 23, p. 3150–3152, 2012.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, London, v. 17, n. 6, p. 333–351, 2016.

HE, Y. et al. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. **Microbiome**, London, v. 3, p. 20, 2015. doi: 10.1186/s40168-015-0081-x.

HOFF, K. J. et al. Orphelia: predicting genes in metagenomic sequencing reads. **Nucleic Acids Research**, Oxford, v. 37, p. W101-105, 2009. Web Server issue. doi: 10.1093/nar/gkp327.

HUG, L. A. et al. A new view of the tree of life. **Nature Microbiology**, London, v. 1, n. 5, p. 16048, 2016.

HUSON, D. H. et al. MEGAN analysis of metagenomic data. **Genome Research**, New York, v. 17, n. 3, p. 377–386, 2007.

IMELFORT, M. et al. GroopM: An automated tool for the recovery of population genomes from related metagenomes. **PeerJ PrePrints**, San Diego, v. 1, e409, 2014. doi: 10.7287/peerj.preprints.409v1.

KANG, D. D. et al. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. **PeerJ**, San Diego, v. 3, e1165, 2015. doi: 10.7717/peerj.1165.

KEEGAN, K. P.; GLASS, E. M.; MEYER, F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. **Methods in Molecular Biology,** Clifton, v. 1399, p. 207–233, 2016.

KELLEY, D. R. et al. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. **Nucleic Acids Research**, Oxford, v. 40, n. 1, e.9, 2012.

KIM, D. et al. Centrifuge: rapid and sensitive classification of metagenomic sequences. **Genome Research**, New York, v. 26, n. 12, p. 1721–1729, 2016.

KOPYLOVA, E. et al. Open-Source Sequence Clustering Methods Improve the State of the Art. **mSystems**, Washington, DC, v. 1, n. 1, e00003-15, 2016.

KUCZYNSKI, J. et al. Experimental and analytical tools for studying the human microbiome. **Nature Reviews Genetics**, London, v. 13, n. 1, p. 47–58, 2012.

LAN, Y. et al. Using the RDP Classifier to Predict Taxonomic Novelty and Reduce the Search Space for Finding Novel Organisms. **PLoS One**, San Francisco, v. 7, n. 3, e32491, 2012.

LEMOS, L. N. et al. Rethinking microbial diversity analysis in the high throughput sequencing era. **Journal of Microbiological Methods**, Amsterdam, v. 86, n. 1, p. 42–51, 2011.

LEMOS, L. N. et al. Genome-Centric Analysis of a Thermophilic and Cellulolytic Bacterial Consortium Derived from Composting. **Frontiers in Microbiology**, Lausanne, v. 8, p. 644, 2017.

LOZUPONE, C.; KNIGHT, R. UniFrac: A New Phylogenetic Method for Comparing Microbial Communities. **Applied and Environmental Microbiology**, Washington, DC, v. 71, n. 12, p. 8228–8235, 2005.

MCCAIG, A. E.; GLOVER, L. A.; PROSSER, J. I. Molecular analysis of bacterial community structure and diversity in unimproved and improved upland grass pastures. **Applied and Environmental Microbiology**, Washington, DC, v. 65, n. 4, p. 1721–1730, 1999.

MCDONALD, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. **The ISME Journal**, London, v. 6, n. 3, p. 610–618, 2012.

NAMIKI, T. et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. **Nucleic Acids Research**, Oxford, v. 40, n. 20, p. e155, 2012.

NAVAS-MOLINA, J. A. et al. Advancing our understanding of the human microbiome using QIIME. **Methods in Enzymology**, Amsterdam, v. 531, p. 371–444, 2013.

OKSANEN, J. et al. Vegan: community ecology package. R package 2.3-3. New York: R Group, 2016. Available from: https://cran.r-project.org/package=vegan.

PAEZ-ESPINO, D. et al. IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses. **Nucleic Acids Research**, Oxford, v. 45, n. D1, p. D457–D465, 2017.

PROSSER, J. I. Dispersing misconceptions and identifying opportunities for the use of “omics” in soil microbial ecology. **Nature Reviews Microbiology**, London, v. 13, n. 7, p. 439–446, 2015.

PRUESSE, E. et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. **Nucleic Acids Research**, London, v. 35, n. 21, p. 7188–7196, 2007.

QUAST, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic Acids Research**, London, v. 41, p. D590–D596, 2013. Database issue.

RHO, M.; TANG, H.; YE, Y. FragGeneScan: predicting genes in short and error-prone reads. **Nucleic Acids Research**, Oxford, v. 38, n. 20, p. e191, 2010.

RINKE, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. **Nature**, London, v. 499, n. 7459, p. 431–437, 2013.

SCHLOSS, P. D. et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. **Applied and Environmental Microbiology**, Washington, DC, v. 75, n. 23, p. 7537–7541, 2009.

SCHLOSS, P. D. Application of a Database-Independent Approach to Assess the Quality of Operational Taxonomic Unit Picking Methods. **mSystems**, Washington, DC, v. 1, n. 2, e00027-16, 2016.

SCHLOSS, P. D.; HANDELSMAN, J. Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. **Applied and Environmental Microbiology**, Washington, DC, v. 71, n. 3, p. 1501–1506, 2005.

SCHLOSS, P. D.; WESTCOTT, S. L. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. **Applied and Environmental Microbiology**, Washington, DC, v. 77, n. 10, p. 3219–3226, 2011.

SEGATA, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. **Nature Methods**, London, v. 9, n. 8, p. 811–814, 2012.

SHANNON, C. E. A Mathematical Theory of Communication. **Bell System Technical Journal**, New York, v. 27, n. 3, p. 379–423, 1948.

SIMPSON, E. H. Measurement of Diversity. **Nature**, London, v. 163, n. 4148, p. 688–688, 1949.

SOKAL, R. R. The Principles and Practice of Numerical Taxonomy. **Taxon**, New York, v. 12, n. 5, p. 190–199, 1963.

WANG, Q. et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. **Applied and Environmental Microbiology**, Washington, DC, v. 73, n. 16, p. 5261-5267, 2007.

WESTCOTT, S. L.; SCHLOSS, P. D. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. **PeerJ**, San Diego, v. 3, p. e1487, 2015. doi: 10.7717/peerj.1487.

WOOD, D. E.; SALZBERG, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. **Genome Biology**, London, v. 15, n. 3, p. R46, 2014.

WU, Y.-W. et al. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. **Microbiome**, London, v. 2, n. 1, p. 26, 2014.

ZHU, W.; LOMSADZE, A.; BORODOVSKY, M. Ab initio gene identification in metagenomic sequences. **Nucleic Acids Research,** London, v. 38, n. 12, p. e132, 2010.

# 3. NEW INSIGHTS ON THE EVOLUTION, POTENTIAL METABOLISM AND DISTRIBUTION OF THE NON-AMMONIA OXIDIZING THAUMARCHAEOTA


## ABSTRACT

The phylum Thaumarchaeota is an important Archaea group and almost all cultivated species of the class Nitrososphaeria have the capacity to oxidize ammonia. However, the use of experimental soil microcosm, the reconstruction of near-complete genomes from metagenomic data, and more recently, the cultivation of the *Conexivisphaera calidus* thaumarchaea, have been demonstrated that they also have the potential for fermentation, iron- and sulfur-reduction metabolism, which are not directly associated with the ammonia oxidation. Here, we described two new non-ammonia oxidizing Thaumarchaeota genomes from Amazon floodplain forest and partially thawed bog sediment. We used an integrated approach to investigate and complement the discussion about the evolution, biogeography and functional metabolism of non-ammonia oxidizing Thaumarchaeota, focusing in the species from soils and sediments. The evolution and potential metabolism were predicted using phylogenomic approaches and functional genome annotation, respectively. While the geographical distribution was analyzed by the public data deposited in the Earth Microbiome Project. Our results suggest that the non-ammonia and ammonia-oxidizing Thaumarchaeota might have emerged from a thermophilic environment, and the mesophilic lifestyle is a derived character. The ammonia oxidation metabolism also seems to be a derived character, appearing later in the diversification of Thaumarchaeota, reinforcing the recent findings. The comparative functional genome annotation indicated a potential to a heterotrophic lifestyle with the capability of acetate fermentation, and organic carbon and nitrogen degradation. Based on the analyses of more than 27,000 public environmental samples, we discovered that this clade has subgroups that may have habitat-specific distribution (*e.g.*, Group 1.1c is more specific of soil and non-saline sediments). Our results expand previous studies, describing the ecophysiology of terrestrial non-ammonia oxidizing thaumarchaea and their potential role in the soil and sediments, as such Amazon floodplain forest and partially thawed permafrosted, and open new questions about the role of Thaumarchaeota in environmental samples.


Keywords: Group 1.1c; Nitrification; Metagenome-assembled genomes (MAGs); Archaea; Sediments

## 3.1. Introduction

The phylum Thaumarchaeota is an important Archaea group and its evolution, biogeography and functional role in the nitrogen cycle (e.g., ammonia oxidation) has been investigated over the last 20 years (LEININGER et al., 2006; KÖNNEKE et al., 2005; TOURNA et al., 2011; KEROU et al., 2016; ALVES et al., 2018). To date (July 2019), there are 117 Thaumarchaeota genomes deposited in the NCBI database, which were assembled using cultivated-dependent methods (e.g., *Nitrososphaera viennesis*) (TOURNA et al., 2011), enrichment cultures (Candidatus *Nitrososphaera everglandensis*) (ZHALNINA et al., 2014), single-amplified genomes (SAGs) (RINKE et al., 2013) and using integrated strategies: cultivation and metagenome-assembled genomes (MAGs) (Candidatus *Nitrosocaldus cavascurensis*) (ABBY et al., 2018).

Although the evolution of ammonia-oxidizing Thaumarchaeota has already been well studied (ALVES et al., 2018), some discrete Thaumarchaeota classes such as Group 1.1c and Marine Benthic Group B, have not still been thoroughly investigated considering their evolutionary history, functional metabolism and biogeographical distribution. The discovered of Dragon and Beowulf thaumarchaeota genomes of hot-springs (BEAM et al., 2014), which do not have the ability to oxidize ammonia, opened new possibilities to explore the origin and evolution of Thaumarchaeota phylum. Further, the cultivation of the first non-ammonia oxidizing (non-AOA) *Conexivisphaera calidus* from a terrestrial acidic hot spring (KATO et al., 2019), showed more evidences to validate the hypothesis that the origin of Thaumarchaeota could be evolved from thermal to moderate temperature habitats, and then originated the ammonia-oxidizing Thaumarchaeota (BROCHIER-ARMANET; GRIBALDO; FROTERRE, 2015; HUA et al., 2018).

The biogeographical distribution of these discrete Thaumarchaeota groups (Group 1.1c, Marine Benthic, and others.) also has been less explored than the ammonia oxidizers Thaumarchaeota. Some studies indicated that the Group 1.1c represents 29% of all 16S rRNA sequences in temperate acid forest soils (KEMNITZ; KOLB; CONRAD, 2007) and in acid forest peat soil (STOPNIŠEK et al., 2010). However, the Thaumarchaeota global distribution has not been reported or explored.

Regarding the Thaumarchaeota functional features, Dragon and Beowulf thaumarchaea have the potential for chemoorganotrophic and growth via the oxidation of sulfide to sulfate or fermentation and litho/organotrophic oxygen or nitrate reduction (BEAM et al., 2014). While, the Fn1 thaumarchaea, which was described by Lin et al. (2015),

has potential to degrade long-chain fatty acids (LCFA) via β-oxidation. Recently, the cultivation of the heterotrophic *Conexivisphaera calidus* opened a new world to explore the metabolism of the non-AOA Thaumarchaeota.

In this context, herein we applied an integrated computational approach to explore multi-omics and complex metagenomic datasets. The description of two new Thaumarchaeota genomes guided us to infer and expand the knowledge about the evolution, metabolism, and biogeography of Thaumarchaeota group. Our results suggest that the non-ammonia oxidizing (non-AOA) Thaumarchaeota evolved in a mesophilic environment, sharing a (hyper)-thermophile last common ancestor with the ammonia-oxidizing group. Based on the analyses of more than 27,000 individual environmental samples, we discovered that this clade has subgroups with habitat-specific distribution (*e.g.*, Group 1.1c is specific of soil and non-saline sediments) and has potential roles in the degradation of the organic carbon and nitrogen, indicating a heterotrophic lifestyle and potential syntrophism in sediments.

## 3.2. Materials and Methods

We developed and applied an integrative computational approach to modeling the evolutionary history, potential metabolic pathways and biogeographical distribution of Thaumarchaeota (Supplementary Figure 1). The first step was to process the raw metagenomic data and filter low quality reads (as described below). After, the metagenomic reads were assembled into contigs and binned to reconstruct individual microbial genomes (MAGs). Once we had the MAGs, from the taxonomic identification we retrieved all Thaumarchaeota genomes deposited in the NCBI database, which was not explored in an evolutionary and metabolic context. To regard the biogeographical distribution, we used the Earth Microbiome Project database and updated the taxonomy using the new version of the Silva database. Lastly, the manual curation by an expert team to check specific informations about the potential metabolic pathways was performed. During the Bioinformatics analyses we carry out state-of-the-art phylogenetic analysis and genome assembly from metagenomics data.

### 3.2.1. Soil sampling, DNA extraction, and metagenomic sequencing.

The floodplain forest sediment was collected in the Tapajós National Forest, in the state of Pará, Eastern Amazon, Brazil, in the dry season (November 2015). The sampling area (S2 49.077 W55 02.077) is completely flooded by the Tapajós river during the wet season and

commonly remains flooded throughout the year. Sampling points were waterlogged, but there was no water column. At time of sampling, the sediments showed a pH ranging from 3,7 to 4,1 and a temperature 37,9 to 38,1. The samples were collected at three points, stored at 4 ºC for chemical analysis and at -20 ºC for DNA extraction and microbial analysis. Total DNA was extracted in duplicate using the PowerLyzer PowerSoil DNA Isolation Kit (MO Bio Laboratories, Carlsbad, CA, USA) and quality was assessed using agarose gel electrophoresis stained with GelRed$^{TM}$ (Biotium, Fremont, CA, USA) and a Nanodrop 2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). Metagenomic sequencing was performed using an Illumina HiSeq 2500 platform (2 x 250 bp) (Illumina, San Diego, CA, USA) at Novogene Corporation, Beijing, China, resulting in more than 155 million paired-end reads.

### 3.2.2. Metagenomic assembly, binning and quality control.

The SICKLE software (JOSHI; FAUS, 2011) was used to remove low-quality reads (parameters: Phred score <30 and minimum size <100 bp), while the assembly was performed with Megahit (LI et al., 2015), using default parameters (Supplementary Table 1). Subsequent processing was performed according to Lemos et al. (2017). Briefly, contigs smaller than 10,000 bp were removed from downstream analyses, and coverage was calculated using Bowtie2 (LANGMEAD; SALZBERG, 2012). Stringent length filtering parameters were used in order to reduce contamination and remove chimeric contigs. Binning was performed with MaxBin 2.0 (WU et al., 2016) and quality control metrics (completeness/contamination) were calculated with CheckM (PARKS et al., 2015). Taxonomy was assigned using the 16S rRNA gene with the RDP Classifier software (LAN et al., 2012) and the SILVA database (version 32) (QUAST et al., 2013). Additional phylogenomic information was inferred by the GTDB-Tk software and GTDB (Genome Taxonomy Database) (PARKS et al., 2018).

### 3.2.3. Thaumarchaeota/Nitrososphaeria genomes in public databases.

We retrieved from the Woodcroft et al. (2018) and Parks et al. (2017) all Thaumarchaeota genomes. Completeness and contamination were assessed with the CheckM software (PARKS et al., 2015) and all genomes with completeness < 85% and contamination

> 5% were filtered out (Supplementary Table 2). The potential non-ammonia oxidizing genomes from the “basal clade” (REN et al., 2019) was identified by phylogenomic analysis (details below) and that from soils or sediments were used in the downstream analysis.

### 3.2.4. Phylogenetic and phylogenomic analyses.

We selected 39 ammonia-oxiziding Thaumarchaeota genomes, the non-AOA genomes already described (Fn1, Beowulf, Dragon and *C. calidus*) and the genomes discovered by Rinke et al. (2013), Plominsky et al. (2018), Anantharaman et al. (2016) and Hua et al. (2018). However, we removed eleven genomes, which did not have a completeness > 85% and contamination < 5%, with the exception of Beowulf and Dragon, because they have a better studied and detailed metabolism (BEAM et al., 2014). We also add in this analysis one genome, from a partially thawed permafrosted sediment (WOODCROFT et al., 2018).

To reconstruct the evolutionary history of Thaumarchaeota, we used the concatenation of single-copy marker genes. Individual single-copy marker genes were identified using an HMM database described by Campbell et al., (2013) in Anvi’o software (EREN et al., 2015). A total of fourteen RP genes was identified in all Thaumarchaeota genomes (Supplementary Table 3), which include draft and near-complete genomes, and each individual gene was aligned using Muscle (EDGAR, 2004) and then they were concatenated. The phylogenomic tree was estimated using FastTree2 (PRICE; DEHAL; ARKIN, 2010). The trees were visualized in the iTol software (LETUNIC; BORK, 2016).

### 3.2.5. Functional annotation of the individual genomes.

Individual Thaumarchaeota genomes were initially annotated using PROKKA pipeline (SEEMANN, 2014), and complemented the annotation using arCOGs (Archaeal Clusters of Orthologous Genes) families (MAKAROVA et al., 2007) using PSI-BLAST (e-value ≤ 0.0005) (ALTSCHUL et al., 1997). The metabolic pathways were mapped using KEGG ontology implemented in BlastKOALA (KEGG Orthology and Links Annotation) (KANEHISA et al., 2016) with default parameters. Annotation was subsequently manually curated according to the guidelines described by Spang and collaborators (2012).

### 3.2.4. Meta-analysis of environmental distribution.

We re-analyzed the taxonomic assignment and updated the archaeal worldwide distribution of more than 27,000 samples deposited on the Earth Microbiome Project (EMP) (THOMPSON et al., 2017). The EMP representative OTUs (emp.90.min25.deblur.seq.fa) were re-assigned taxonomically using the most recent SILVA database (version 132) (QUAST et al., 2013) in the mothur software (SCHLOSS et al., 2009). The new OTU table was updated using the biom package (MCDONALD et al., 2012).

## 3.3. Results and discussions

### 3.3.1. Two new uncultivated non-ammonia oxidizing Thaumarchaeota

We recovered one near-complete Thaumarchaeota genome from a floodplain Amazon rainforest sediment metagenome (Table 1). The Amazon floodplain forest metagenome was assembled into 6,164,645 contigs from 156,056,159 paired-end reads (Supplementary Table 1). To recovery the individual genomes in the computational binning step, we used only the contigs ≥ 10,000 bp, which included 3,113 contigs totalizing 0.04 Gbp. The maximum contig length was 248,272 bp. We used very stringent length-filter parameters in order to reduce the contamination and remove probable chimeric contigs (LEMOS et al., 2017). The Thaumarchaeota genome was identified using the 16S rRNA gene taxonomy (Table 1). This genome has 86% of completeness with 1.94% of contamination, has a size of 1.44 Mbp (58 contigs and 39.83 of GC content), 1,587 predicted Coding Sequences (CDS), and a GC content of 39.83% (Table 1). We provisionally name this MAG Saci.

One new, previously undescribed Thaumarchaeal MAG was retrieved from NCBI (July/2018), with a completeness greater than ~95% and less than ~2% contamination (Table 1). This MAG was reconstructed originally from partially thawed bog samples (WOODCROFT et al., 2018) and its size is 2.84 Mbp, respectively, and was not explored in an evolutionary and metabolic context. We will refer to it throughout the manuscript as Bog.

### 3.3.2. Phylogenomic analysis of non-ammonia oxidizing Thaumarchaeota

The reconstruction of the evolutionary history revealed the presence of three monophyletic groups that were supported with bootstrap values (> 95%) (Figure 1). One group consists of Beowulf, Dragon and *Conexivisphaera calidus* (Basal group I), a second group is related with the new genomes described here (Saci and Bog) and other (Saci, Bog, UBA 141, YP1 and Fn1) (Basal Group II), and a third group composed only by ammonia-oxidizing Thaumarchaeota (Figure 1). The Basal Group I and II members do not have *amoABC* key-genes for ammonia oxidation and complete carbon fixation pathways (Supplementary Table 1 and REN et al., 2019). The Basal Group I is deep-branching taxa to the other Thaumarchaeota groups, while Saci, Bog and other (Basal Group II) are the closest relatives of the ammonia-oxidizing group. The phylogeny presented here highlights new data about the evolution of Thaumarchaeota phyla, once we corroborated the hypothesis of a thermophilic ancestor for the Thaumarchaeota and post adaptation to mesophilic biomes (BARNS et al., 1996; EME et al., 2013; ADAM et al., 2017). We also associated Beowulf, Dragon and *Conexivisphaera calidus* thaumarchaea genomes (Basal Group I) with the most deep-branching groups and the mesophilic lifestyle as a derived character (BROCHIER-ARMANET; GRIBALDO; FORTERRE, 2012). Furthermore, the ammonia oxidation metabolism is also a derived character, appearing later in the diversification of the Thaumarchaeota group, as predicted by Brochier-Armanet, Gribaldo and Forterre (2012), which suggest that the origin of the Thaumarchaeota and Aigarchaeota (the sister group of Thaumarchaeota) might be emerged from a thermophilic environment and the mesophilic lifestyle is a derived character, and reforces the recent findings about the evolution of this group (REN et al., 2019).

Table 1. Genome features and predicted lifestyle of the non-ammonia oxidizing Thaumarchaeota species

| Genome (bin) name | Saci | Dragon[a] | Beowulf[b] | Fn1[c] | Bog [d] | *Conexivisphaera calidus* |
|---|---|---|---|---|---|---|
| Genome Accession | This study. | 2263082000 (IMG) | 2519899514 (IMG) | 2558309099 (IMG) | GCA_003164815 (NCBI) | AP018732 (NCBI) |
| Habitat | Floodplain forest sediment | Acidic geothermal sediment | Acidic geothermal mat | Peatland | Partially thawed bog | Terrestrial acid hot spring |
| Taxonomy (16S rRNA) - Phylum - SILVA | Thaumarchaeota | Crenarchaeota | Thaumarchaeota | Thaumarchaeota | Thaumarchaeota | Aigarchaeota |
| Taxonomy (16S rRNA) - Class - SILVA | SCGC AB-179 | Crenarchaeota Incertae Sedis | SCGC AB-179 | 1.1c | 1.1c | Terrestrial Hot Spring Gp(THSCG) |
| Taxonomy (Phylogenomic) – Phylum - GTDB | Crenarchaeota | Crenarchaeota | Crenarchaeota | Crenarchaeota | Crenarchaeota | Crenarchaeota |
| Taxonomy (Phylogenomic) – Class - GTDB | Nitrososphaeria | Nitrososphaeria | Nitrososphaeria | Nitrososphaeria | Nitrososphaeria | Nitrososphaeria |
| Estimated Genome Size (bp) | 1,447,552 | 1,485,980 | 1,202,119 | 1,716,974 | 2,842,388 | 1,593,902 |
| Number of contigs | 58 | 38 | 110 | 90 | 115 | 1 |
| Estimated Completeness (%) | 86.33 | 89.16 | 83.98 | 97.25 | 99.03 | 98.1 |
| Estimated Contamination (%) | 1.94 | 0.97 | 0.97 | 2.91 | 1.94 | 0.00 |
| G+C content (%) | 39.83 | 40.20 | 41.92 | 56.33 | 59.44 | 62.1 |
| Maximum scaffold length (bp) | 139,074 | 153,490 | 51,712 | 108,222 | 132,590 | 1,593,902 |
| N50 contig length | 30,165 | 59,663 | 11,435 | 33,134 | 45,416 | 1,593,902 |
| CDS number | 1,587 | 1,827 | 1,437 | 1,920 | 3,224 | 1,610 |
| Metabolism | Fermentation | Fermentation or Organotrophic sulfur reduction | Litho/organotrophic oxygen or nitrate reduction | Fatty acid oxidation | Chemoheterotrophic | sulfur- and iron-reducing organoheterotroph |

[a,b]Data from Beam et al. (2014) and JGI (Joint Genome Institute) database; [c]Data from Lin et al. (2015) and JGI (Joint Genome Institute) database; [d]Data from Woodcroft et al. (2018) and NCBI (National Center for Biotechnology Information) database.

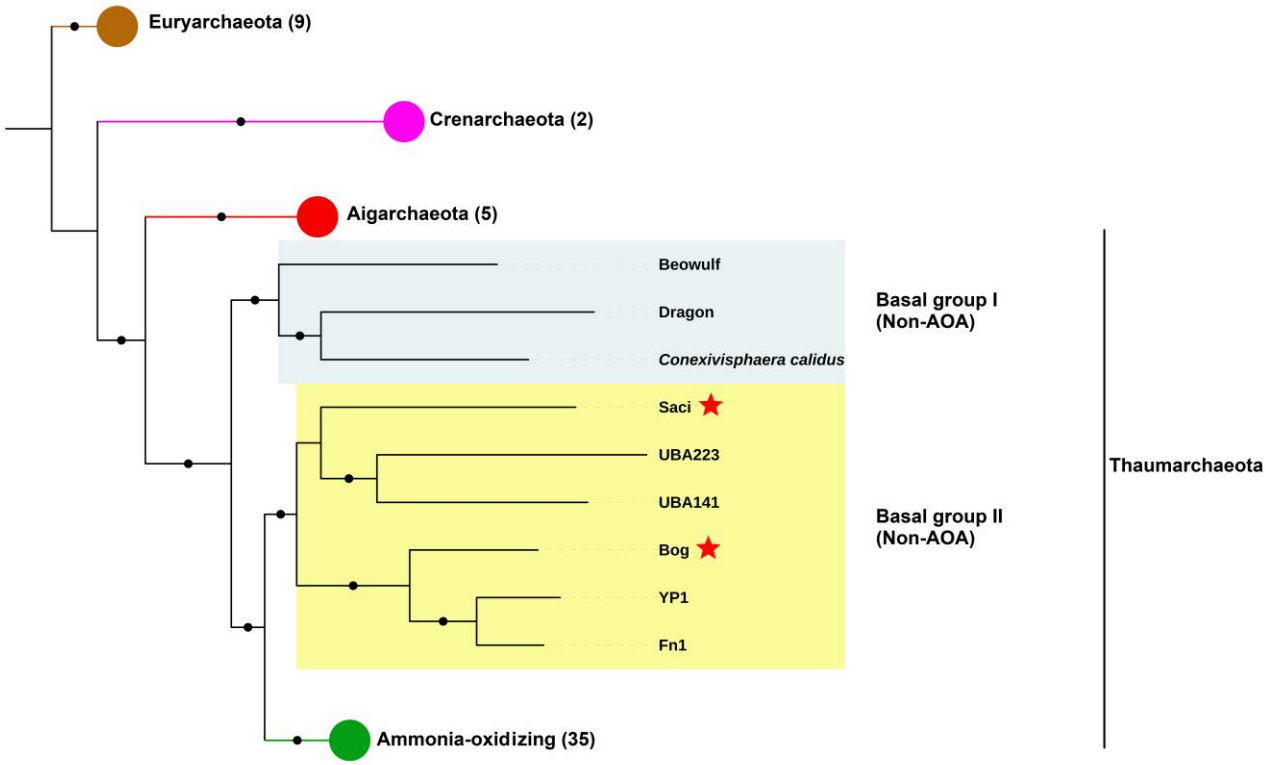



**Figure 1. Phylogenomic tree showing the evolutionary position of the ammonia and potential non-ammonia oxidizing Thaumarchaeota species.** The phylogenomic tree was infered using the alignment and the concatenation of 14 single-copy marker genes under the Jones-Taylor-Thorton model and CAT approximation with 20 rate categories. The new Thaumarchaeota genomes described here are assigned with a red star. Blue and yellow represent the potential non-ammonia oxidizing Thaumarchaeota and the green collapsed clade represents the ammonia-oxidizing Archaea. The nodes that showed a bootstrap support ≥95% are assigned with a black point in the trees

### 3.3.3. Biogeography distribution of Archaea and non-ammonia oxidizing Thaumarchaeota

To check the biogeographical distribution of the non-ammonia oxidizing Thaumarchaeota, we re-analyzed the relative abundance of the Archaea groups using more than 27,000 samples deposited on the Earth Microbiome Project database (THOMPSON et al., 2017) (Figure 2). Thaumarchaeota was the most abundant Archaea group of the microbiomes analyzed, which include soil (non-saline), sediment (saline) and water (saline). These findings are in accordance to other studies (BATES et al., 2011; FIERER, 2017; WEBSTER et al., 2015; PETRO et al., 2017). In addition, the clade related to the ammonia

oxidation metabolism (Nitrososphaeria) is the most abundant Thaumarchaeota class, and more abundant than the Group 1.1c, Marine Benthic Group A and SCGC_AB-179 (non-ammonia oxidizing groups related to Basal Group I and II) (Figure 2A). However, when we checked the subcategories of the Soil category, we can observe that the Group 1.1c had the relative abundance greater than 10% in soils from Coniferous forest, Forest, Montane Shrubland, Temperate coniferous forest, Tropical grassland, Tropical broadleaf forest and Tropical Shrubland (Figure 2B). This discrepancy was associated with the presence of more than 1,000 samples of Cropland and the absence of the Group 1.1c in these samples. The SCGC-AB-179 group was detected with a small proportion (0.006%)) in the Sediment (non-saline) samples (Figure 2C).

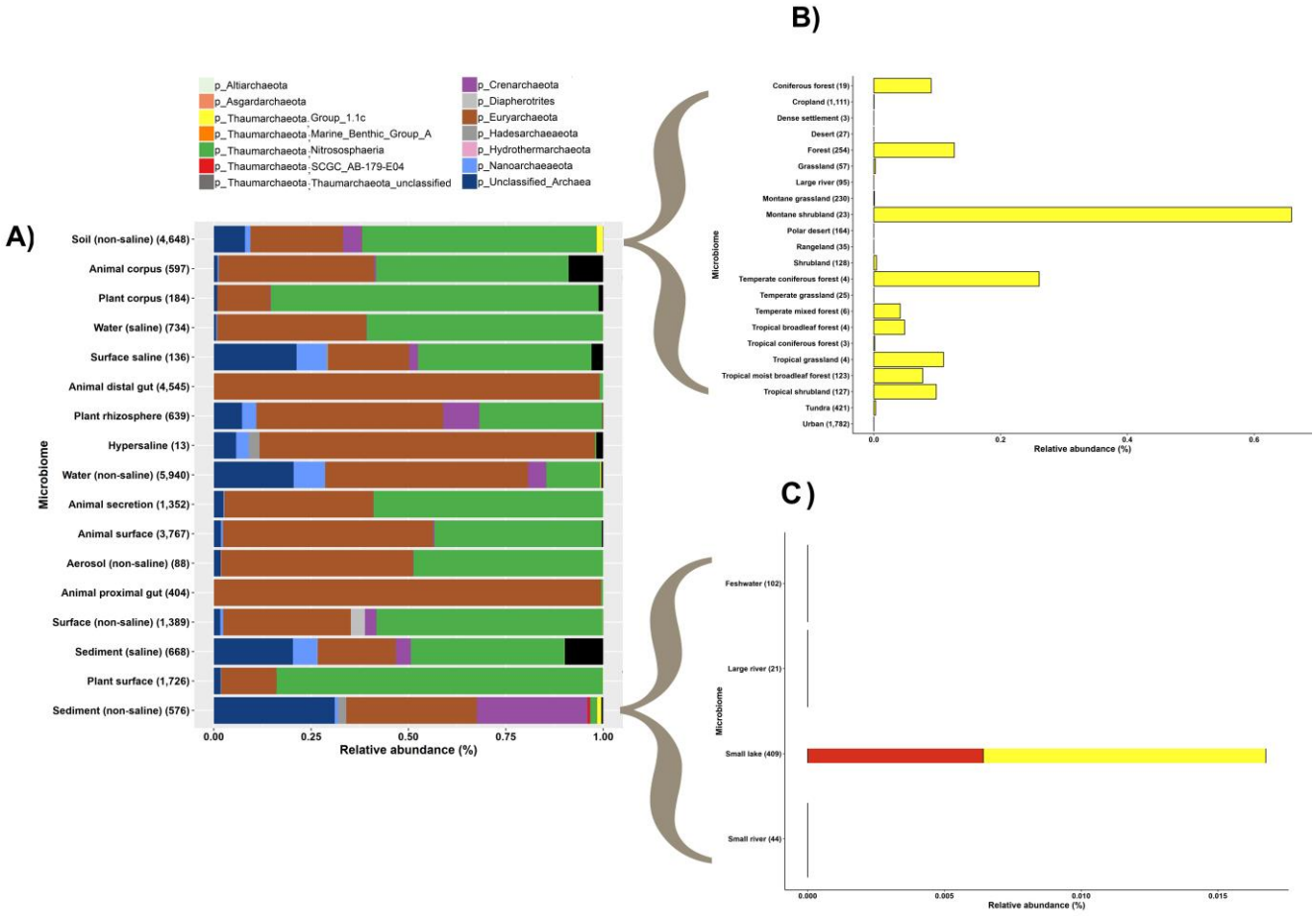



**Figure 2. Environmental distribution of the Archaea groups.** (A) Relative abundance of Archaea phylum of Earth Microbiome Project dataset using more than 27,000 environmental samples. (B) Soil (non-saline) sub-category. (C) Sediment (non-saline) sub-category. The taxonomy predictions were updated using RDP Classifier and the SILVA Database version 132

The high abundance of Nitrosophaeria group may be correlated with the process of nitrification in soils (ZHANG et al., 2012; ALVES et al., 2013). On the other hand, the Group 1.1c, which do not have potential to ammonia oxidizing (BEAM et al., 2014; WEBER et al., 2015), may have functional activity in other processes, such as the degradation of long-chain fatty acids (LCFA) (LIN et al., 2015), and soil organic nitrogen compounds (WEBER et al., 2015) or soil carbon degradation or both, as discussed below. The abundance of Group 1.1c suggests that it might have a functional potential in these environments, once it has been described as a group that respond to the variation of soil pH (LEHTOVIRTA; PROSSER; NICOL, 2009) and nutrient levels (NICOL et al., 2005). For the other hand, the SCGC-AB-179 group, represented by Saci, was identified with a significant relative abundance only in non-saline sediment.

### 3.3.4. New lineages of Thaumarchaeota may be associated with the heterotrophic lifestyle

The two new genomes (Saci and Bog) described here have potential to degrade organic carbon and/or nitrogen (Figure 3). Furthermore, we did not identify key functional traits linked to ammonia oxidation metabolism (e.g., amoABC genes, 4-hydroxybutyryl-CoA synthetase and 4-hydroxybutyryl-CoA dehydratase) (Supplementary Table 4).

#### 3.3.4.1. Saci thaumarchaea metabolism (Anaerobic acetate fermentation)

Saci has genes of the Glycolysis/Gluconeogensis (Embden-Meyeholf) and Pentose Phosphate Pathways (Figure 3A and Supplementary Table 4), which are part of the degradation of complex sugar polymers in simple sugars. The sugar uptake may occur by membrane transport proteins (e.g., ABC.SS.S - simple sugar transport system, and msmX - multiple sugar transport system), which were also identified. Although the glycolytic pathway is fragmented, which may be associated with the genome completeness, some key-enzymes were identified, such as phosphofructokinase, indicating the potential to degrade sugars.

The energy in Saci may be generated by the fermentation of acetate from AMP-forming acyl-CoA synthetase (Figure 3A) or from the use of the Fumarate as a Terminal Electron Acceptor (TEA) by succinate dehydrogenase/fumarate reductase. Furthermore, we identified only a small part of the genes related to TCA, suggesting that is not functionally in this genome. Saci genome described in this study has a potential capacity to hydrolyse

organic carbon using the Glycolytic Fermentation of Glucose to Acetate as carbon and energy sources. The glucose fermentation generates acetate from acetyl-CoA by one archaeal acetyl-CoA-synthetase (ADP-forming) [E.C. 6.2.1.13] acetyl-CoA + ADP + Pi → acetate + ATP + CoA) (SCHAFER; SELIG; SCHONHEIT, 1993) and has been described experimentally in many archaeal species, such as the strict anaerobe *Pyrococcus furiosus* (SAPRA; BAGRAMYAN; ADAMS, 2003) and *Haloarcula marismortui* (BRASEN; SCHONHEIT, 2004). Dragon thaumarchaea also has potential to strach degradation (BEAM et al., 2014), as such the recent described Bathyarchaeota (LAZAR et al., 2016) and Thoarchaeaota (SEITZ et al., 2016) phylum. The potential generation of acetate in floodplain forest adds Saci thaumarchaea in the complex syntrophic network, as potential acetate production, and involved in the plant-litter decomposition and soil organic matter formation (YARWOOD, 2018).

On the other hand, the Saci thaumarchaea has a fragmented-genome (completeness of 86.33% and other parts of this genome could be incomplete.). We found a few number of enzymes involved in the TCA cycle (Supplementary Table 4). However, Thoarchaeota, a recently described Archaea phylum, does not show the majority of enzymes that are necessary for this pathway (SEITZ et al., 2016). Similar to the Fn1, Saci has potential to generate energy from the anaerobic respiration using Fumarate as a terminal electron acceptor (TEA) under anoxic conditions (LIN et al., 2015). We also found one enzyme ribulose bisphosphate carboxylase/oxygenase (RuBisCO), which may be used to generate Glyceraldehyde-3P and contribute in the Glycolysi pathway, as already described for the Dragon and Beowulf thaumarchaeas (BEAM et al., 2014).

#### 3.3.4.2. Bog thaumarchaea metabolism (heterotrophic metabolism)

Bog has potential to grow using organic carbon (e.g., glucose) and maybe organic nitrogen compounds (e.g., casamino acids) (Figure 3B). Almost all genes of the Glycolysis/Gluconeogenesis Pathway were identified, with the exception of Phosphofructokinase (Supplementary Table 4). We also identified all genes of the Non-oxidative Phosphate Pathway, which could supplement the effectiveness of this pathway. The same peptide/amino acid/sugar transporters identified in Saci are also presented in Bog.

In addition, the major difference between Saci to Bog is the presence of a major metabolic potential capacity to degrade amino acids (Figure 3B). We identified 33 enzymes, which have potential function to degrade amino acids (Supplementary Table 5),

while Saci has only 13. The amino acids degradation can be used to the generation of intermediates of the Glycolysis (e.g., Alanine or Aspartate to Pyruvate) or TCA cycle (Valine, Isoleucine or Methionine to Succinyl-CoA) (Figure 3B). Weber et al. (2015) suggested that the Group 1.1c soil Thaumarchaeota grow with addition of organic nitrogen compounds (glutamate and casamino acids), but not only with organic carbon. We found the potential degradation of 11 amino acids (organic nitrogen in the form of amino acids) and organic carbon, indicating a heterotrophic growth and predicting that the amino acid degradation routes feed into the TCA cycle. The amino acids might be transported via the general amino acid permeases (e.g., ABC-type dipeptide/oligopeptide/nickel transport system") (SLACK et al., 1991) or specific amino acid permeases (e.g., ABC-type branched-chain amino acid transport system) (KOYANAGI et al., 2004). We hypothesized that the metabolism of amino acid degradation could improve the capacity to generate TCA cycle intermediates and improve the energy generation. Hobbie and Hobbie (2013) described that the uptake of amino acids intermediates, as such protein and oligopeptides, generate an extreme competition between the members of the soil microbial communities and the ability to uptake every amino acid could improve the fitness and survive. Furthermore, to complement the potential role of the Thaumarchaeota on C cycle, we also found a gene that encodes one endoglucanase (GH5), indicating its role in the breakdown of polysaccharides into simple sugars. The microbial cellulase is important for the decomposition of plant litter in wetland environments (YARWOOD, 2018) and has been identified in Euryarchaeota (e.g., *Thermococcus* sp. And *Pyrococcus horikoshii*) (WU; CONRAD, 2001; KANG; ISHIKAWA, 2007).

To our knowledge, this is the second genome assigned taxonomically into the Group 1.1c. The first described genome, Fn1 (LIN et al., 2015), has a metabolism associated with the degradation of long-chain fatty acids via beta-oxidation. In the Bog, we found almost all of the genes involved in this pathway, unless the gene that encode the 3-ketoacyl-CoA thiolase. Thiolase is important in the final step of the beta oxidation, producing Acetyl-CoA, which could be used in the energy generation (FUJITA; MATSUOKA; HIROOKA, 2007). We do not discard the possibility of on an anaerobic respiration, wherein Bog could use this pathway to regenerate NAD+ and a succinate dehydrogenase/fumarate reductase as TEA (LIN et al., 2015).

### 3.4. Conclusion

The phylum Thaumarchaeota has been studied for its important ecological role in the Nitrogen Cycle (e.g, ammonia oxidation). The integrated data analyzes presented here highlights new information about the phylogeny, potential metabolism and biogeography of the discrete and uncultivated non-ammonia oxidizing Thaumarchaeota class, focusing in two new near-complete genomes. This group also may be a habitat-specific and not a generalist, as such the ammonia-oxidizing clade (Nitrososphaeria). We found strongly potential functional evidences, which associated this new thaumarchaea with the soil organic nitrogen and carbon decomposition and the heterotrophic lifestyle. Our results expand previous studies describing the ecophysiology of non-ammonia oxidizing Thaumarchaeota, and open up new questions about the role of Thaumarchaeota in environmental samples.

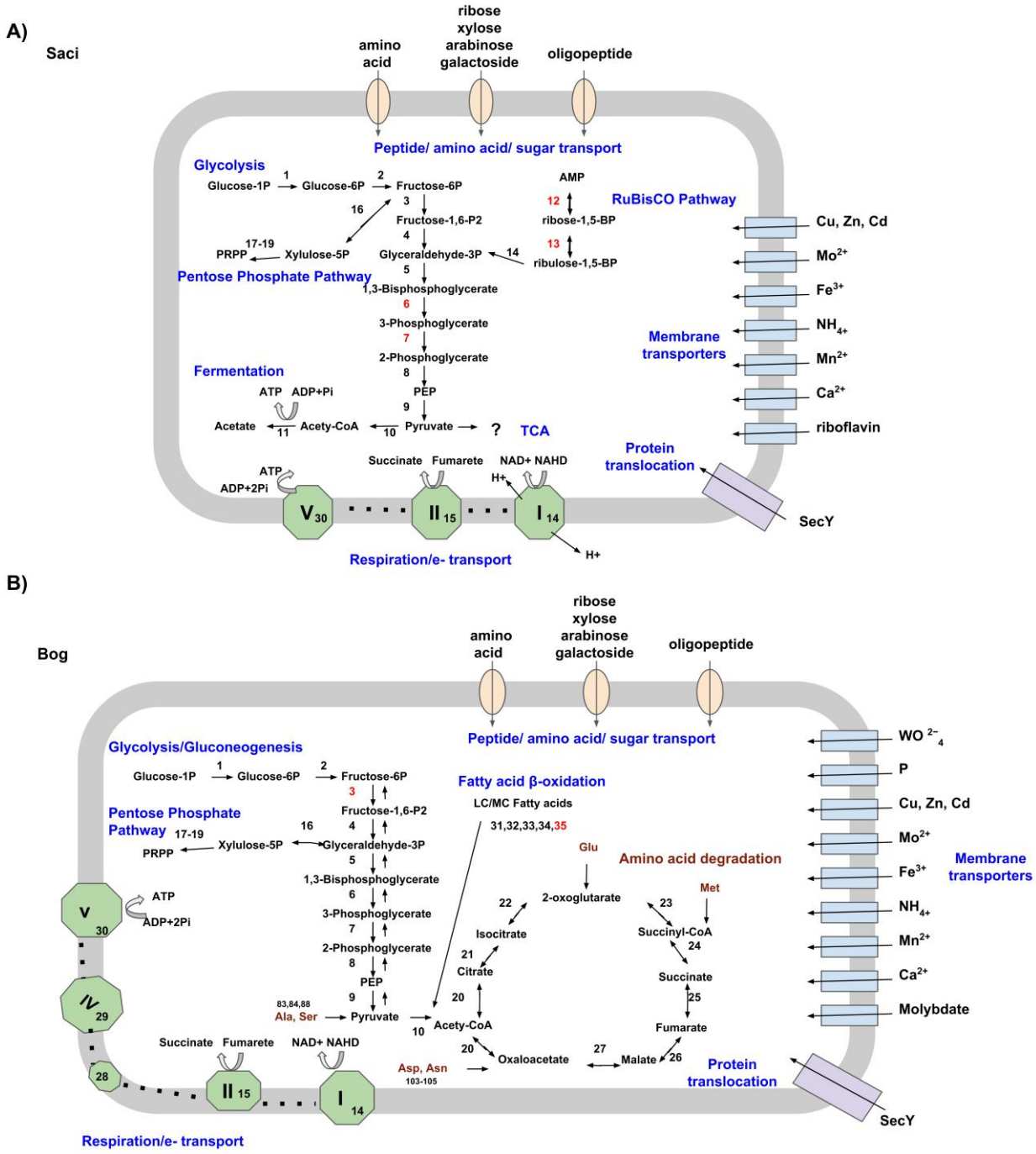



**Figure 3. Reconstructed key metabolic pathways of three new potential non-ammonia oxidizing Thaumarchaeota.** (A) Saci and (B) Bog thaumarchaea. Black numbers indicate the presence of enzymes of each individual metabolic pathway, and red numbers indicate the absence. Each number correspond to annotation detailed in Supplementary Table 4

**References**

ABBY, S. S. et al. Candidatus Nitrosocaldus cavascurensis, an Ammonia Oxidizing, Extremely Thermophilic Archaeon with a Highly Mobile Genome. **Frontiers in Microbiology**, Lausanne, v. 9, p. 28, 2018. doi: 10.3389/fmicb.2018.00028.

ADAM, P. S. et al. The growing tree of Archaea: new perspectives on their diversity, evolution and ecology. **The ISME Journal**, London, v. 11, n. 11, p. 2407–2425, 2017.

ALTSCHUL, S. F. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Research**, London, v. 25, n. 17, p. 3389-402, 1997.

ALVES, R. J. E. et al. Nitrification rates in Arctic soils are associated with functionally distinct populations of ammonia-oxidizing archaea. **The ISME Journal**, London, v. 7, n. 8, p. 1620–1631, 2013.

ALVES, R. J. E. et al. Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on amoA genes. **Nature Communications**, London, v. 9, n. 1, p. 1–17, 2018.

ANANTHARAMAN, K. et al. Metabolic handoffs shape biogeochemical cycles mediated by complex microbial communities. **Nature Communications**, London, v. 7, p. 13219, 2016.

ARÍSTEGUI, J. et al. Microbial oceanography of the dark ocean's pelagic realm. **Limnology and Oceanography**, Portland, v. 54, n. 5, p. 1501–1529, 2009.

BARNS, S. M. et al. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 93, n. 17, p. 9188–9193, 1996.

BATES, S. T. et al. Examining the global distribution of dominant archaeal populations in soil. **The ISME Journal**, London, v. 5, n. 5, p. 908–917, 2011.

BEAM, J. P. et al. Niche specialization of novel Thaumarchaeota to oxic and hypoxic acidic geothermal springs of Yellowstone National Park. **The ISME Journal**, London, v. 8, n. 4, p. 938–951, 2014.

BRÄSEN, C.; SCHÖNHEIT, P. Unusual ADP-forming acetyl-coenzyme A synthetases from the mesophilic halophilic euryarchaeon Haloarcula marismortui and from the hyperthermophilic crenarchaeon Pyrobaculum aerophilum. **Archives of Microbiology**, London, v. 182, n. 4, p. 277-287, 2004.

BROCHIER-ARMANET, C.; GRIBALDO, S.; FORTERRE, P. Spotlight on the Thaumarchaeota, **The ISME Journal**, London, v. 6, n. 2, p. 227-230, 2012.

CAMPBELL, J. et al. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 110, n. 14, p. 5540–5545, 2013.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. **Nucleic Acids Research**, London, v. 32, n. 5, p. 1792–1797, 2004.

ELOE-FADROSH, E. A. et al. Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. **Nature Communications**, London, v. 7, p. 10476, 2016.

EME, L. et al. Metagenomics of Kamchatkan hot spring filaments reveal two new major (hyper)thermophilic lineages related to Thaumarchaeota. **Research in Microbiology**, Amsterdam, v. 164, n. 5, p. 425–438, 2013.

EREN, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. **PeerJ**, San Diego, v. 3, e1319, 2015. doi: 10.7717/peerj.1319.

FIERER, N. Embracing the unknown: disentangling the complexities of the soil microbiome. **Nature Reviews Microbiology**, London, v. 15, n. 10, p. 579–590, 2017.

FUJITA, Y; MATSUOKA, M; HIROOKA, K. Regulation of fatty acid metabolism in bacteria. **Molecular Microbiology**, London, v. 66, n. 4, p. 829-839, 2007.

HOBBIE, J. E.; HOBBIE, E. A. Microbes in nature are limited by carbon and energy: the starving-survival lifestyle in soil and consequences for estimating microbial rates. **Frontiers in Microbiology**, Lausanne, v. 4, p. 324, 2013. doi: 10.3389/fmicb.2013.00324.

HUA, Z.-S. et al. Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. **Nature Communications**, London, v. 9, n. 1, p. 1–11, 2018.

JOSHI, N. A.; FASS, J. N. **Sickle**: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. San Francisco, 2011. Available at: https://github.com/najoshi/sickle.

KANEHISA, M. et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. **Nucleic Acids Research**, London, v. 45, n. D1, p. D353–D361, 2017.

KANG, H.-J.; ISHIKAWA, K. Analysis of active center in hyperthermophilic cellulase from Pyrococcus horikoshii. **Journal of Microbiology and Biotechnology**, Seoul, v. 17, n. 8, p. 1249–1253, 2007.

KATO, S. et al. Isolation and characterization of a thermophilic sulfur- and iron-reducing thaumarchaeote from a terrestrial acidic hot spring. **The ISME Journal**, London, 2019. doi: 10.1038/s41396-019-0447-3.

KEMNITZ, D.; KOLB, S.; CONRAD, R. High abundance of Crenarchaeota in a temperate acidic forest soil. **FEMS Microbiology Ecology**, Amsterdam, v. 60, n. 3, p. 442–448, 2007.

KEROU, M. et al. Proteomics and comparative genomics of Nitrososphaera viennensis reveal the core genome and adaptations of archaeal ammonia oxidizers. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 113, n. 49, p. E7937–E7946, 2016.

KÖNNEKE, M. et al. Isolation of an autotrophic ammonia-oxidizing marine archaeon. **Nature**, London, v. 437, n. 7058, p. 543–546, 2005.

KOYANAGI, T. et al. Identification of the LIV-I/LS system as the third phenylalanine transporter in Escherichia coli K-12. **Journal Bacteriology**, Washington, DC, v. 186, p. 343–350, 2004.

LAN, Y. et al. Using the RDP Classifier to Predict Taxonomic Novelty and Reduce the Search Space for Finding Novel Organisms. **PLoS One**, San Francisco, v. 7, n. 3, e32491, 2012.

LANGMEAD, B.; SALZBERG, S. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, London, v. 9, p. 357-359, 2012.

LEHTOVIRTA, L. E.; PROSSER, J. I.; NICOL, G. W. Soil pH regulates the abundance and diversity of Group 1.1c Crenarchaeota. **FEMS Microbiology Ecology**, Amsterdam, v. 70, n. 3, p. 367-376, 2009.

LEININGER, S. et al. Archaea predominate among ammonia-oxidizing prokaryotes in soils. **Nature**, London, v. 442, p. 806–809, 2006.

LEMOS, L. N. et al. Genome-Centric Analysis of a Thermophilic and Cellulolytic Bacterial Consortium Derived from Composting. **Frontiers in Microbiology**, Lausanne, v. 8, p. 1-16, 2017.

LETUNIC, I.; BORK, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. **Nucleic Acids Research**, Oxford, v. 44, p. W242–W245, 2016.

LI, D. et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. **Bioinformatics**, Oxford, v. 31, n. 10, p. 1674-1676, 2015.

LIN, X. et al. Metabolic potential of fatty acid oxidation and anaerobic respiration by abundant members of Thaumarchaeota and Thermoplasmata in deep anoxic peat. **The ISME Journal**, London, v. 9, n. 12, p. 2740–2744, 2015.

MAKAROVA, K. S. et al. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. **Biology Direct**, London, v. 2, p. 33, 2007. doi: 10.1186/1745-6150-2-33.

MCDONALD, D. et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. **GigaScience**, Oxford, v. 1, n. 1, p. 7, 2012.

MUELLER, P. et al. Global-change effects on early-stage decomposition processes in tidal wetlands – implications from a global survey using standardized litter. **Biogeosciences**, Göttingen, v. 15, n. 10, p. 3189–3202, 2018.

NICOL, G. W. et al. Primary succession of soil Crenarchaeota across a receding glacier foreland. **Environmental Microbiology**, Baltimore, v. 7, n. 3, p. 337–347, 2005.

PARKS, D. H. et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. **Genome Research**, New York, v. 25, n. 7, p. 1043–1055, 2015.

PARKS, D. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. **Nature Microbiology**, London, v.2, p. 1533–1542, 2017.

PARKS, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. **Nature Biotechnology**, London, v. 36, n. 10, p. 996–1004, 2018.

PETRO, C. et al. Microbial community assembly in marine sediments. **Aquatic Microbial Ecology**, Oldendorf, v. 79, n. 3, p. 177-195, 2017.

PESTER, M.; SCHLEPER, C.; WAGNER, M. The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. **Current Opinion in Microbiology**, London, v. 14, n. 3, p. 300–306, 2011.

PLOMINSKY, A. et al. Metabolic potential and in situ transcriptomic profiles of previously uncharacterized key microbial groups involved in coupled carbon, nitrogen and sulfur cycling in anoxic marine zones. **Environmental Microbiology**, Baltimore, v. 20, n. 8, p. 2727-2742, 2018.

PRICE, M.; DEHAL, P. S.; ARKIN, A. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. **PLoS One**, San Francisco, v. 5, n. 3, e9490, 2010. doi: 10.1371/journal.pone.0009490.

QUAST, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic Acids Research**, Oxford, v. 41, p. D590–D596, 2013. Database issue.

RAYMANN, K.; BROCHIER-ARMANET, C.; GRIBALDO, S. The two-domain tree of life is linked to a new root for the Archaea. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 112, n. 21, p. 6670–6675, 2015.

REN, M. et al. Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. **The ISME Journal**, London, v. 13, p. 2150–2161, 2019.

RINKE, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. **Nature**, London, v. 499, n. 7459, p. 431–437, 2013.

SAPRA, R.; BAGRAMYAN, K.; ADAMS, M. W. W. A simple energy-conserving system: Proton reduction coupled to proton translocation. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 100, n. 13, p. 7545-7550, 2003.

SCHÄFER, T. et al. Acetyl-CoA synthetase (ADP forming) in archaea, a novel enzyme involved in acetate formation and ATP synthesis. **Archives of Microbiology**, London, v. 159, n. 1, p. 72–83, 1993.

SCHLOSS, P. D. et al. Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. **Applied and Environmental Microbiology**, Washington, DC, v. 75, n. 23, p. 7537–7541, 2009.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, Oxford, v. 30, n. 14, p. 2068–2069, 2014.

SEITZ, K. W. et al. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. **The ISME Journal**, London, v. 10, n. 7, p. 1696-705, 2016.

SLACK, F. J. et al. Transcriptional regulation of a Bacillus subtilis dipeptide transport operon. **Molecular Microbiology**, Oxford, v. 5, n. 8, p. 1915–1925, 1991.

SPANG, A. et al. The genome of the ammonia- oxidizing CandidatusNitrososphaera gargensis: insights into metabolic versatility and environmental adaptations, **Environmental Microbiology**, Washington, DC, v. 14, n. 12, p. 3122–3145, 2012.

STOPNIŠEK, N. et al. Thaumarchaeal Ammonia Oxidation in an Acidic Forest Peat Soil Is Not Influenced by Ammonium Amendment. **Applied and Environmental Microbiology**, Baltimore, v. 76, n. 22, p. 7626–7634, 2010.

SUNAGAWA, S. et al. Computational eco-systems biology in Tara Oceans: translating data into knowledge. **Molecular Systems Biology**, Heidelberg, v. 11, n. 5, p.809, 2015. doi: 10.15252/msb.20156272.

THOMPSON, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. **Nature**, London, v. 551, n. 7681, p. 457–463, 2017.

TOURNA, M. et al. Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 108, n. 20, p. 8420–8425, 2011.

VANWONTERGHEM, I. et al. Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. **Nature Microbiology**, London, v. 1, n. 12, p. 16170, 2016.

VENTURINI, A. M. **Conversão floresta-pastagem na Amazônia Oriental**: impactos sobre as comunidades microbianas do metano do solo. 2019. Tese (Doutorado em Ciências) – Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, 2019.

WEBER, E. B. et al. Ammonia oxidation is not required for growth of Group 1.1c soil Thaumarchaeota. **FEMS Microbiology Ecology**, Amsterdam, v. 91, n. 3, 2015. doi: 10.1093/femsec/fiv001.

WEBSTER, G. et al. Archaeal community diversity and abundance changes along a natural salinity gradient in estuarine sediments. **FEMS Microbiology Ecology**, Amsterdam, v. 91, n. 2, p. 1–18, 2015.

WU, X. L.; CONRAD, R. Functional and structural response of a cellulose-degrading methanogenic microbial community to multiple aeration stress at two different temperatures. **Environmental Microbiology**, Baltimore, v. 3, n. 6, p. 355–362, 2001.

WU, Y.-W.; SIMMONS, B. A.; SINGER, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. **Bioinformatics**, Oxford, v. 32, n. 4, p. 605–607, 2016.

YARWOOD, S. A. The role of wetland microorganisms in plant-litter decomposition and soil organic matter formation: a critical review. **FEMS Microbiology Ecology**, Amsterdam, v. 94, n. 11, 2018. doi: 10.1093/femsec/fiy175.

ZHALNINA, K. V. et al. Genome sequence of Candidatus Nitrososphaera evergladensis from group I.1b enriched from Everglades soil reveals novel genomic features of the ammonia-oxidizing archaea. **PloS One**, San Francisco, v. 9, n. 7, p. e101648, 2014.

ZHANG, L.-M. et al. Ammonia-oxidizing archaea have more important role than ammonia-oxidizing bacteria in ammonia oxidation of strongly acidic soils. **The ISME Journal**, London, v. 6, n. 5, p. 1032–1045, 2012.

WOODCROFT, B.J. Genome-centric view of carbon processing in thawing permafrost. **Nature**, London, v. 560, p. 49-54, 2018.

# 4. WHEN IT COMES TO THE SOIL MICROBIAL GENOMES, DOES SIZE REALLY MATTER?


## ABSTRACT

Soil microbiome is one of the most complex biological systems. State-of-the-art molecular approaches such those based on single-amplified genomes (SAGs) and metagenome assembled-genomes (MAGs) are now improving our capacity for disentailing soil microbial mediated-process. The complexity of soil microbial functions is usually related to increased genome sizes, which may improve the microbial fitness in a scenario with diverse but scarce resources. However, we contend that small-genome microorganisms may play a role in soil but are usually neglected by most of the studies. Here, we explored two reference soil microbiota datasets on the basis of the genome size of their representatives. Additionally, we also used two MAGs belonging to the new CPR/Patescibacteria phylum reconstructed from a cattle-pasture Amazon soil metagenome to complement our comparative analyzes. Our results suggest that microorganisms hosting small genomes exert peculiar functions in soil. Additionally, the use of MAGs may be a better choice over SAGs to expand the soil microbial databases.


Keywords: Patescibacteria; Candidate Phyla Radiation (CPR); Metagenome-assembled genomes (MAGs); Amazon; Soil

## 4.1. Short communication

Many hypotheses may explain the complexity and high diversity of soil microbiomes. In the genomic context, Raes and colleagues (2007) argued that each habitat selects a specific range of microbial genome sizes, regarding the environment stability (ANGLY et al., 2009), where more stable environments select microorganisms with small genomes and less non-redundant functions (MORRIS; LENSKI; ZINSER, 2012), such as parasites (e.g., *Mycoplasma pneumoniae*) (HIMMELREICH et al., 1996) and symbionts (MCCUTCHEON et al., 2009). On the other hand, complex environments favor microorganisms with larger genomes and accessory genes, with greater metabolic versatility, have the ability to survive and acclimate in a changing-environment with diverse but limited resources, like soil (DINI-ANDREOTE et al., 2012; KONSTANTINIDIS; TIEDJE, 2004). In fact, the estimated genomes sizes from public available soil metagenomes, including natural (e.g., Amazon rainforest) and agriculture soils (e.g., soybean), ranges from 4.5 to 8.0 Mbp (SORENSEN et al., 2019). However, these values may be biased due the methods applied for generating this data.

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. For example, the recently proposed Candidate Phyla Radiation (CPR)/Patescibacteria (BROWN et al., 2015) has not yet been cultivated in higher numbers (up to date, only the strain TM7x has been cultivated (HE et al., 2015)). However, they represent nearly 15% of the domain Bacteria. Furthermore, the small genome size (<1.5 Mbp) is a common genomic trait shared between all members of the CPR/Patescibacteria group, including the lack of biosynthetic capabilities (BROWN et al., 2015) and potential for co-metabolism interdependencies (HE et al., 2015). These biological traits could prove to be the strong challenge for cultivating these organisms.

To alleviate these issues, massive DNA sequencing methods and bioinformatics tools have been developed to reconstruct complete or near-complete microbial genomes from metagenomic datasets, and access their potential functional role (WU et al., 2014). Different approaches are used with this aim, e.g., (i) single amplified genome (SAG), a strategy to sequence genomes of individual cells, and (ii) metagenome-assembled genomes (MAGs), by the use of metagenomics approaches. Strategies to recovery MAGs, also known as “binning”, are based on using compositional signatures (e.g., GC content and coverage) for clustering

post-assembled sequences (WU et al., 2014). Each subpopulation derived from this analysis represents a potential individual genome (referred as MAG). The use of these approaches has revolutionized our view about microbial metabolism, diversity and evolution of soil microbial diversity (WOODCROFT et al., 2018; SORENSEN et al., 2019). Similarly, single-cell genomics has also been applied to target the unknown microbial diversity, but usually for less complex environments such as aquatic microbiomes (RINKE et al., 2013) and acid mine drainage samples (MEDEIROS et al., 2017). The application of the single-cell sequencing on soil microbiome studies is limited, since microbial communities are more heterogeneous in soil particles, and other unsolved challenges related to cell capture and downstream analysis due to the complex nature of the ecosystem (EICHORST et al., 2015).

Here, we applied an integrated meta-analysis of public available genomes and metagenomes, aiming to explore the genome size features of soil microorganisms using two datasets: (i) the recently launched RefSoil database (CHOI et al., 2017), and (ii) the metagenome assembled-genomes (MAGs) from the thawing permafrost deposited on NCBI (WOODCROFT et al., 2018). We also used two reconstructed genomes belonging to the new CPR/Patescibacteria phylum reconstructed from a cattle-pasture Amazon soil metagenome to complement our comparative analyzes (Supplementary Material and Methods).

Our analysis revealed that the average size of the microbial genomes available in the RefSoil (CHOI et al., 2017) was 4.5 ± 1.0 Mbp (Figure 1A). Similarly, almost all MAGs retrieved from the thawing permafrost dataset had their average genome size close to those observed in the RefSoil (Figure 1B). However, CPR/Patescibacteria genomes had an average genome size of 0.9 ± 0.2 Mbp (985.282 ± 283.457 bp). The size of the soil CPR/Patescibacteria genomes is 4-fold smaller than the mean identified in the RefSoil and the thawing permafrost databases. The same pattern was also found when we checked the genome size of the two new CPR/Patescibacteria MAGs reconstructed using a cattle-pasture of Amazon soil metagenome dataset (Supplementary Table 1 and 2), which are also similar to the soil CPR/Patescibacteria described by Kroeger and collaborators (2018) using tropical soil metagenomes. We provisionally name these MAGs Caipora and Curupira.

To better understand the relationship between the genome size and the potential functions performed by soil microorganisms, we also deeply explore the functions of 18 new CPR/Patescibacteria reconstructed in the thawing permafrost metagenomes (WOODCROFT et al., 2018), together with the two new CPR/Patescibacteria described here and the genomes described by Kroeger et al. (2018). The functional profile of the CPR/Patescibacteria representatives was based on COG (Clusters of Orthologous Groups) genome annotation data,

through a set of multivariate statistics, and compared with the functional profile of all other genomes deposited on RefSoil database. We observed that members of this phylum harbor similar functional profiles, but very different from other microbial phyla (Figure 2A), indicating a functional redundancy.

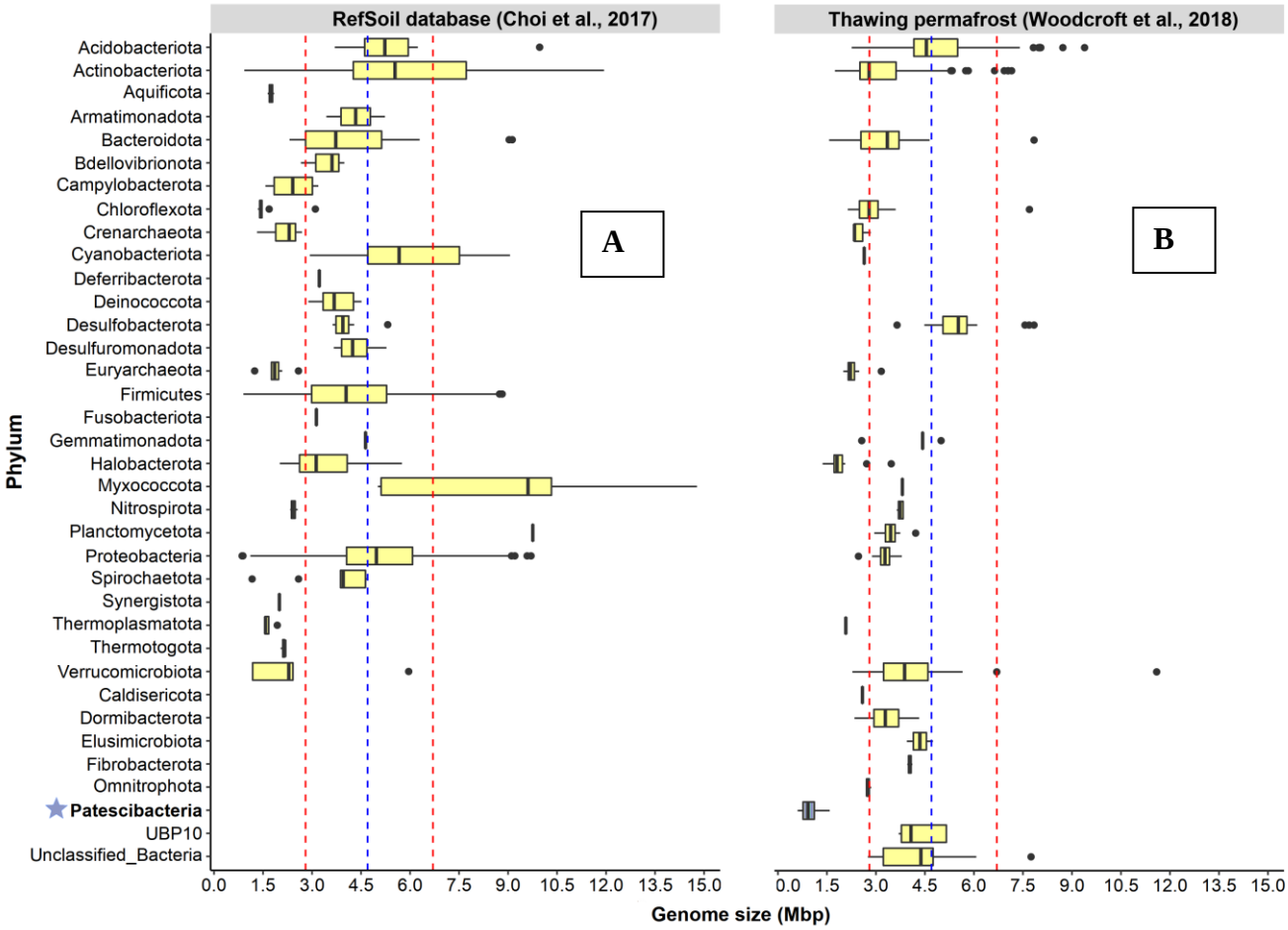



**Figure 1. Soil genome size distributions.** (A) RefSoil database and (B) Metagenome-assembled genomes (MAGs) from thawing permafrost metagenomes

Regarding to the potential central metabolism, we identified that all soil CPR/Patescibacteria studied here lack the functional Tricarboxylic Acid Cycle (TCA) pathway and electron transport chain to generate ATP, but some of them may ferment organic compounds via glycolysis pathway, generating lactate as final products (Figure 2B and 2C). Overall, the soil CPR/Patescibacteria genomes described here also lacks genes required for the *De novo* biosynthesis of nucleotides, amino acids and cofactors (Supplementary Table 3 and Figure 2B-C). These metabolic limitations could suggest an episymbiotic lifestyle

(CASTELLE et al., 2018) 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., 2014). On the other hand, a new soil bacterium called Candidatus *Udaeobacter copiosus* (Verrucomicrobia) was recently described (BREWER et al., 2017) as being free-living and hosting a 2.81 Mbp genome. Metabolic predictions indicated that C. *U. copiosus* could keep a reduced genome by acquiring costly amino acids and vitamins from the environment (BREWER et al., 2016). A few number of genomes shorter than 1.5 Mbp are available on RefSoil (Supplementary Table 4), and all of them have a parasitic lifestyle (for example *Neorickettsia* and *Tropheryma*). These findings reinforce the hypothesis that soil CPR/Patescibacteria could also be associated with a symbiotic/parasitic lifestyle.

Our data also highlight the importance of binning methods for the expansion of RefSoil database. In the original paper, which described the RefSoil database (CHOI et al., 2017), the authors recommended the use of single-cell methods. However, the single-cell genomes presented by Choi et al. (2017) did not present good quality recommended by the Minimum information about a single amplified genome (MISAG) standards (BOWERS et al., 2017) (Supplementary Table 5). We argue that the binning approaches may complement the single-cell approaches to expanding the RefSoil database allowed us a more complete and informative soil microbial reference database.

In conclusion, the small-sized genome is a peculiar trait of the CPR/Patescibacteria phyla members living in thawing permafrost and cattle-pasture soils. Here, we expanded the range of environments within the radiation of this bacteria group and their ecological role, including two distinct soils that showed CPR/Patescibacteria with similar functions (e.g., fermentation). These findings indicate a possible syntrophy between CPR and other microorganisms, such as methanogenic archaea or acetogenic bacteria during the soil organic matter degradation, revealing a functional redundancy between the CPR/Patescibacteria in the soil microbiome. Furthermore, soil CPR/Patescibacteria lacks essential biosynthetic functions (e.g. *de novo* amino acids and nucleotide biosynthesis) indicating a symbiotic lifestyle (e.g. cell surface attached). Also, further study is required to better elucidate the ecology of CPR/Patescibacteria, such as the design of new 16S rRNA primers to measure the abundance and structure of CPR/Patescibacteria in soil microbial communities, and their metabolism using metatranscriptomics and/or RNA-SIP.

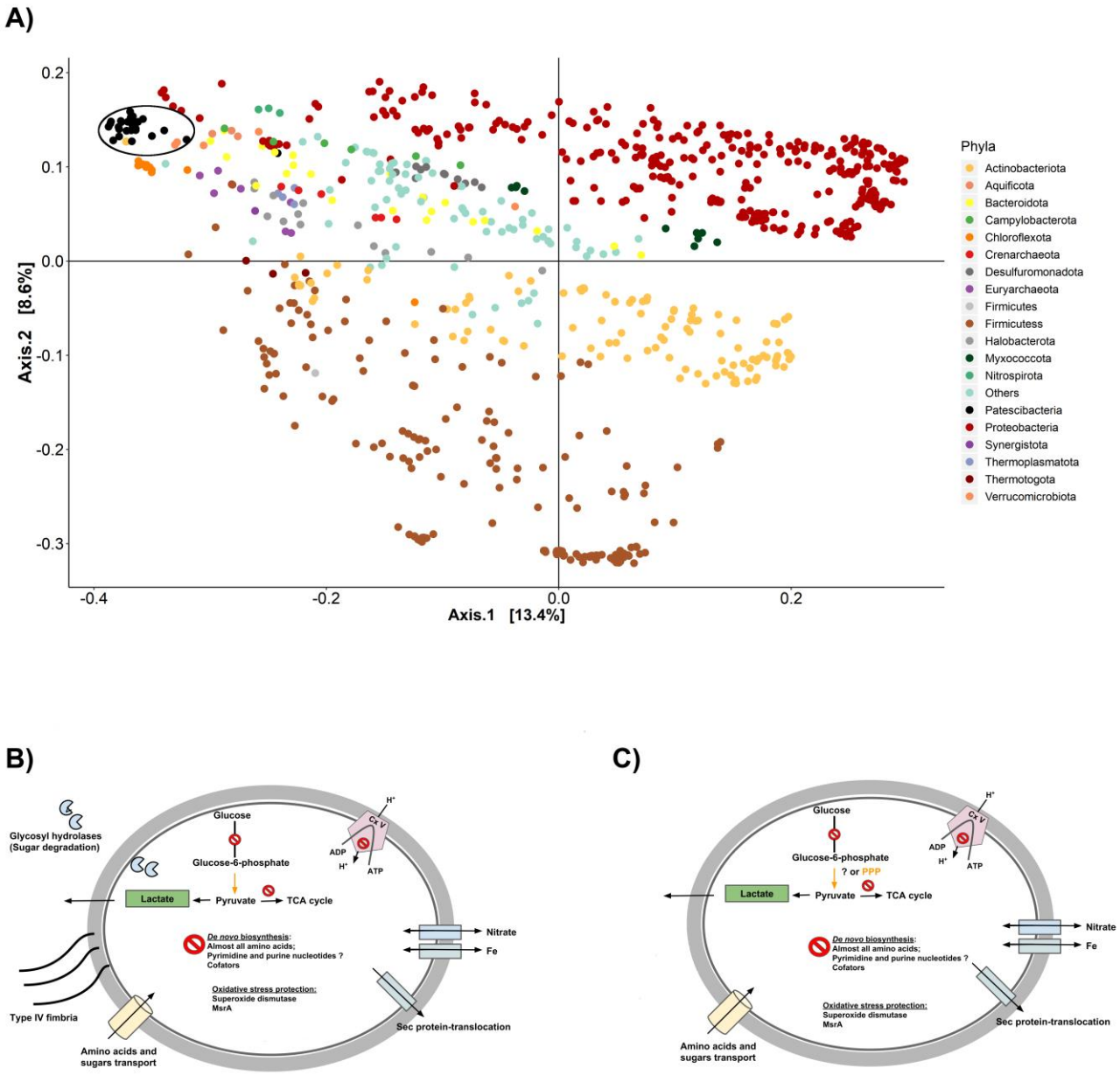



Figure 2. CPR/Patescibacteria functional genome profile. (A) Similarity between the soil microbial genomes and the Patescibacteria phyla. (B-C) Patescibacteria genome from Amazon cattle-pasture (Caipora) (B) and thawing permafrost (GCA_003151615.1) (C) soils

## References

ANGLY, F. et al. The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. **PloS Computational Biology**, San Francisco, v. 5, n. 12, 2009.

BROWERS, R.M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. **Nature Biotechnology**, London, v. 35, n. 8, p. 725-731, 2017.

BREWER, T. E. et al. Genome reduction in an abundant and ubiquitous soil bacterium 'Candidatus Udaeobacter copiosus'. **Nature Microbiology**, London, v. 2, n. 2, p. 16198, 2017.

BROWN, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. **Nature**, London, v. 523, n. 7559, p. 208–211, 2015.

CASTELLE, C. J. et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. **Nature Reviews Microbiology**, London, v. 16, n. 10, p. 629–645, 2018.

CHOI, J. et al. Strategies to improve reference databases for soil microbiomes. **The ISME Journal**, London, v. 11, n. 4, p. 829–834, 2017.

DANIEL, R. The metagenomics of soil. **Nature Reviews Microbiology**, London, v. 3, n. 6, p. 470–478, 2005.

DINI-ANDREOTE, F. et al. Bacterial genomes: habitat specificity and uncharted organisms. **Microbial Ecology**, Amsterdam, v. 64, n. 1, p. 1–7, 2012.

EICHORST, S. A. et al. Advancements in the application of NanoSIMS and Raman microspectroscopy to investigate the activity of microbial cells in soils. **FEMS Microbiology Ecology**, Amsterdam, v. 91, n. 10, 2015.

HE, X. et al. Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 112, n. 1, p. 244–249, 2015.

HIMMELREICH, R. et al. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. **Nucleic Acids Research**, Oxford, v. 24, n. 22, p. 4420–4449, 1996.

KONSTANTINIDIS, K. T.; TIEDJE, J. M. Towards a Genome-Based Taxonomy for Prokaryotes. **Journal of Bacteriology**, Washington, DC, v. 187, n. 18, p. 6258–6264, 2005.

KROEGER, M. E. et al. New biological insights into how deforestation in Amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. **Frontiers in Microbiology**, Lausanne, v. 9, p. 1635, 2018. doi: 10.3389/fmicb.2018.01635.

MCCUTCHEON, J. P. The bacterial essence of tiny symbiont genomes. **Current opinion in Microbiology**, London, v. 13, n. 1, p. 73-78, 2010.

MEDEIROS, J. D. et al. Single-cell sequencing unveils the lifestyle and CRISPR-based population history of Hydrotalea sp. in acid mine drainage. **Molecular Ecology**, Amsterdam, v. 26, n. 20, p. 5541–5551, 2017.

MORRIS, J. J.; LENSKI, R. E.; ZINSER, E. R. The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. **MBio**, Washington, DC, v. 3, n. 2, e00036-12, 2012.

RAES, J. et al. Prediction of effective genome size in metagenomic samples. **Genome Biology**, London, v. 8, n. 1, p. R10, 2007.

RINKE, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. **Nature**, London, v. 499, n. 7459, p. 431–437, 2013.

ROESCH, L. F. W. et al. Pyrosequencing enumerates and contrasts soil microbial diversity. The **ISME Journal**, London, v. 1, n. 4, p. 283–290, 2007.

SORENSEN, J. W. et al. Ecological selection for small microbial genomes along a temperate-to-thermal soil gradient. **Nature Microbiology**, London, v. 4, n. 1, p. 55–61, 2019.

WOODCROFT, B. J. et al. Genome-centric view of carbon processing in thawing permafrost. **Nature**, London, v. 560, n. 7716, p. 49–54, 2018.

WU, Y.-W. et al. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. **Microbiome**, London, v. 2, n. 1, p. 26, 2014.

# 5. EFFECTS OF FOREST-TO-PASTURE CONVERSION AND INCREASE IN SOIL MOISTURE LEVELS ON ARCHAEA COMPOSITION IN AMAZON SOILS


## ABSTRACT

The Amazon rainforest is one of the most diverse biomes of the Earth. However, recent predictions suggest an increase in precipitation during the Amazonian wet season as a result of global climate change, and these changes can be intensified with the forest-to-pasture conversion, altering important soil properties, such as soil moisture. Here, the hypothesis that the forest-to-pasture conversion and the increase of soil moisture levels modify the Archaea community composition (potential methanogens and non-methanogens) was tested. To test this hypothesis, forest and pasture soil samples from the Eastern Amazon Forest were collected and a 30-day microcosm incubation experiment with four moisture levels (control, 60, 80, and 100% of field capacity) to simulate extreme rainfall events (floods) in the rainy season was conducted. Further, the structure and composition of the archaeal communities using 16S rRNA amplicon sequencing were analysed using bioinformatics and statistical models. The soil methane fluxes were also analyzed on a gas chromatograph from the same samples, and the results were integrated with the Archaea abundance. As already described for this microcosm experiment, forest and pasture soils under 60% and 80% FC presented negative/or neutral emission values, but with the increase of soil moisture to 100% FC, primary forest and pasture soils started acting as methane source, with the latest soil presenting the highest emission. Similar to the methane emission in pasture soils, there was a significant correlation in decreased relative abundance of Thaumarchaeota and the increased potential methanogenic phyla Bathyarchaeota with increasing in field capacity of the pasture soils. Analysis based on beta diversity indicated that archaeal communities were first strongly determined by long-term (land-use change) and then by short-term (moisture level) perturbation. Furthermore, soil moisture controlled more the increase of beta diversity in pasture than in forest soils. Our results indicated that the community alterations caused by the higher soil moisture levels were most pronounced in pasture, where communities from pasture were more sensitive, enhancing the potential of methanogenesis, while forest may act as buffers during the rainy season and harbor more stable communities. Given the intensification of forest-to-pasture conversion in the Amazon region and the possible prolongation of the rainy season as a result of climate change, methane production could increase because of the effects of these perturbations on the archaeal community composition.


Keywords: Deforestation; Microcosms experiment; Methanomassiliicoccales; Bathyarchaeota

## 5.1. Introduction

The Amazon rainforest is a great reservoir of biological diversity, including 25% of the world's terrestrial animal and plant species, and is responsible for regulating biogeochemical cycles (WILSON et al., 2016) with effects on the climate (MALHI et al., 2008). This also includes the importance of microbial communities in maintaining a functional equilibrium in native forests in this region (MENDES et al., 2015). However, previous studies demonstrated that land-use changes had altered the abundance, composition and diversity of specific bacterial taxa detected in these soils, such as Acidobacteria (NAVARRETE et al., 2015), Verrucomicrobia (Ranjan et al., 2015), and also Fungi (MUELLER et al., 2014).

Up to date, only a small number of studies have investigated the diversity of Archaea in Amazon soils (HAMAOUI et al., 2016; NAVARRETE et al., 2011; TUPINAMBÁ et al., 2016). This group is widespread in most environments and has many important functions in soil ecosystems, such as methanogenesis through anaerobic degradation of organic matter and N cycling through oxidation of ammonia to nitrite (OFFRE; SPANG; SCHLEPER, 2013). Methanogenic Archaea are found in anaerobic environments and can grow and produce methane through multiple processes. These processes include reduction of different carbon compounds such as carbon dioxide and hydrogen (hydrogenotrophic methanogens), methyl compounds (methylotrophic methanogens) and acetate (acetoclastic methanogens) (JABLÓNSKI; RODOWICZ; LUKASZEWICZ, 2015). Classical microbiological studies based on the culture-dependent methods indicated that only the seven Euryachaeota classes (Methanococcales, Methanopyrales, Methanobacteriales, Methanosarcinales, Methanomicrobiales, Methanocellale and Methanomassiliicoccales) are methanogens (HEDDERICH; WHITMAN, 2006; BORREL et al., 2013). The orders Methanococcales, Methanopyrales, Methanobacteriales, Methanomicrobiales, and Methanocellales are strictly hydrogenotrophic (LIU; WHITMAN, 2008), while Methanosarcinales are more versatile, which includes *Methanosarcina* sp., with the the ability to use the three pathways (SPRENGER et al., 2000). The Methanomassiliicoccales order is able to uses many methylated compounds (e.g., methylamines, monomethylamine, dimethylamine and trimethylamine) (BORREL et al., 2013). With the advent of the high-throughput sequencing, it has been possible to reconstruct genomes from metagenomes and explored the metabolic potential through genome-centric analysis of the new yet not culturable microorganisms

(EVANS et al., 2015; LEMOS et al., 2017). This approach allowed the discovery of new taxonomic groups of potential methanogenic Archaea, namely Bathyarchaeota (EVANS et al., 2015) and Verstraetearchaeota (VANWONTERGHEM et al., 2016), which expanded the groups with the ability to produce methane by the fermentation of various methylated substrates. The phylum Thaumarchaeota, on the other hand, has metabolic traits associated with ammonia oxidation (KEROU et al., 2016) and the most studied class is Nitrososphaeria (TOURNA et al., 2011). However, recent studies have indicated that the Group 1.1c within Thaumarchaeota is not capable of oxidizing ammonia and they are widespread in soils (WEBER et al., 2015), hot springs (BEAM et al., 2014) and marine sediments (LIN et al., 2015).

Land-use change in the Amazon region has been associated with alterations in cloudiness and precipitation (WANG et al., 2009a), with a recent prediction suggesting increased precipitations during the wet season (GLOOR et al., 2015). The intensification of deforestation in the Amazon region and the possible prolongation of the wet season as a result of climate change may alter soil properties. In this case, the variability of soil parameters like moisture, which is also linked to land-use changes (VERCHOT et al., 2010), could affect specific microbial processes such as ammonia oxidation (DI et al., 2014) and methanogenesis (BREWER et al., 2018). Thus, alterations in the precipitation, and moisture in the Amazon soils could affect the structure of the archaeal community linked to methanogenesis, further altering the flux of methane in this region. This group of microorganisms also respond to changes in multiple environmental factors including soil pH (TRIPATHI et al., 2013), moisture content and C:N ratio (SHI et al., 2016). However, knowledge on biology of the archaeal communities and their roles in biotic and abiotic factors of Amazon soil is very poorly understood.

Amazon soils play an important role in the global methane cycle, and changes in land-use and precipitation may also alter its balance (source to sink or vice and verse). In general, there are many evidences that demonstrate the fact that Amazon forest soils are a sink for atmospheric methane (FERNANDES et al., 2002; KELLER et al., 2005) and that these soils might have the potential to consume approximately 470 mg $C\text{-}CH_4$ per $m_2$ per year (STEUDLER et al., 1996). Whereas advances in deforestation and establishment of pastures leads to decrease in their potential to act as carbon sink. Several studies have demonstrated this fact, with pastures potentially emitting large amounts of methane of about 270 mg $C\text{-}CH_4$ per $m_2$ per year (FERNANDES et al., 2002 and STEUDLER et al., 1996). Furthermore, studies have also revealed that increasing water content in these pasture soils may further

stimulate their methane production through changes (e.g. taxonomic and functional) in the archaeal community (KELLER et al., 2005; MEYER et al., 2017; MUTSCHLECHNER et al., 2018; PREM et al., 2014). Predicted intensifications of both land-use (i.e. forest-to-pasture conversion) and rainfall in the Amazon region are major concerns, as these changes may lead to an increased methane production, further affecting the global climate.

Here, the responses of archaeal communities to long-term (land-use change) and short-term (soil moisture level alterations) disturbance in Amazon soils were explored. We hypothesized that land-use change (forest-to-pasture conversion) and the increase of soil moisture levels modify the archaea community composition (e.g., potential methanogens and non-methanogens). To test our hypothesis, we collected soil samples from a native forest and a cattle pasture in the Eastern Amazon (Tapajos National Forest) and conducted a 30-day microcosm incubation experiment using four moisture levels. The experimental treatments simulated an increase in moisture levels during the rainy season as well as extreme rainfall events (floods), as this has been predicted to be more frequent in Amazon (HARPER et al., 2010; ARVOR et al., 2017). The 16S rRNA amplicon sequencing was used to assess the archaeal community structure and composition and analyzed the data using the integration between metataxonomy and environmental metadata (soil methane fluxes).

## 5.2. Material and Methods

The microcosm experiment described here also was used by VENTURINI (2019), which included the soil methane fluxes analyzed on a gas chromatograph.

### 5.2.1. Site description and soil sampling

The sites are located in the Belterra municipality, in the state of Pará, Brazil. The climate of the region is classified as Am (Köppen-Geiger classification), with an average annual air temperature of 26 °C and average annual precipitation of 2150 mm. The predominant soil type is Oxisol, with clay texture and low fertility. Soil sampling was carried out in July 2015 in the Tapajós National Forest (3°17'44.4"S, 54°57'46.7"W), a well-preserved primary forest with no evidence of logging, fire and other disturbances. A cattle pasture (3°18'46.7"S, 54°54'34.8"W) next to the forest, which was established more than 20 years ago, after the slash-and-burning of the natural vegetation with subsequent seeding of

fast-growing non-native grass *Urochloa brizantha* (Supplementary Figure 1). At each site, a 150 m transect was established with three equally spaced sampling points (50 m apart). First, the litter layer was removed, and then, soil samples were collected from 0-10 cm depth. These included (1) 500 g of loose soil for chemical properties, (2) 50 g of loose soil for molecular analysis, and (3) 2000 g of loose soil for the microcosm experiment. For each analysis, a total of 6 soil samples were collected in the field (2 sites × 3 sampling points per site). Then, soil samples were transported to the research facility on ice, where samples for molecular analysis were stored at -80 °C and samples for chemical analysis at 4°C.

### 5.2.2. Microcosm experiment and gas chromatography

For each site, three soil samples (2000 g of loose soil) were mixed to form one composite sample, totalling 6000 g of soil. Then, the soil was sieved through a 5 mm mesh to remove litter material prior to the microcosm experiment, which consisted of a 2 X 4 factorial design: 2 soils (forest and pasture) x 4 moisture levels (original moisture, determined as 22% for forest and 24% for pasture; and 60%, 80% and 100% of moisture at field capacity.). Each treatment was established in triplicate in 1.5 L jars filled with 350 g of soil. The jars were maintained for 30 days at 25 °C in a BOD incubator, in which the soil moisture of each jar was checked and corrected daily by weighing. Before testing for moisture content at 1, 2, 3, 6, 9, 12, 15, 18, 21, 24, 27 and 30 days after the start of the experiment, gas samples from each jar were collected with a syringe for 30 minutes (1, 10, 20 and 30 minutes after the jars were closed). The soil from each jar was frozen in liquid nitrogen and stored at -80 °C at the end of the experiment. The gas samples were analyzed on a SRI 8610c gas chromatograph (SRI Instruments, Torrance, CA, USA). Methane fluxes from each jar were calculated according to the change in the jar concentration over time. Based on these results, the total accumulated emissions were determined by the linear interpolation of the daily emission. As described by VENTURINI (2019), primary forest and pasture soils under 60% and 80% FC presented negative/or neutral emission values compared to their respective control/untreated soils throughout the experimental period, thus demonstrating that under these conditions these soils can act as methane sink (Supplementary Table 1).

### 5.2.3. DNA extraction, quantification and sequencing

Total DNA from each soil sample was extracted using PowerLyzer PowerSoil DNA Isolation Kit (MO Bio Laboratories, Carlsbad, CA, USA), according to the manufacturer's protocol, except at the initial stage after the addition of the C1 Solution, in which the samples were vortexed for 15 minutes at maximum speed and centrifuged for 3 minutes at 10,000 x g. The quality and quantity of the DNA samples were evaluated using agarose gel electrophoresis and Nanodrop 2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The V4 region of the archaeal 16S rRNA gene was amplified using the primer pair 519F and 915R (COOLEN et al., 2004); the resulting PCR products were sequenced with Illumina HiSeq 2500 platform (2 x 250 bp) (Illumina, San Diego, CA, USA) at Novogene Corporation, Beijing, China.

### 5.2.4. Bioinformatics analysis

Bioinformatics analyses were performed using Divisive Amplicon Denoising Algorithm (DADA2) version 1.8 pipeline (CALLAHAN et al., 2016a) to remove low-quality sequences, model and correct Illumina-sequencing errors, merge paired-end reads, identify and quantify ASVs (Amplicon Sequence Variants). The paired-end reads were truncated to 240 bp (forward) and 160 bp (reverse), and reads shorter than the minimum length were discarded. The low-quality sequences were removed that presented the maximum number of expected errors (maxEE) equal or greater than 2, and after truncation, sequences with that contained Ns were also discarded. To estimate the error rates, a parametric error model was used after de-replicating the data using a core sample inference algorithm. Finally, the complete denoised data were obtained after merging the forward and reverse reads with an overlap criterion in which at least 12 bases were identical in the overlapping region. The taxonomy was predicted using the naive Bayesian classifier method (WANG et al., 2007b), where the sequences were compared against a trained dataset based on the Silva database (v. 132) (PRUESSE et al., 2007). Sequences assigned as Bacteria were removed from the dataset. Multivariate statistics based on Detrended Correspondence Analysis (DCA) were calculated using phyloseq R package (MCMURDIE; HOLMES, 2013). Similar to the statistics on methane fluxes, the Aligned-rank transformation (ART) was performed on the abundance for each taxonomic level of the Archaea. To integrate the phylogenetic and the functional dimensions, and to identify which ASVs were correlated with the methane emission,

linear Pearson correlations were estimated with a threshold value of 0.7 and p-value ≤ 0.05 (R Core Team) based on the relative abundance values. To better visualize the changes in the relative abundance on a heatmap log transformations (log(x+1)) were applied.

## 5.3. Results

### 5.3.1. Community structure and composition of archaea

A total of about 2.5 million 16S rRNA sequences were obtained, with an average of 45,210 ± 24,368 sequences per sample. After the sequence processing, 108 archaeal ASVs were identified.

In the microcosm experiment, the archaeal communities in forest and pasture soils were mainly influenced by long-term (land-use change; 73.8% of variation) followed by short-term (moisture level; 15.3% of variation) changes (Figure 1). Although the increased moisture levels altered the archaeal community structure in both soils, it affected pasture soils more than the forest soils. In addition, the field samples were highly similarity to the control treatments in the microcosm experiment. There was an increase in beta diversity in pasture soils, as shown by a greater dispersion and distance between the clusters of each treatment.

The three most abundant archaeal phyla were Thaumarchaeota (97.7±3.4% of the total sequences), followed by Bathyarchaeota (1.2±2.5%) and Euryarchaeota (1.1±0.9%) (Figure 2). At a deeper taxonomic level, Thaumarchaeota (Nitrososphaerales, Nitrosotalestes and Group 1.1c) and Euryarchaeota (Methanobacteriales, Methanocellales, Methanosarcinales and Methanomassiliicoccales) were affected by the interactions between their land-use change and increase in moisture of the soils ($p \leq 0.05$) (Table 1). Similar results were observed for the Bathyarchaeota phylum as well (Table 1).

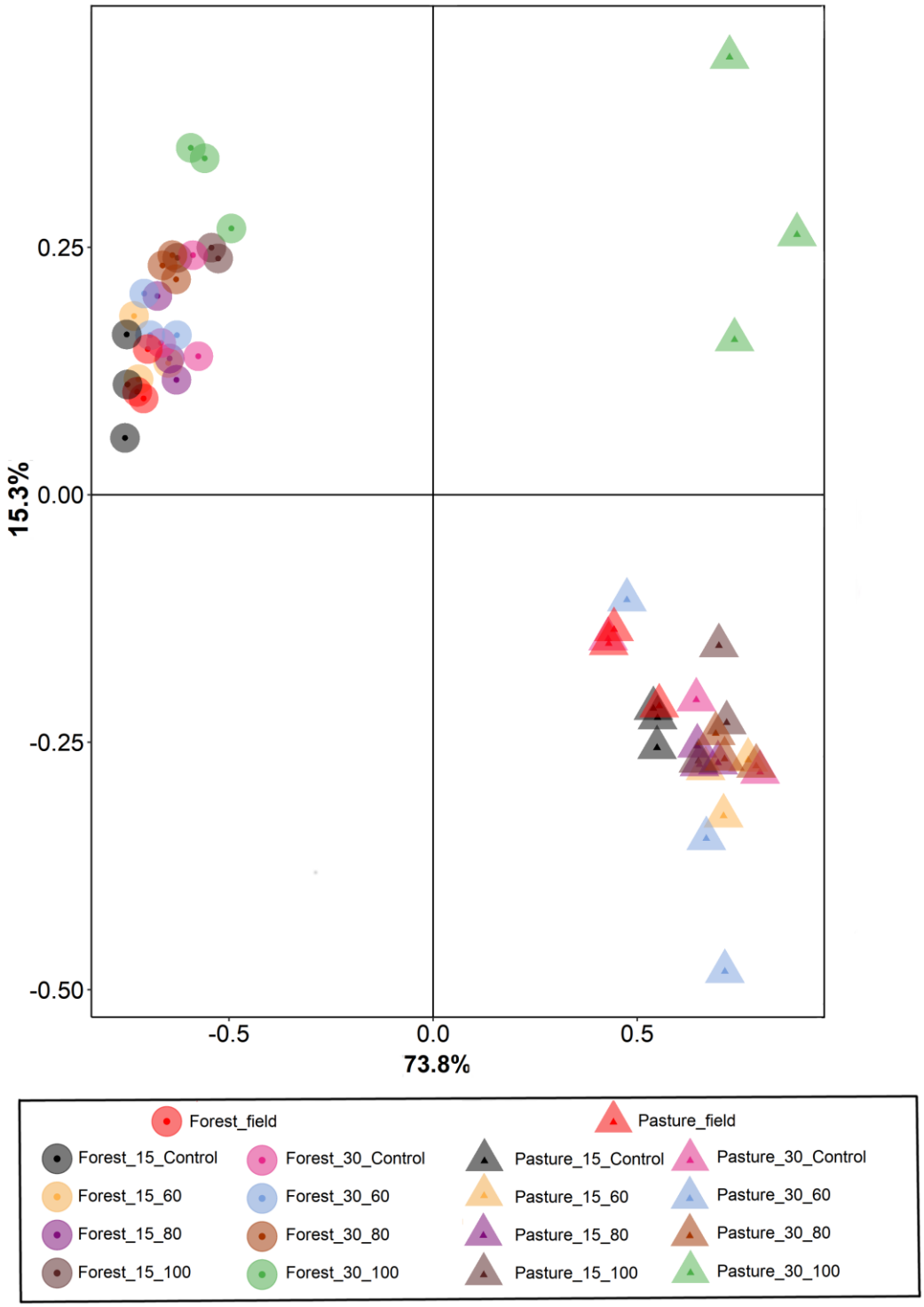



**Figure 1. Detrended correspondence analysis for the archaeal communities in forest and pasture soils from Eastern Amazon under different moisture levels (control, 60, 80, and 100% of field capacity) after 30 days of experiment**. Each colour represents one specific treatment based on the moisture level; while each shape represents the land-use

Table 1. Aligned-rank transformation (ART) for non-parametric ANOVA of the relative abundance of archaea group as a function of land-use (forest-to-pasture conversion) and soil moisture levels, along with interaction

| Archaea groups | Land-use | | Moisture levels | | Land-use x Moisture levels | |
|---|---|---|---|---|---|---|
| | F | *p* | F | *p* | F | *p* |
| Thaumarchaeota | | | | | | |
| Group 1.1c | 3.7 | 0.06 | 9.2 | **≤0.0005** | 7.3 | **≤0.005** |
| Nitrososphaerales | 11.0 | **≤0.005** | 0.4 | 0.69 | 9.3 | **≤0.0005** |
| Nitrosotaleales | 40.9 | **≤0.005** | 1.7 | 0.16 | 4.1 | **≤0.05** |
| Bathyarchaeota | | | | | | |
| Bathyarchaeota | 28.6 | **≤0.0005** | 4.5 | **≤0.005** | 9.9 | **≤0.0005** |
| Euryarchaeota | | | | | | |
| Methanobacteriales | 104.8 | **≤0.0005** | 7.2 | **≤0.0005** | 11.7 | **≤0.0005** |
| Methanocellales | 20.6 | **≤0.0005** | 9.1 | **≤0.0005** | 13.9 | **≤0.0005** |
| Methanosarcinales | 14.3 | **≤0.0005** | 5.8 | **≤0.005** | 3.7 | **≤0.05** |
| Methanomassicoccales | 27.2 | **≤0.0005** | 4.2 | **≤0.05** | 2.7 | **≤0.05** |

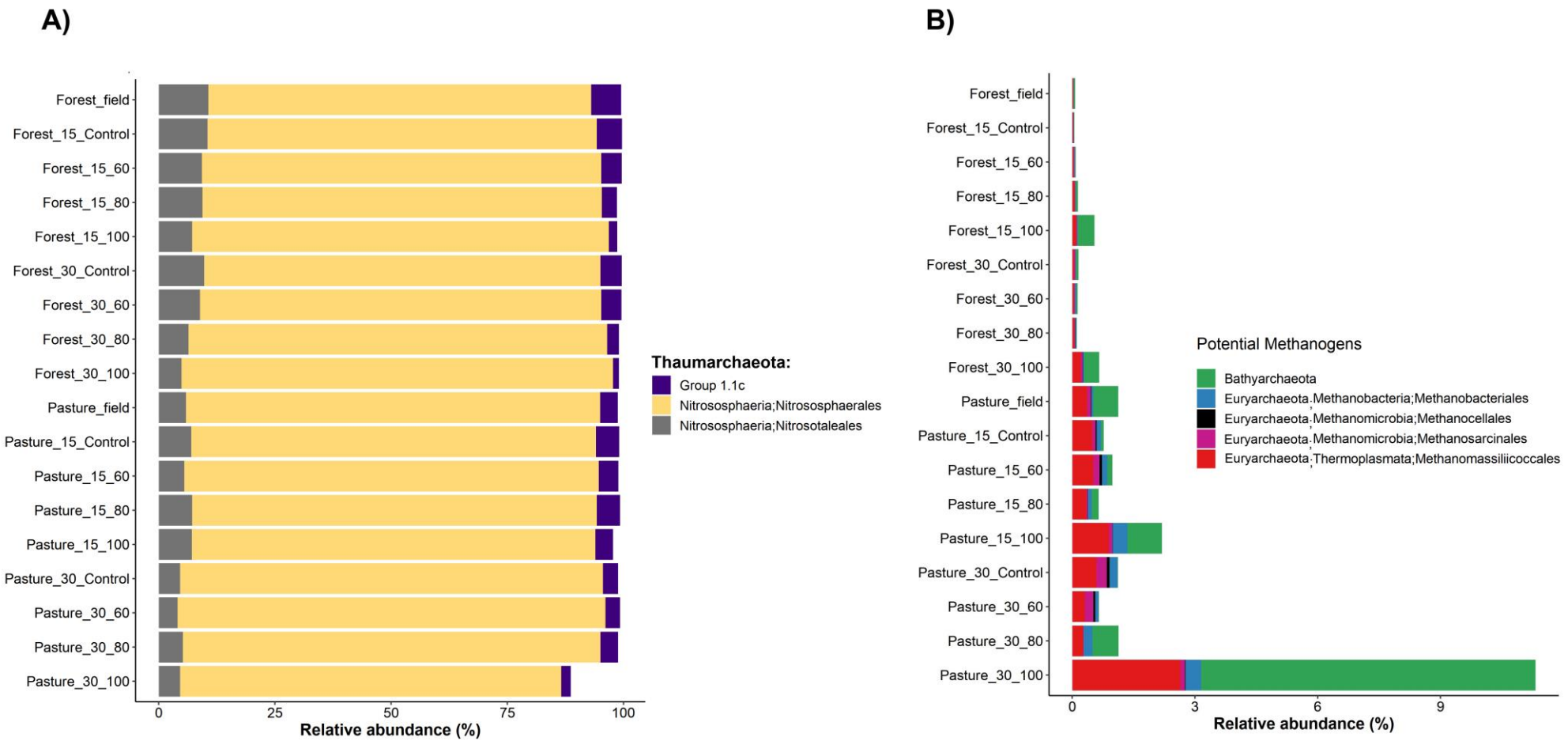



**Figure 2. Relative abundance of Archaea in forest and pasture soils from Eastern Amazon under different moisture levels (control, 60, 80, and 100% of field capacity) after 15$^{th}$ and 30$^{th}$ days of microcosm experiment.** The three most abundant Archaea phylum: Thaumarchaeota – non-methanogens (A) and Potential methanogens taxa (B).

The Thaumarchaeota groups responded differently to the changes in soil land-use and their moisture contents (Figure 2A). For example, the relative abundance of Nitrososphareles at 30-day time point from forest soils with 100% FC increased (91.9±0.01%) compared to control (86±1.0%), whereas it decreased in pasture soils (control: 91.1±1.3% to 100% FC: 82.6±0.02%). However, the reverse was true for the Nitrosotaleales group, where the relative abundance was higher in forest (7.8±1.2%) than pasture (4.4±0.1%), and the effects of the moisture (100% of FC) was stronger in forest (5.3±0.6%) than pasture (4.6±0.8%). Even though the interaction between the two factors (land-use and moisture content) was significant ($p \leq 0.05$) in affecting the relative abundance of the Nitrosotaleales, the moisture content alone was not significant in affecting their composition which can be seen by the similar relative abundance in the control and 100% FC treatment in pasture soil. Also, the Group 1.1c was more abundant in forest (4.19±0.9%) than pasture (3.6±1.6%), and the changes in soil moisture content reduced their relative abundance in both forest (1.3±0.1%) and pasture (2.3±1.0%) soils (100% FC).

The potential methanogenic groups Bathyarchaeota and Euryarchaeota (Methanobacteriales, Methanocellales, Methanosarcinales and Methanomassiliicoccales) were also affected by both soil land-use change and increase in their moisture content (Figure 2B). The same pattern was observed for both time points of the experiment ($15_{th}$ and $30_{th}$ day), but unlike the forest, the pasture was more influenced by the time. Stronger effect was observed as a result of the interaction of the factors (land-use and moisture content) with the increase in relative abundances of Bathyarchaeota and Methanomassiliicoccales. Bathyarchaeota increased seven times more in pasture (0.01±0.02 to 7.2±2.2%) than in forest (0.06±0.02 to 0.28±0.08%) under 100% FC on the $30_{th}$ day. The same pattern was observed for Methanomassiliicoccales and Methanobacteriales. However, the Methanocellales and Methanosarcinales decreased in pasture with 100% of FC. It is also to be noted that the overall incubation time period had affected the archaeal community structure more in the pasture than the forest soils (Figure 2A and 2B).

### 5.3.2. Archaeal communities and methane emission

We analyzed the correlation between the relative abundance of the ASVs and methane emissions in both soils. Seven ASVs were positively correlated with methane emissions, being four in pasture and three in both soils (Figure 3). All of them belong to either Euryarchaeota (Thermoplasmata; Methanomassiliicoccales) or Bathyarchaeota phyla. ASV_45, ASV_32, and ASV_36 were associated with methane emissions in both forest and pasture soils. ASV_45 was assigned to the potential methanogenic phyla Bathyarchaeota and ASV_31 and ASV_36 to the methanogenic Euryarchaeota *Methanomassiliicossus* (Figure 3). The best BLAST identity (100%) for ASV_45 was a nucleotide sequence characterized in a stable-isotope probing (SIP-DNA) study based on the methanogenic communities (NCBI Id: AJ879013) in the rice rhizosphere (LU; CONRAD, 2005).

Soil specific associations between the increase in the methane emissions and ASVs were also identified in pasture, but not in forest. For example, ASV_25, ASV_41, ASV_84, assigned to Bathyarchaeota, and ASV_69, assigned to *Methanomassiliicoccus*, were significantly correlated to $CH_4$ emissions only in pasture (Figure 3).

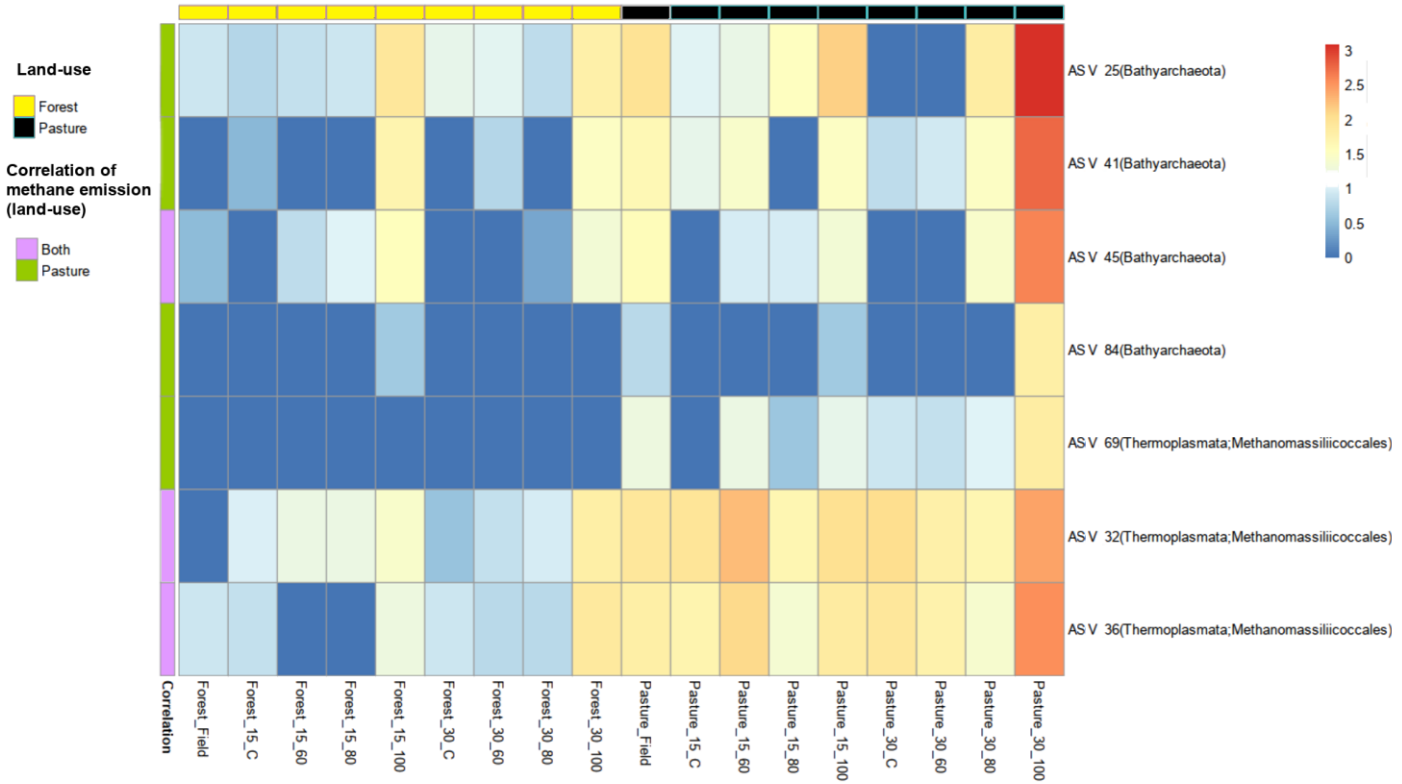



**Figure 3. Abundance and correlation between *Amplicon Sequence Variants* (ASV) and methane emissions in forest and pasture soils from Eastern Amazon under different moisture levels (control, 60, 80, and 100% of field capacity) after 30 days of experiment.** The relative abundance and taxonomy assignment of each ASV, which was correlated with the methane increase, was log+1 transformed to better visualization the values. The correlation column indicates for which land-use each ASV was correlated ($r \geq 0.7$ and $p \leq 0.05$) with the increase in methane emission

## 5.3. Discussion

Our data demonstrated that archaea community dynamics were altered with the long-term (forest-to-pasture conversion) and short-term (increase in soil moisture level) changes in the Amazon soils. The data showed that archaeal community in pasture was more responsive to the increase of soil moisture levels than forest soil. We hypothesize a scenario where forest soils may buffer the archaeal communities and influence their resistance against the increase of soil moisture levels. In the dry season (equivalent to ~ 60 % FC), the precipitation and humidity are reduced, but the forest soils were able to withdraw water from deep layers, maintaining the soil moisture content than pasture soils (VON RANDOW et al., 2004; JUÁREZ et al., 2007). In addition, the soil moisture content is also controlled by the increase of litterfall during the dry season in Amazon rainforest (CAMARGO et al., 2015; XU et al., 2013). The forest soils may select microorganisms more adapted to survive the rapid increase in moisture content during the wet season, as suggested by Evans and Wallenstein (2014), who also showed that the environmental history can influence changes in microbial communities. These archaeal groups may also have adaptive mechanisms to tolerate other physiological stresses (e.g., osmolyte accumulation, control of ion flux and change in protein expression) caused by the rapid change in water potential (MARTIN; CIULLIA; ROBERTS, 1999).

As already described by Venturini (2019), using the same microcosm experiment used here, the forest-to-pasture conversion and the increase in soil moisture level affect the abiotic factors of the soil, especially their methane fluxes. Similar to other studies performed in this region, it was observed that forest-to-pasture conversion lead to an increase in soil pH and nutrients availability, mainly by the addition of large amount of ashes derived slash and burn events of the original vegetation (FERNANDES et al., 2002; KROEGER et al., 2018; PEDRINHO et al., 2018). Furthermore, studies suggest that forest-to-pasture conversion may also alter soil physical properties (i.e. soil texture, macro- and microporosity, density and water content) and consequently affect microbial communities and further important processes like including ammonia oxidation and methane production (SONG et al., 2003; KELLER; REINERS, 1994; KELLER et al., 2005). Some studies have demonstrated that forest soils are an important sink of methane under low moisture content (~60% FC) and crucial for minimizing the impacts of global climate change (MUTSCHLECHNER et al., 2018). However, the increase in moisture content in these soils (equivalent to intensive

rainfalls), increased the production of methane. It is important also to note that the pasture soils generally presented the highest emission values

The low resistance to the increase of moisture content in pasture soils indicated that the stability of this community was reduced, changing the relative abundance of Thaumarchaeota (Nitrososphaerales, Nitrosotalestes and Group 1.1c), Euryarchaeota (Methanobacteriales, Methanocellales, Methanosarcinales and Methanomassiliicoccales) and Bathyarchaeota, resulting in higher methane emissions. Our data indicated that Thaumarchaeota was the most abundant archaeal group in all samples. When we increased the water content (from 60% to 100% FC), the Thaumarchaeota groups decreased in 10% in pasture soils. In temperate forest soils, Szukics et al. (2012) described that the abundance of ammonia-oxidizing archaea and the *amo*A gene are dependent on soil properties. Similar to the results described here, Szukics and collaborators (2012) identified that the increased of soil moisture provided a non-optimal condition for Thaumarchaeota survival, suggesting a sensitivity to anaerobic conditions. In this case, the low dissolved oxygen may control the relative abundance of ammonia-oxidizing archaea (ERGUDER et al., 2009) and the rates of nitrification (DONG et al., 2011).

The relative abundance of the phyla Euryarchaeota and Bathyarchaeota were higher in pasture soils compared to forest soils and the changes were further intensified by the increase in soil moisture content (e.g., 100% of the field capacity). Euryarchaeota (MORAN et al., 2005) and their role in methane production are very well studied in soils (ANGEL; CLAUS; CONRAD, 2012). Furthermore, Meyer et al. (2017) also observed shifts in the methanogenic communities after the deforestation of the Amazon forest. Bathyarchaeota, a new proposed archaeal phylum with a potential role in methanogenesis (EVANS et al., 2015) and acetogenesis (HE et al., 2016), had not yet been identified in Amazon soils. Bathyarchaeota and Euryarchaeota have a versatile metabolism in low-oxygen environments, and theirs high abundance in pasture soil under 100% of the field capacity may be associated with changes in soil properties (e.g., soil compaction) after the forest-to-pasture conversion. These changes may have been caused as a consequence of the poor grazing management and subsequent compaction of soil (BRAZ et al., 2013), since soil compaction alters the structure and size of soil pores occupied by water (TORBERT; WOOD, 1992), limiting the air and water conductivity (HARTMANN et al., 2014). In this case, continuous cow grazing and trampling on pastures reduces the macropore space, decreasing soil oxygenation and increasing methane emissions. A similar pattern was described by Frey and collaborators (FREY et al., 2009), in a study about the effect of heavy-machinery traffic on the abundance of methanogens in

oxic forest soils, and Radl (2007) described an increase in the soil methane production stimulated by cattle grazing in grasslands.

We also identified seven potential ASVs associated with methane emission in Amazon soils. For the first time, high-abundant Bathyarchaeota were identified in Amazon soils under wet conditions, which have the potential to emit methane. One Bathyarchaeota ASV, which had 100% sequence identity to an unculturable methanogenic archaea from anoxic rice soils (LU; CONRAD, 2005), increased its abundance in pasture soils at 100% of the field capacity, indicating a potential methanogenic role. One Euryarchaeota ASV also presented the same pattern of identity and abundance, but we were able to taxonomically assign it only at the genus level (*Methanomassiliicoccus*). There are two sequenced genomes for this genus: *Candidatus Methanomassiliicoccus intestinalis* (BORREL et al., 2013) and *Methanomassiliicoccus luminyensis* (GORLAS et al., 2012) and they were isolated from the human microbiome but are also identified in paddy soils (REIM et al., 2017) and wetlands (SÖLLINGER et al., 2015).

Our results reveal a potential emerging contribution of Bathyarchaeota and *Methanomassiliicoccus* to the methane emission, as consequence of the forest-to-pasture conversion and the increase in soil moisture levels. The methylotrophic methanogenesis could be started with the use of methanol, an important methylated compound, which is produced by the degradation of pectin (WARNEKE et al., 1999; SCHINK et al., 1980) and is very common in low-oxygened soils and sediments (SCHINK; ZEIKUS, 1982). This substrate may be used by Bathyarchaeota and *Methanomassillicoccus*, as already reported by Evans and collaborators (2015), during the *Brachiaria* (pasture) degradation under floodplain soil conditions. The results described here indicated only an association between the relative abundance of these groups and the methane emission, and not the specific fraction that was produced for each taxonomic group. Furthermore, we do not discard the potential metabolic activity of Methanobacteriales, Methanocellales and Methanosarcinales, once they also were identified in the treatments which had more methane emission. New experimental assays, such as stable-isotope probing (SIP) (HUNGATE et al., 2015), new investigations about the metabolism of methanogens by metagenome-assembled genomes studies (EVANS et al., 2015), and possibly metatranscriptomics are necessary to validate the role of methylotrophic methane metabolism in amazon soils.

The results reported here suggest that (I) the effect of forest-to-pasture conversion on soil microbial communities were intensified when the moisture levels were increased, affecting the archaeal community structure; (II) the archaeal communities from forest were

more resistant to the increase of the soil moisture levels, while the communities from pasture were more sensitive, enhancing the potential of methanogenesis in this soil. Furthermore, with the intensification of forest-to-pasture conversion in the Amazon region and the possible prolongation of the wet season as a result of climate change, may result in more methane production in the future, thus altering its global biogeochemical cycle. In this sense, a better understanding of the impacts of forest-to-pasture conversion on archaeal groups can help the development of a more sustainable management strategy, aiming to reduce methane emissions.

**References**

ANGEL, R; CLAUS, P; CONRAD, R. Methanogenic archaea are globally ubiquitous in aerated soils and become active under wet anoxic conditions. **The ISME Journal**, London, v. 6, p. 847–862, 2012.

ARVOR, D. et al. Monitoring Rainfall Patterns in the Southern Amazon with PERSIANN-CDR Data: Long-Term Characteristics and Trends. **Remote Sensing**, Basel, v. 9, p. 1-20, 2017.

BEAM, J. P. et al. Niche specialization of novel Thaumarchaeota to oxic and hypoxic acidic geothermal springs of Yellowstone National Park. **The ISME Journal**, London, v. 8, n. 4, p. 938–951, 2014.

BORREL, G. et al. Phylogenomic Data Support a Seventh Order of Methylotrophic Methanogens and Provide Insights into the Evolution of Methanogenesis. **Genome Biology & Evolution**, Oxford, v. 5, p 1769–1780, 2013.

BREWER, P.E. et al. Impacts of moisture, soil respiration, and agricultural practices on methanogenesis in upland soils as measured with stable isotope pool dilution. **Soil Biology and Biochemistry**, Oxford, v. 127, p. 239-251, 2018.

BRAZ, A. M et al. Soil Attributes After the Conversion from Forest to Pasture in Amazon. **Land Degradation & Development**, New Jersey, v. 24, p. 33-38, 2013.

CALLAHAN, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. **Nature Methods**, London, v. 13, n. 7, p. 581–583, 2016a.

CAMARGO, M.; GIARRIZZO, T.; JESUS, A. J. S. Effect of seasonal flooding cycle on litterfall production in alluvial rainforest on the middle Xingu River (Amazon basin, Brazil). **Brazilian Journal of Biology**, São Carlos, v. 75, p 250-256, 2015.

DI, D.I. et al. (2014). Effect of soil moisture status and a nitrification inhibitor, dicyandiamide, on ammonia oxidizer and denitrifier growth and nitrous oxide emissions in a grassland soil. **Soil Biology and Biochemistry**, Oxford, v. 7, p. 59-68, 2014.

ERGUDER, T. et al. Environmental factors shaping the ecological niches of ammonia-oxidizing archaea. **FEMS Microbiology Ecology**, Amsterdam, v. 33, p. 855-869, 2009.

EVANS, S.; WALLENSTEIN, M. D. Climate change alters ecological strategies of soil bacteria. **Ecology Letters**, New Jersey, v. 17, p. 155-164, 2014.

EVANS, P.N. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. **Science**, New York, v. 350, p. 434–438, 2015.

FERNANDES, S. A. P. et al. Seasonal variation of soil chemical properties and CO2 and CH4 fluxes in unfertilized and P-fertilized pastures in an Ultisol of the Brazilian Amazon. **Geoderma**, Amsterdam, v. 107, p. 227–241, 2002.

FREY, B. et al. Compaction of forest soils with heavy logging machinery affects soil bacterial community structure. **European Journal of Soil Biology**, Amsterdam, v. 45, n. 4, p. 312-320, 2009.

GLOOR, M. et al Recent Amazon climate as background for possible ongoing and future changes of Amazon humid forests. **Global Biogeochemestry Cycles**, New Jersey, v. 29, p. 1384–1399, 2015.

GORLAS, A. ET AL. Complete genome sequence of Methanomassiliicoccus luminyensis, the largest genome of a human-associated Archaea species. **Journal of Bacteriology**, Washington, DC, v. 194, p. 4745, 2012.

HAMAOUI, G.S. et al. Land-use change drives abundance and community structure alterations of thaumarchaeal ammonia oxidizers in tropical rainforest soils in Rondônia, Brazil. **Applied Soil Ecology**, Amsterdam, v. 107, p. 48–56, 2016.

HARPER, A.B. et al, Role of deep soil moisture in modulating climate in the Amazon rainforest. **Geophysical Research Letters**, New Jersey, v. 37, 2010.

HARTMANN, A. et al. Karst water resources in a changing world: Review of hydrological modeling approaches. **Reviews of Geophysics**, New Jersey, v. 52, p. 218-242, 2014.

HEDDERICH, R.; WHITMAN, W.B. Physiology and biochemistry of the methane-producing archaea. In: ROSEMBERG, E. et al. (Ed.). **The Prokaryotes**", Berlin: Spring-Verlag, 2013. p. 1050-1079.

HUNGATE, B. A. et al. Quantitative Microbial Ecology through Stable Isotope Probing. **Applied and Environmental Microbiology**, Baltimore, v. 81, n. 21, p. 7570-7581, 2015.

JABLÓNSKI, S; RODOWICZ, P; LUKASZEWICZ, M. Methanogenic archaea database containing physiological and biochemical characteristics. **International Journal of Systematic and Evolutionary Microbiology**, London, v. 65, p. 1360-1368, 2015.

JUÁREZ, R. I. N. Control of Dry Season Evapotranspiration over the Amazonian Forest as Inferred from Observations at a Southern Amazon Forest Site. **Journal of Climate**, Washington, DC, v. 20, p. 2827-2839, 2007.

KELLER et al. Soil–Atmosphere Exchange of Nitrous Oxide, Nitric Oxide, Methane, and Carbon Dioxide in Logged and Undisturbed Forest in the Tapajos National Forest, Brazil. **Earth Interactions**, Washington, DC, v. 9, n. 23, p. 1-28, 2005.

KELLER, M; REINERS, W.A. Soil-atmosphere exchange of nitrous oxide, nitric oxide, and methane under secondary succession of pasture to forest in the Atlantic lowlands of Costa Rica. **Global Biogeochemical Cycles**, New Jersey, v. 8, p. 399–409, 1994.

KEROU, M. et al. Proteomics and comparative genomics of Nitrososphaera viennensis reveal the core genome and adaptations of archaeal ammonia oxidizers. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 113, n. 49, p. E7937–E7946, 2016.

KROEGER, M. E. et al. New biological insights into how deforestation in Amazonia affects soil microbial communities using metagenomics and metagenome-assembled genomes. **Frontiers in Microbiology**, Lausanne, v. 9, p. 1635, 2018. doi: 10.3389/fmicb.2018.01635.

LEMOS, L. N. et al. Genome-Centric Analysis of a Thermophilic and Cellulolytic Bacterial Consortium Derived from Composting. **Frontiers in Microbiology**, Lausanne, v. 8, p. 1-16, 2017.

LIN, X. et al. Metabolic potential of fatty acid oxidation and anaerobic respiration by abundant members of Thaumarchaeota and Thermoplasmata in deep anoxic peat. **The ISME Journal**, London, v. 9, n. 12, p. 2740–2744, 2015.

LIU, Y.; WHITMAN, W. Metabolic, Phylogenetic, and Ecological Diversity of the Methanogenic Archaea. **Annals of the New York Academy of Sciences**, New Jersey, v. 1125, n. 1, p. 171-189, 2008.

LU, Y; CONRAD, R. In situ stable isotope probing of methanogenic archaea in the rice rhizosphere. **Science**, New York, v. 309, p. 1088-1090, 2005.

McMURDIE, P.J. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. **PLoS One**, San Francisco, v. 8, p. e61217, 2013.

MALHI, Y. et al. Climate Change, Deforestation, and the Fate of the Amazon. **Science**, New York, v. 319, n. 169-172, 2008.

MARTIN, D; CIULLA, R.A; ROBERTS, M.F. Osmoadaptation in archaea. **Applied and Environmental Microbiology**, Washington, DC, v. 65, p. 1815-1825, 1999.

MENDES, L.W. Soil-borne microbiome: linking diversity to function. **Microbial Ecology**, Amsterdam, v. 70, p. 255-265, 2015.

MEYER, K. M. et al. Conversion of Amazon rainforest to agriculture alters community traits of methane-cycling organisms. **Molecular Ecology**, Amsterdam, v. 26, n. 6, p. 1547–1556, 2017.

MORAN, J. J. et al. Trace methane oxidation studied in several Euryarchaeota under diverse conditions. **Archaea**, Vancouver, v. 1, n. 5, p. 303–309, 2005.

MUELLER, R. C. et al. Links between plant and fungal communities across a deforestation chronosequence in the Amazon rainforest. **The ISME Journal**, London, v. 8, n. 7, p. 1548–1550, 2014.

MUTSCHLECHNER, M.; PRAEG, N.; ILLMER, P. The influence of cattle grazing on methane fluxes and engaged microbial communities in alpine forest soils. **FEMS Microbiology Ecology**, Amsterdam, v. 94, n. 5, 2018.

NAVARRETE, A. A. et al. Land-use systems affect Archaeal community structure and functional diversity in western Amazon soils. **Revista Brasileira de Ciência do Solo**, Viçosa, v. 35, n. 5, p. 1527–1540, 2011.

NAVARRETE, A. A. et al. Differential Response of Acidobacteria Subgroups to Forest-to-Pasture Conversion and Their Biogeographic Patterns in the Western Brazilian Amazon. **Frontiers in Microbiology**, v. 6, 2015.

OFFRE, P.; SPANG, A.; SCHLEPER, C. Archaea in biogeochemical cycles. **Annual Review of Microbiology**, Palo Alto, v. 67, p. 437–457, 2013.

PEDRINHO, A. et al. Forest-to-pasture conversion and recovery based on assessment of microbial communities in Eastern Amazon rainforest. **FEMS Microbiology Ecology**, Amsterdam, v. 95, n. 3, 2019.

PREM, E. M.; REITSCHULER, C.; ILLMER, P. Livestock grazing on alpine soils causes changes in abiotic and biotic soil properties and thus in abundance and activity of microorganisms engaged in the methane cycle. **European Journal of Soil Biology**, Amsterdam, v. 62, p. 22–29, 2014.

PRUESSE, E. et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. **Nucleic Acids Research**, Oxford, v. 35, n. 21, p. 7188–7196, 2007.

RADL, V. et al. Effects of cattle husbandry on abundance and activity of methanogenic archaea in upland soils. **The ISME Journal**, London, v. 1, n. 5, p. 443–452, 2007.

RANJAN, K. et al. Forest-to-pasture conversion increases the diversity of the phylum Verrucomicrobia in Amazon rainforest soils. **Frontiers in Microbiology**, Lausanne, v. 6, 2015.

REIM, A. et al. Response of Methanogenic Microbial Communities to Desiccation Stress in Flooded and Rain-Fed Paddy Soil from Thailand. **Frontiers in Microbiology**, Lausanne, v. 8, 2017.

SCHINK, B.; ZEIKUS, J. G. Microbial methanol formation: A major end product of pectin metabolism. **Current Microbiology**, London, v. 4, n. 6, p. 387–389, 1980.

SHI, Y. et al. The biogeography of soil archaeal communities on the eastern Tibetan Plateau. **Scientific Reports**, London, v. 6, p. 38893, 2016.

SÖLLINGER, A. et al. Phylogenetic and genomic analysis of Methanomassiliicoccales in wetlands and animal intestinal tracts reveals clade-specific habitat preferences. **FEMS Microbiology Ecology**, Amsterdam, v. 92, p. 1-11, 2015.

SONG, J. et al. Novel use of soil moisture samplers for studies on anaerobic ammonium fluxes across lake sediment-water interfaces. **Chemosphere**, Oxford, v. 50, n. 6, p. 711–715, 2003.

SPRENGER, W. et al. Methanomicrococcus blatticola gen. nov., sp. nov., a methanol- and methylamine-reducing methanogen from the hindgut of the cockroach Periplaneta americana. **International Journal of Systematic Microbiology**, London, v. 50, p. 1989-1999, 2000.

STEUDLER, P. A. et al. Consequence of forest-to-pasture conversion on CH4 fluxes in the Brazilian Amazon Basin. **Journal of Geophysical Research: Atmospheres**, New Jersey, v. 101, n. D13, p. 18547–18554, 1996.

SZUKICS, U. et al. Rapid and dissimilar response of ammonia oxidizing archaea and bacteria to nitrogen and water amendment in two temperate forest soils. **Microbiological Research**, Amsterdam, v. 167, n. 2, p. 103–109, 2012.

TORBERT, H. A.; WOOD, C. W. Effects of soil compaction and water- filled pore space on soil microbial activity and N losses. **Communications in Soil Science and Plant Analysis**, New York, v. 23, n. 11–12, p. 1321–1331, 1992.

TRIPATHI, B. M. et al. pH dominates variation in tropical soil archaeal diversity and community structure. **FEMS Microbiology Ecology**, Amsterdam, v. 86, n. 2, p. 303–311, 2013.

TOURNA, M. et al. Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 108, n. 20, p. 8420–8425, 2011.

TUPINAMBÁ, D. D. et al. Archaeal Community Changes Associated with Cultivation of Amazon Forest Soil with Oil Palm. **Archaea**, Vancouver, v. 2016, p. 3762159, 2016.

VANWONTERGHEM, I. et al. Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. **Nature Microbiology**, London, v. 1, n. 12, p. 16170, 2016.

VENTURINI, A. M. **Conversão floresta-pastagem na Amazônia Oriental: impactos sobre as comunidades microbianas do metano do solo**. 2019. Tese (Doutorado em Ciências) – Centro de Energia Nuclear na Agricultura, Universidade São Paulo, Piracicaba, 2019.

VERCHOT, A.; L.V. Impacts of forest conversion to agriculture on microbial communities and microbial function. In: **Soil biology and agriculture in the tropics**. London Springer, 2010.

VON RANDOW, C. et al. Comparative measurements and seasonal variations in energy and carbon exchange over forest and pasture in South West Amazonia. **Theoretical and Applied Climatology**, London, v. 78, n. 1, p. 5–26, 1 jun. 2004.

WANG, J. et al. Impact of deforestation in the Amazon basin on cloud climatology. **Proceedings of the National Academy of Sciences of the USA**, Washington, DC, v. 106, n. 10, p. 3670–3674, 10 mar. 2009a.

WANG, Q. et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. **Applied and Environmental Microbiology**, Washington, DC, v. 73, n. 16, p. 5261–5267, 2007b.

WARNEKE, C. et al. Acetone, methanol, and other partially oxidized volatile organic emissions from dead plant matter by abiological processes: Significance for atmospheric HOx chemistry. **Global Biogeochemical Cycles**, New Jersey, v. 13, n. 1, p. 9–17, 1999.

WEBER, E. B. et al. Ammonia oxidation is not required for growth of Group 1.1c soil Thaumarchaeota. **FEMS Microbiology Ecology**, Amsterdam, v. 91, n. 3, 2015.

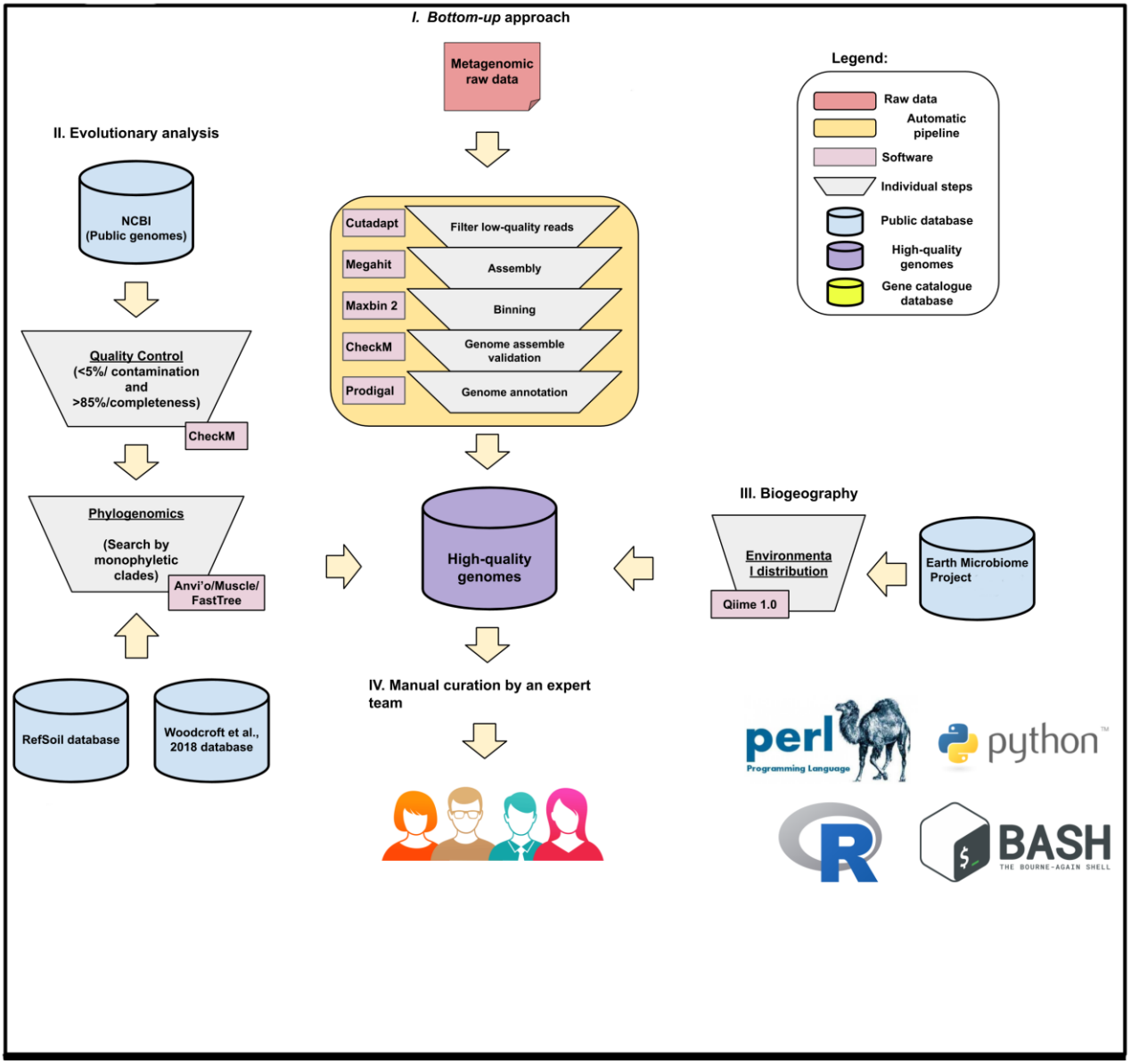
WILSON, C. et al. Contribution of regional sources to atmospheric methane over the Amazon Basin in 2010 and 2011. **Global Biogeochemical Cycles**, New Jersey, v. 30, n. 3, p. 400–420, 2016.

WOBBROCK, J. et al. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. In**:** ACM Conference on Human Factors in Computing Systems (CHI ’11), 11., 2011, Vancouver. **Proceedings…** New York: ACM Press, 2011. p. 143-146.

XU, S.; LIU, L. L.; SAYER, E. J. Variability of above-ground litter inputs alters soil physicochemical and biological processes: a meta-analysis of litterfall-manipulation experiments. **Biogeosciences**, Göttingen, v. 10, n. 11, p. 7423–7433, 2013.

# APPENDIX

## Appendix A. Supplementary Material of Chapter 3




Supplementary Figure 1. Integrative computational approach applied to study the microbial functional traits in metagenomes datasets. Bottom-up approach to study the evolution, potential metabolism and distribution of the non-ammonia oxidizing Thaumarchaeota (Chapter 3) and to explore the genome size traits of the soil CPR/Patescibacteria bacteria phyla (Chapter 4). The metagenomic raw data (amazon soil metagenomes) used to reconstruct the Metagenome-Assembled Genomes (MAGs) in the Study 1 and Study 2 were the same. (I) Bottom-up approach to reconstruct MAGs from Amazon soil metagenomes. (II). Evolutionary analysis methods used to discovery new microbial genomes. These informations were updated in 'High quality genomes' database. (III) Biogeography distribution of specific Archaea/Bacteria groups using the Earth Microbiome Project database (IV) Manual curation by an expert team to check specific information about the evolution, potential metabolic pathways and environmental distribution. The manual curation of traits was very important step of this proposal, because was necessary a multidisciplinary team to interpret the amount of biological complexity and information generated in previously steps.

Supplementary Table 1. General information about the metagenomic Amazon floodplain forest dataset used in this study to reconstruct individual microbial genomes.

| Parameter | |
|---|---|
| Sample description | Amazon floodplain forest sediment collected during the wet season |
| Number of paired-end reads | 156,056,159 |
| N50 (bp) | 600 |
| Longest contig (bp) | 248,200 |
| Total assembled length (bp) | 6,164,645 |
| Number of assembled contigs > 10,000 bp | 3,113 |
| Total assembled contigs length > 10,000 bp | 28,923,227 |

Supplementary Table 2. Potential non-AOA Thaumarchaeota genomes deposited on NCBI and JGI databases.

| DDBJ/ENA/GenBank Accession | Completeness (%) | Contamination (%) | Reference |
|---|---|---|---|
| **GCA_002494485** | **86.83** | **0.97** | **PARKS et al., 2017** |
| GCA_002494565 | 62.86 | 0.00 | PARKS et al., 2017 |
| GCA_002494985 | 79.21 | 0.00 | PARKS et al., 2017 |
| GCA_002495205 | 66.23 | 2.91 | PARKS et al., 2017 |
| **GCA_002495315** | **91.91** | **0.00** | **PARKS et al., 2017** |
| **GCA_002495905** | **87.08** | **0.00** | **PARKS et al., 2017** |
| GCA_002495965 | 68.77 | 0.97 | PARKS et al., 2017 |
| GCA_002496625 | 72.65 | 1.94 | PARKS et al., 2017 |
| GCA_002498345 | 69.90 | 2.91 | PARKS et al., 2017 |
| GCA_002499005 | 84.14 | 0.97 | PARKS et al., 2017 |
| GCA_002499525 | 81.91 | 0.00 | PARKS et al., 2017 |
| GCA_002505305 | 75.24 | 0.97 | PARKS et al., 2017 |
| GCA_002506605 | 69.74 | 0.97 | PARKS et al., 2017 |
| **GCA_002506665** | **92.23** | **0.00** | **PARKS et al., 2017** |
| GCA_002508305 | 82.45 | 0.00 | PARKS et al., 2017 |
| **GCA_002508395** | **95.15** | **0.00** | **PARKS et al., 2017** |
| **GCA_003135575** | **92.89** | **2.91** | **WOODCROFT et al., 2018** |
| **GCA_003139715** | **94.66** | **2.91** | **WOODCROFT et al., 2018** |
| **GCA_003164815** | **99.03** | **1.94** | **WOODCROFT et al., 2018** |
| **YP1** | **92.23** | **0.97** | **HUA et al 2018** |
| **UBA223** | **87.08** | **0.00** | **ANANTHARAMAN et al 2016** |
| UBA164 | 84.14 | 0.97 | ANANTHARAMAN et al 2016 |
| AD-613-B23 | 79.13 | 1.94 | PLOMINSKY et al 2018 |
| RBG_16_49_8 | 71.25 | 0.00 | ANANTHARAMAN et al 2016 |
| UBA160 | 69.74 | 0.97 | ANANTHARAMAN et al 2016 |
| UBA183 | 66.23 | 2.91 | ANANTHARAMAN et al 2016 |
| UBA57 | 62.86 | 0.00 | ANANTHARAMAN et al 2016 |
| DRTY-7 | 45.63 | 0.00 | HUA et al 2018 |
| AB-179-E04 | 33.50 | 0.00 | RINKE et al 2013 |
| EAC691 | 30.58 | 0.00 | ANANTHARAMAN et al 2016 |
| SP3992 | 18.93 | 0.00 | ANANTHARAMAN et al 2016 |
| SAT139 | 16.02 | 0.00 | ANANTHARAMAN et al 2016 |

Supplementary Table 3. Single marker genes used in the Phylogenomic analysis.

| Gene | Protein |
|---|---|
| *rpb11* | RNA pol L |
| *rpsI* | Ribosomal S9 |
| *rplR* | Ribosomal L18e |
| *rplE* | Ribosomal L5 |
| *rplN* | Ribosomal L14 |
| *rpsM* | Ribosomal S13 |
| *rpsJ* | Ribosomal S10 |
| *rplP* | Ribosomal L16 |
| *rpsS* | Ribosomal S19 |
| *rplE* | Ribosomal L5 C |
| *RNA polymerase beta subunit* | RNA pol A bac |
| *rplM* | Ribosomal L13 |
| *rplF* | Ribosomal L6 |
| *SecY* | SecY |

Supplementary Table 4. General features of metabolic pathway

| Number | ENZYME NAME | EC Number | Saci | Bog | METABOLIC PATHWAY |
|---|---|---|---|---|---|
| 1 | phosphoglucomutase | 5.4.2.2 | 1 | 1 | GLYCOLYSIS – EMP pathway |
| 2 | glucose-6-phosphate isomerase | 5.3.1.9 | 1 | 1 | GLYCOLYSIS – EMP pathway |
| 3 | 6-phosphofructokinase | 2.7.1.11 | 1 | 0 | GLYCOLYSIS – EMP pathway |
| 4 | fructose-bisphosphate aldolase | 4.1.2.13 | 1 | 1 | GLYCOLYSIS – EMP pathway |
| 5 | glyceraldehyde 3-phosphate dehydrogenase | 1.2.1.12/1.2.1.591 | 0 | 1 | GLYCOLYSIS – EMP pathway |
| 6 | phosphoglycerate kinase | 2.7.2.3 | 0 | 1 | GLYCOLYSIS – EMP pathway |
| 7 | 2,3-bisphosphoglycerate-independent phosphoglycerate mutase | 5.4.2.12 | 1 | 1 | GLYCOLYSIS – EMP pathway |
| 8 | enolase | 4.2.1.11 | 1 | 1 | GLYCOLYSIS – EMP pathway |
| 9 | pyruvate kinase/pyruvate phosphate dikinase | 1/2.7.9.1 | 1 | 1 | GLYCOLYSIS – EMP pathway |
| 36 | Fructose-1,6-bisphosphatase | 3.1.3.11 | 0 | 1 | GLUCONEOGENESIS |
| 10 | pyruvate ferredoxin oxidoreductase | 1.2.7.1 | 1 | 0 | ACETATE FORMATION |
| 11 | acetate---CoA ligase (ADP-forming) | 6.2.1.13 | 1 | 0 | ACETATE FORMATION |
| 12 | AMP phosphorylase | 2.4.2.57 | 0 | 0 | AMP METABOLISM |
| 13 | Ribose-1,5-bisphosphate isomerase | 5.3.1.29 | 0 | 0 | AMP METABOLISM |
| 14 | Ribulose-1,5-bisphosphate carboxylase | 4.1.1.39 | 1 | 0 | AMP METABOLISM |
| 16 | transketolase | 2.2.1.1 | 1 | 1 | PENTOSE PHOSPHATE PATHWAY |
| 17 | ribose-phosphate pyrophosphokinase | 2.7.6.1 | 1 | 1 | PENTOSE PHOSPHATE PATHWAY |
| 18 | ribose 5-phosphate isomerase | 5.3.1.6 | 1 | 1 | PENTOSE PHOSPHATE PATHWAY |
| 19 | ribulose-phosphate 3-epimerase | 5.1.3.1 | 0 | 1 | PENTOSE PHOSPHATE PATHWAY |
| 14 | NADH-quinone oxidoreductase | 1.6.5.11 | 1 | 1 | ELECTRON TRANSPORT CHAIN |
| 15 | Succinate dehydrogenase/fumarate reductase | 1.3.5.1 | 1 | 1 | ELECTRON TRANSPORT CHAIN |
| 28 | ubiquinol-cytochrome c reductase | 1.10.2.2 | 0 | 1 | ELECTRON TRANSPORT CHAIN |
| 29 | cytochrome c oxidase cbb3-type subunit I | 1.9.3.1 | 0 | 1 | ELECTRON TRANSPORT CHAIN |
| 30 | F-type H+-transporting ATPase | 7.1.2.2/7.1.2.2. | 1 | 1 | ELECTRON TRANSPORT CHAIN |

| | | | | |
|---|---|---|---|---|
| 20 | citrate synthase | 2.3.3.1 | 0 | 1 TCA CYCLE |
| 21 | aconitate hydratase | 4.2.1.3 | 0 | 1 TCA CYCLE |
| 22 | Isocitrate dehydrogenase | 1.1.1.42/1.1.1.41 | 1 | 1 TCA CYCLE |
| 23 | α-ketoglutarate dehydrogenase | 1.2.4.2/2.3.1.61/1.8.1.4 | 0 | 1 TCA CYCLE |
| 24 | Succinyl coenzyme A synthetase (succinate thiokinase) | 6.2.1.5 | 0 | 1 TCA CYCLE |
| 25 | Succinate dehydrogenase | 1.3.5.1 | 1 | 1 TCA CYCLE |
| 26 | Fumarase (or fumarate hydratase) | 4.2.1.2 | 1 | 1 TCA CYCLE |
| 27 | Malate dehydrogenase | 1.1.1.37 | 1 | 1 TCA CYCLE |
| | | | | |
| 31 | long-chain acyl-CoA synthetase | 6.2.1.3 | 0 | 1 FATTY ACID OXIDATION |
| 32 | acyl-coA dehydrogenase | 1.3.8.1 | 0 | 1 FATTY ACID OXIDATION |
| 33 | Enoyl-CoA hydratase | 4.2.1.17/4.2.1.150 | 0 | 1 FATTY ACID OXIDATION |
| 34 | 3-hydroxyacyl-CoA dehydrogenase | 1.1.1.35 | 0 | 1 FATTY ACID OXIDATION |
| 35 | 3-ketoacyl-CoA thiolase | 2.3.1.16 | 0 | 0 FATTY ACID OXIDATION |
| | | | | |
| 37 | acetyl-CoA carboxylase | 6.4.1.2 | 0 | CARBON FIXATION (3-HYDROXYPROPIONATE/4-0 HYDROXYBUTYRATE CYCLE) |
| 38 | malonyl-CoA reductase (NADPH) | 1.2.1.75 | 0 | CARBON FIXATION (3-HYDROXYPROPIONATE/4-0 HYDROXYBUTYRATE CYCLE) |
| 39 | malonate semialdehyde reductase (NADPH) | 1.1.1.298 | 0 | CARBON FIXATION (3-HYDROXYPROPIONATE/4-0 HYDROXYBUTYRATE CYCLE) |
| 40 | 3-hydroxypropionyl-CoA synthetase (AMP-forming) | 6.2.1.36 | 0 | CARBON FIXATION (3-HYDROXYPROPIONATE/4-0 HYDROXYBUTYRATE CYCLE) |
| 41 | hydroxypropionyl-CoA dehydratase | 4.2.1.1.16 | 0 | CARBON FIXATION (3-HYDROXYPROPIONATE/4-0 HYDROXYBUTYRATE CYCLE) |
| 42 | acryloyl-CoA reductase (NADPH) | 1.3.1.84 | 0 | CARBON FIXATION (3-HYDROXYPROPIONATE/4-0 HYDROXYBUTYRATE CYCLE) |
| 43 | propionyl-CoA carboxylase | 6.2.1.2/3 | 0 | 0 CARBON FIXATION (3- |

| | | | | |
|---|---|---|---|---|
| | | | | HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 44 | methylmalonyl-CoA epimerase | 5.1.99.1 | 0 | 0 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 45 | methylmalonyl-CoA mutase | 5.4.99.2 | 0 | 0 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 46 | succinyl-CoA reductase (NADPH) | 1.2.1.76 | 0 | 0 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 47 | succinate semialdehyde reductase (NADPH) | 1.1.1.- | 0 | 0 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 48 | 4-hydroxybutyryl-CoA synthetase (AMP-forming) | 6.2.1.- | 0 | 0 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 49 | 4-hydroxybutyryl-CoA dehydratase | 4.2.1.1.20 | 0 | 0 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 50 | crotonyl-CoA hydratase | 4.2.1.17 | 0 | 0 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 51 | 3-hydroxybutyryl-CoA dehydrogenase(NAD + ) | 1.1.1.157 | 0 | 1 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 52 | acetoacetyl-CoA b-ketothiolase | 2.3.1.16 | 0 | 0 CARBON FIXATION (3-HYDROXYPROPIONATE/4-HYDROXYBUTYRATE CYCLE) |
| 53 | Ribulose-1,5-bisphosphate carboxylase | 4.1.1.39 | 1 | 0 REDUCTIVE PENTOSE PHOSPHATE-CYCLE (CALVIN-BENSON-BASSHAM CYCLE) |
| 54 | Phosphoribulokinase | 2.7.1.19 | 0 | 0 REDUCTIVE PENTOSE PHOSPHATE-CYCLE (CALVIN-BENSON-BASSHAM CYCLE) |
| 55 | 2-Oxoglutarate synthase | 1.2.7.3 | 1 | 0 REDUCTIVE CITRIC ACID CYCLE (ARNON-BUCHANAN CYCLE) |
| 56 | ATP-citrate lyase | 2.3.3.8 | 0 | 0 REDUCTIVE CITRIC ACID CYCLE |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | (ARNON-BUCHANAN CYCLE) |
| 57 | Acetyl-CoA synthase | 2.3.1.169 | 0 | 0 | REDUCTIVE ACETYL-COA PATHWAY (WOOD-LJUNGDAHL) PATHWAY |
| 58 | CO dehydrogenase | 1.2.7.4 | 0 | 0 | REDUCTIVE ACETYL-COA PATHWAY (WOOD-LJUNGDAHL) PATHWAY |
| 59 | Malonyl CoA reductase | 1.2.1.75 | 0 | 0 | 3-HYDROXYPROPIONATE BICYCLE |
| 60 | Propionyl-CoA synthase | 6.2.1.17 | 0 | 0 | 3-HYDROXYPROPIONATE BICYCLE |
| 61 | Malyl-CoA lyase | 4.1.3.24 | 0 | 0 | 3-HYDROXYPROPIONATE BICYCLE |
| 62 | Acetyl-CoA synthase | 2.3.1.169 | 0 | 0 | 3-HYDROXYPROPIONATE-4-HYDROXYBURYRATE CYCLE |
| 63 | CO dehydrogenase | 1.2.7.4 | 0 | 0 | 3-HYDROXYPROPIONATE-4-HYDROXYBURYRATE CYCLE |
| 64 | methylmalonyl-CoA mutase | 5.4.99.2 | 0 | 1 | 3-HYDROXYPROPIONATE-4-HYDROXYBURYRATE CYCLE |
| 65 | 4-HydroxybutyrylCoA dehydratase | 4.2.1.120 | 0 | 0 | 3-HYDROXYPROPIONATE-4-HYDROXYBURYRATE CYCLE |
| 66 | 4-HydroxybutyrylCoA dehydratase | 4.2.1.120 | 0 | 0 | DICARBOXYLATE-4-HYDROXYBUTYRATE CYCLE |
| 67 | amoA (PF12942) | 1.14.99.39 | 0 | 0 | AMMONIA OXIDATION |
| 68 | amoB (PF04744) | 1.14.99.39 | 0 | 0 | AMMONIA OXIDATION |
| 69 | amoC | - | 0 | 0 | AMMONIA OXIDATION |
| **NAR** | Respiratory nitrate reductase (NirK) | 1.7.5.1 | 0 | 0 | ELECTRON TRANSPORT CHAIN |

Supplementary Table 5. General features of biosynthesis/or degradation of amino acids

| | Enzyme name | EC Number | Saci | Bog |
|---|---|---|---|---|
| | **Pyruvate degradation family (alanine, serine, glycine, cysteine, trytophan)** | | | |
| 70 | Alanine dexydrogenase | 1.4.1.1 | 0 | 0 |
| 71 | Serine deaminase | 4.3.1.17 | 0 | 0 |
| 72 | glycine betaine transmethylase | 2.1.1.5 | 0 | 0 |

| | | | | |
|---|---|---|---|---|
| 73 | dimethylglycine dehydrogenase | 1.5.8.4 | 0 | 0 |
| 74 | sarcosine oxidase | 1.5.3.1 | 0 | 1 |
| 75 | serine hydroxymethyltransferase | 2.1.2.1 | 1 | 1 |
| 76 | Serine desaminase | 4.3.1.17 | 0 | 0 |
| 77 | cystathionine β-lyase | 4.4.1.1/4.4.1.2 8 | 1 | 1 |
| 78 | cysteine dioxygenase | 1.13.11.20 | 0 | 0 |
| 79 | 3-sulfinoalanine aminotransferase | 2.6.1 | 1 | 1 |
| 80 | cysteine aminotransferase | 2.6.1.3 | 0 | 0 |
| 81 | 3-mercaptopyruvate sulfutransferase | 2.8.1.2 | 0 | 1 |
| 82 | tryptophanase | 4.1.99.1 | 0 | 0 |
| 83 | 2.6.1.44 Alanine—glyoxylate transaminase | 2.6.1.44 | 0 | 1 |
| 84 | 2.6.1.45 Serine—glyoxylate transaminase | 2.6.1.45 | 0 | 1 |
| 85 | 4.1.2.5 L-threonine aldolase | 4.1.2.5 | 0 | 0 |
| 86 | 2.1.2.1 Glycine hydroxymethyltransferase | 2.1.2.1 | 0 | 1 |
| 87 | 4.3.1.19 Threonine ammonia-lyase | 4.3.1.19 | 0 | 1 |
| 88 | Aspartate transaminase | 4.4.1.24 | 1 | 1 |
| | **Oxaloacetate and fumarate degradation family (aspartate, asparagine, tyrosine)** | | | |
| 99 | Aspartate aminotransferase | 2.6.1.1 | 0 | 0 |
| 100 | Malate dehydrogenase | 1.1.1.37 | 0 | 1 |
| 101 | Aspartate transaminase | 2.6.1.1 | 0 | 1 |
| 102 | Aspartate ammonia-lyase | 4.3.1.1 | 0 | 0 |
| 103 | Asparginase | 3.5.1.1 | 1 | 1 |
| 104 | Aspargine aminotransferase | 2.6.1.14 | 1 | 0 |
| 105 | 2-oxosuccinamate deamidase (Ps) | 3.5.1 | 1 | 1 |
| 106 | Tyrosine aminotransferase | 2.6.1.5/2.6.1.2 7/2.6.1.57 | 0 | 1 |
| 107 | 4-hydroxyphenylpyruvate dioxygenase | 1.13.11.27 | 0 | 0 |
| 108 | Homogentisate oxygenase | 1.13.11.5 | 0 | 0 |
| 109 | Maleylacetoacetate isomerase | 5.2.1.2 | 0 | 0 |
| 110 | Fumarylacetoacetate hydrolase | 3.7.1.2 | 0 | 0 |
| | **α-ketoglutarate degradation family (glutamate, glutamine, proline, arginine,** | | | |

| | | | | |
|---|---|---|---|---|
| | **histidine, aspartate, tryptophan)** | | | |
| 111 | Glutamate decarboxylase | 4.1.1.15 | 0 | 0 |
| 112 | Glutamic dehydrogenase | 1.4.1.2 | 0 | 0 |
| 113 | Glutamate mutase | 5.4.99.1 | 0 | 0 |
| 114 | Glutamate dehydrogenase | 1.4.1.3 | 1 | 1 |
| 115 | Glutaminase | 3.5.1.2/3.5.1.3 8 | 0 | 0 |
| 116 | Glutamate synthase | 1.4.1.13 | 1 | 0 |
| 117 | Proline dehydrogenase | 1.5.5.2 | 0 | 0 |
| 118 | Arginine deiminase | 3.5.3.6 | 0 | 0 |
| 119 | Arginase | 3.5.3.1 | 0 | 0 |
| 120 | Ornithine aminotransferase | 2.6.1.13 | 0 | 0 |
| 121 | L-glutamate-dehydrogenase | 1.2.1.88 | 0 | 1 |
| 122 | Pyrroline-5-carboxylate reductase | 1.5.1.2 | 1 | 0 |
| 123 | Ornithine cyclodeaminase | 4.3.1.12 | 0 | 0 |
| 124 | Histidase | 4.3.1.3 | 0 | 0 |
| 125 | Uroconase | 4.2.1.49 | 0 | 0 |
| 126 | Imidazolone-5-propionate hydrolase | 3.5.2.7 | 0 | 0 |
| 127 | Formiminoglutamate formiminohydrolase | 3.5.3.8 | 0 | 0 |
| 128 | N-formylglutamate amidohydrolase | 3.5.1.68 | 0 | 0 |
| 129 | Aspartate aminotransferase | 2.6.1.1 | 0 | 1 |
| 130 | Malate dehydrogenase | 1.1.1.37 | 0 | 0 |
| | | | | |
| | **Succinyl-CoA degradation family (valine, isoleucine, methionine)** | | | |
| 131 | L-valine:2-oxoglutarate aminotransferase | 2.6.1.42 | 0 | 1 |
| 132 | 2-oxoisovalerate dehydrogenase | 1.2.1.25 | 0 | 0 |
| 133 | Isobutyryl-CoA:FAD oxidoreductase | 1.3.8 | 0 | 0 |
| 134 | 3-hydroxy-isobutyryl-CoA hydro-lyase | 4.2.1.17 | 0 | 1 |
| 135 | 3-hydroxyisobutyrtyl-CoA hydrolase | 3.1.2.4 | 0 | 0 |
| 136 | 3-hydroxyisobutyrate dehydrogenase | 1.1.1.31 | 1 | 1 |
| 137 | Methylmalonate-semialdehyde dehydrogenase | 1.2.1.27 | 1 | 1 |
| 138 | L-isoleucine:2-oxoglutarate aminotransferase | 2.6.1.42 | 0 | 1 |
| 139 | 3-methyl-2-oxopentanoate dehydrogenase | 1.2.1.25 | 0 | 0 |
| 140 | S-2-methylbutryl-CoA:FAD oxidoreductase | 1.3.8.5 | 0 | 0 |

| | | | | |
|---|---|---|---|---|
| 141 | tiglyl-CoA hydrase | 4.2.1.17 | 0 | 1 |
| 142 | 3-hydroxy-2-methylbutyryl-CoA dehydrogenase | 1.1.1.178 | 0 | 0 |
| 143 | 2-methylacetoacetyl-CoA thiolase | 2.3.1.16 | 0 | 0 |
| 144 | S-adenosylmethionine synthase | 2.5.1.6 | 1 | 1 |
| 145 | Adenosylhomocysteinase | 3.3.1.1 | 1 | 1 |
| | **Acetyl-CoA, acetoacetyl-CoA and acetoacetate degradation family (leucine, threonine, isoleucine, lysine, phenylalanine, tyrosine)** | | | |
| 146 | L-leucine:2-oxoglutarate aminotransferase | 2.6.1.6/2.6.1.42 | 0 | 1 |
| 147 | 4-methyl-2-oxopentanoate dehydrogenase | 1.2.1.25 | 0 | 0 |
| 148 | Isovaleryl-CoA:FAD oxidoreductase | 1.3.8.4 | 0 | 0 |
| 149 | 3-methylcotonyl-CoA carboxylase | 6.4.1.4 | 0 | 0 |
| 150 | 3-methylglutaconyl-CoA hydratase | 4.2.1.18 | 0 | 0 |
| 151 | Hydroxymethylglutaryl-CoA lyase | 4.1.3.4 | 0 | 0 |
| 152 | Threonine aldolase | 4.1.2.5/1.2.48 | 0 | 0 |
| 153 | Acetaldehyde dehydrogenase | 1.2.1.10 | 0 | 0 |
| 154 | L-isoleucine:2-oxoglutarate aminotransferase | 2.6.1.42 | 0 | 0 |
| 155 | 3-methyl-2-oxopentanoate dehydrogenase | 1.2.1.25 | 0 | 0 |
| 156 | S-2-methylbutyryl-CoA:FAD oxidoreductase | 1.3.8.5 | 0 | 0 |
| 157 | Tiglyl-CoA hydrase | 4.2.1.17 | 0 | 1 |
| 158 | 3-hydroxy-2-methylbutyryl-CoA dehydrogenase | 1.1.1.178 | 0 | 0 |
| 159 | 2-methylacetoacetyl-CoA thiolase | 2.3.1.16 | 0 | 0 |
| 160 | L-lysine monooxygenase | 1.13.12.2 | 0 | 0 |
| 161 | Aminovaleramidase | 3.5.1.30 | 0 | 0 |
| 162 | 5-aminovalerate transaminase | 2.6.1.48 | 0 | 0 |
| 163 | Glutarate semialdehyde dehydrogenase | 1.2.1 | 0 | 1 |
| 164 | Succinyl-CoA-glutarate CoA-transferase | 2.8.3.13 | 0 | 0 |
| 165 | Phenylalanine transaminase | 2.6.1.1 | 0 | 0 |
| 166 | Phenylpyruvate | 4.1.1.43 | 0 | 0 |
| 167 | Phenylacetaldehyde dehydrogenase | 1.2.1.39 | 0 | 0 |
| 168 | Tyrosine aminotransferase | 2.6.1.5/2.6.1.57/2.6.1.57 | 0 | 1 |
| 169 | Fumarylacetoacetate hydrolase | 3.7.1.2 | 0 | 0 |

# Appendix B. Supplementary Material of Chapter 4

**Supplementary Material and Methods**

**Candidate Phyla Radiation (CPR)/Patescibacteria in public soil metagenome datasets.** We retrieve from Genbank (August 2018) all CPR/Patescibacteria identified in the Woodcroft and colaboradors (2018) study about Metagenome-assembled genomes (MAGs) from a thawing permafrost metagenome dataset (n=19). Up to date, this was the most complete soil MAGs dataset. A specific 51 CPR/Patescibacteria single-marker copy genes was used to re-estimate the completeness and contamination parameters and retrieve only the genomes with the high-quality draft (completeness > 90% and contamination < 5%) recommended by the Minimum information about a metagenome-assembled genome (MIMAG) of bacteria and archaea (BOWERS et al., 2017). Due to the reduced size of the CPR/Patescibacteria genomes, during the evolution, some universal single-marker genes were lost, and to get a better estimation of the genome quality, we selected only all single-marker genes (51 genes) were presented in all complete CPR/Patescibacteria genomes deposited on NCBI (Genome Taxonomy Database classification/August 2018) (Supplementary Table 6). To build the specific single-marker gene database, we calculated the completeness and contamination using the universal single-maker genes HMMs dataset from Albertsen and collaborators (2013). The CheckM (PARKS et al., 2015) was used to identify the specific CPR/Patescibacteria single-marker copy and we considered to the downstream genome quality analysis only the genes which were presented in all complete genomes. To complement our comparative analyzes, we also used three high-quality soil CPR/Patescibacteria reconstructed by KROEGER AND COLLABORATORS (2018).

**Cattle-pasture soil sampling, DNA extraction, and metagenomic sequencing.** The soil was collected in an adjacent area of the Tapajós National Forest, in the state of Pará, Eastern Amazon. The topsoil 0-10 cm from a cattle-pasture was collected in July 2015. This soil was used in a microcosm experiment (VENTURINI, 2019; unpublished data) to increase the soil moisture level to 100% at field capacity under 30-day. Total DNA was extracted using PowerLyzer PowerSoil DNA Isolation Kit (MO Bio Laboratories,Carlsbad, CA, USA). The quality and quantity of the DNA samples were evaluated using agarose gel electrophoresis stained with ethidium bromide and a Nanodrop 2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The metagenome was sequenced on an Illumina HiSeq 2500 platform (2x 250 bp) (Illumina, San Diego, CA, USA) at Novogene Corporation, Beijing, China.

**Metagenomic assembly, binning and quality control.** To remove the low-quality reads (parameters: Phred score (<30) and minimum size (<100 bp)) the SICKLE software (JOSHI; FAUS, 2011) was used. The cattle-pasture metagenome was assembled in Megahit (LI et al., 2015) with default parameters (Supplementary Table 1). The downstream analysis was performed according to as in Lemos et al. (2017). Briefly, contigs smaller than 10,000 bp were removed from downstream analyses and coverage was calculated using Bowtie2 (LANGMEAD; SALZBERG, 2012). Stringent length filtering parameters were used in order to reduce contamination and remove chimeric contigs. Binning was performed with MaxBin 2.0 (WU et al., 2016) and quality control metrics

(completeness/contamination) were calculated with CheckM (PARKS et al., 2015). Taxonomy was assigned using the 16S rRNA gene with the RDP Classifier software (WANG et al., 2007) and the SILVA database (version 32) (QUAST et al., 2013). Additional phylogenomic information was inferred by the GTDB-Tk software and GTDB (Genome Taxonomy Database) (PARKS et al., 2018).

**Functional annotation and multivariate statistics.** We annotated 891 microbial genomes, including 857 deposited on RefSoil database (CHOI et al., 2017), 18 from the permafrost metagenome dataset (WOODCROFT et al., 2018), ten symbiotic/parasitic microbial genome, three soil CPR/Patescibacteria genomes, the Udaeobacter copiosus (BREWER et al., 2017) and two new CPR/Patescibacteria described here (Supplementary Table 4), using the PROKKA pipeline (SEEMANN et al., 2014) to identify the ORFs (Open Reading-frames) and the COG (cluster of orthologous groups) categories. The query proteins were blasted with RPS-BLAST+ (Reverse Position-Specific BLAST) against NCBI Conserved Domain Database (CDD). The multivariate analysis was performed on R platform.

References:

BROWERS, R.M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. **Nature Biotechnology**, London, v. 35, n. 8, p. 725-731, 2017.

CHOI, J. et al. Strategies to improve reference databases for soil microbiomes. **The ISME Journal**, London, v. 11, n. 4, p. 829–834, 2017.

KROEGER, M. E. et al. New Biological Insights into How Deforestation in Amazonia Affects Soil Microbial Communities Using Metagenomics and Metagenome-Assembled Genomes. **Frontiers in Microbiology**, Lausanne, v. 9, 2018.

JOSHI, N.A.; FASS, J.N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle., 2011.

PARKS, D. H. et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. **Genome Research**, New York, v. 25, n. 7, p. 1043–1055, 2015.

LANGMEAD, B.; SALZBERG, S. Fast gapped-read alignment with Bowtie 2. **Nature Methods**, London, v. 9, p. 357-359, 2012.

LEMOS, L.N. et al. Genome-Centric Analysis of a Thermophilic and Cellulolytic Bacterial Consortium Derived from Composting. **Frontiers in Microbiology**, Lausanne, v. 8, p. 1-16, 2017.

LI, D. et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. **Bioinformatics**, Oxford, v. 31, n. 10, p. 1674-1676, 2015.

QUAST, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic Acids Research**, Oxford, v. 41, n. Database issue, p. D590–D596, 2013.

PARKS, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. **Nature Biotechnology**, London, v. 36, n. 10, p. 996–1004, 2018.

WANG, Q. et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. **Applied and Environmental Microbiology**, Washington, v. 73, n. 16, p. 5261-5267, 2007.

WOODCROFT, B.J. Genome-centric view of carbon processing in thawing permafrost. **Nature**, London, v. 560, 2018.

WU, Y.-W. et al. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. **Microbiome**, London, v. 2, n. 1, p. 26, 2014.

Supplementary Table 1. General informations about the metagenomic dataset used in this study to reconstruct microbial genomes.

| **Parameter** | **Pasture metagenome** |
|---|---|
| Sample description | Pasture soil incubated by 30-days under a moisture-controlled experiment. |
| Location | 2°25′48″S 54°43′12″W |
| Number of paired-end reads | 149,386,627 |
| Number of assembled contigs | 6,231,429 |
| N50 (bp) | 719 |
| Longest contig (bp) | 321,699 |
| Total assembled length (bp) | 4,272,955,462 |
| Number of assembled contigs ≥ 10,000 bp | 7,309 |
| Total assembled contigs length ≥ 10,000 bp | 122,505,934 |

Supplementary Table 2. Genomic features of two new CPR/Patescibacteria

| **Genome (bin) name** | Curupira | Caipora |
|---|---|---|
| Phyla | Patescibacteria | Patescibacteria |
| Class/Order/Family | Microgenomatia/Shapirobacterales/UBA 12405 | Doudnabacteria/UBA920/UBA920 |
| Estimated Genome Size (bp) | 1,306,057 | 1,302,885 |
| Number of contigs | 23 | 24 |
| Best hit (16S rRNA) (NCBI Accession Number) | Uncultured bacterium clone YH79 [JQ861406.1] | Uncultured bacterirum clone C-134 [KC836053] |
| Coverage/Identity (%) | 97/92 | 97/93 |
| Estimated Completeness (%) | 100.0 | 100.0 |
| Estimated Contamination (%) | 0.0 | 0.0 |
| G+C content (%) | 34.33 | 43.01 |
| Maximum scaffold length (bp) | 321,669 | 192,499 |
| N50 contig length | 91,855 | 67,807 |
| CDS number | 1,276 | 1,332 |

Supplementary Table 3. General features of metabolic pathway.

| Pathway | Genes | COG | Caipora | Curupira | GCA_003151615.1 |
|---|---|---|---|---|---|
| EMP pathway | Glucokinase | COG0837 | 0 | 0 | 0 |
| | Glucose-6-phosphate isomerase | COG0166 | 1 | 0 | 0 |
| | Glucose-6-phosphate 1-dehydrogenase | COG0364 | 1 | 0 | 0 |
| | Phosphofructokinase | COG0205 | 0 | 0 | 0 |
| | Fructose-bisphosphate aldolase | COG3588 | 0 | 0 | 1 |
| | Triosephosphate isomerase | COG0149 | 1 | 0 | 1 |
| | Glyceraldehyde-3-phosphate dehydrogenase | COG0057 | 1 | 1 | 0 |
| | Phosphoglycerate kinase | COG0126 | 0 | 0 | 1 |
| | Phosphoglycerate mutase | COG0696 | 2 | 1 | 1 |
| | Enolase | COG0148 | 1 | 1 | 1 |
| | Pyruvate kinase | COG0469 | 1 | 0 | 0 |
| | Phosphoenolpyruvate synthase | COG0574 | 2 | 2 | 0 |
| Pentose phosphate patway | Transketolase | COG3959 | 0 | 0 | 1 |
| | ribulose-5-phosphate 3-epimerase | COG3623 | 0 | 0 | 1 |
| | Ribose-5-phosphate isomerase | COG0120 | 0 | 0 | 0 |
| | Ribokinase | COG0524 | 2 | 1 | 0 |
| Tricarboxylic acid cycle | Citrate synthase | COG0372 | 0 | 0 | 0 |
| | Aconitase | COG1048 | 0 | 0 | 0 |
| | Isocitrate dehydrogenase | COG0538 | 0 | 0 | 0 |
| | 2-oxoglutarate dehydrogenase E1 component | COG0567 | 0 | 0 | 0 |
| | Succinyl-CoA synthetase | COG0045 | 0 | 0 | 0 |
| | Succinate dehydrogenase | COG0479 | 0 | 0 | |
| | Fumarase | COG0114 | 0 | 0 | 1 |
| Eletron transport | | | | | |
| | NADH-quinone oxidoreductase | COG1034 | 0 | 0 | 0 |

| | | | | | |
|---|---|---|---|---|---|
| | Succinate dehydrogenase/fumarate reductase | COG2009 | 0 | 0 | 0 |
| | ubiquinol-cytochrome c reductase | COG5605 | 0 | 0 | 0 |
| | cytochrome c oxidase cbb3-type subunit I | COG2993 | 0 | 0 | 0 |
| | F-type H+-transporting ATPase | COG0356 | 0 | 0 | 0 |
| Oxidative stress | Superoxide dismutase | COG0605 | 1 | 1 | 1 |
| | Catalase | COG0376 | | | |
| | Peptide methionine sulfoxide reductase MsrA | COG0225 | 1 | 0 | 4 |
| Type IV pili | | | | | |
| | Type IV fimbrial assembly, ATPase PilB | COG2804 | 2 | 2 | 0 |
| Protein translocation Sec dependent | | | | | |
| | SecA | COG0653 | 2 | 1 | 1 |
| | SecY | COG0201 | 1 | 1 | 1 |
| | SecE | COG0690 | 1 | 1 | 1 |
| | SecG | COG1314 | 0 | 0 | 0 |
| | YidC | COG0759 | 0 | 0 | 0 |
| | Sortase | COG3764 | 0 | 0 | 0 |
| Sugar degradation | Alpha-amylase | COG1449 | | | |
| | Glucoamylase | COG3387 | 0 | 0 | 0 |
| | Beta-glucosidase/6-phospho-beta-glucosidase/beta-galactosidase | COG2723 | 1 | 0 | 0 |
| | Beta-glucosidase (EC 3.2.1.21) | COG2723 | 1 | 0 | 0 |
| Amino acids biosynthesis | Asparagine (aspartate--ammonia ligase) | COG2502 | 0 | 0 | 0 |
| | asparagine synthase (glutamine-hydrolysing) | COG0367 | 0 | 0 | 0 |
| | Glutamine (glutamine synthetase) | COG0174 | 0 | 0 | 0 |
| | Proline (glutamate 5-kinase) | COG0263 | 0 | 0 | 0 |
| | glutamate-5-semialdehyde dehydrogenase | COG0014 | 0 | 0 | 0 |

| | | | | |
|---|---|---|---|---|
| pyrroline-5-carboxylate reductase | COG0345 | 0 | 1 | 0 |
| Arginine (ornithine carbamoyltransferase) | COG0078 | 0 | 0 | 0 |
| argininosuccinate synthase | COG0137 | 0 | 0 | 0 |
| argininosuccinate lyase | COG0165 | 0 | 0 | 0 |
| Cysteine (serine O-acetyltransferase) | COG1045 | 0 | 0 | 0 |
| cysteine synthase | COG0031 | 0 | 0 | 0 |
| O-acetylhomoserine (thiol)-lyase | COG2873 | 0 | 0 | 0 |
| Methionine (aspartate kinase) | COG0527 | 0 | 0 | 0 |
| aspartate-semialdehyde dehydrogenase | COG0136 | 0 | 0 | 1 |
| 5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase | COG0620 | 0 | 0 | 0 |
| Serine (D-3-phosphoglycerate dehydrogenase / 2-oxoglutarate reductase) | COG0111 | 0 | 0 | 0 |
| phosphoserine aminotransferase | COG1932 | 0 | 0 | 0 |
| Threonine (aspartate kinase) | COG0527 | 0 | 0 | 0 |
| aspartate-semialdehyde dehydrogenase | COG0136 | 0 | 0 | 1 |
| threonine synthase | COG0498 | 0 | 0 | 0 |
| Lysine (aspartate kinase) | COG0527 | 0 | 0 | 0 |
| diaminopimelate decarboxylase | COG0019 | 1 | 0 | 1 |
| Tryptophan (3-dehydroquinate dehydratase I) | COG0710 | 0 | 0 | 0 |
| 3-phosphoshikimate 1-carboxyvinyltransferase | COG0128 | 0 | 0 | 0 |
| Phenylalanine/Tyrosine (chorismate mutase) | COG1605 | 0 | 0 | 0 |
| aromatic-amino-acid transaminase | COG1448 | 0 | 0 | 0 |
| Alanine (cysteine desulfurase) | COG1104 | 1 | 1 | 2 |
| Histidine (ATP phosphoribosyltransferase) | COG0040 | 0 | 0 | 0 |
| histidinol dehydrogenase | COG0040 | 0 | 0 | 0 |

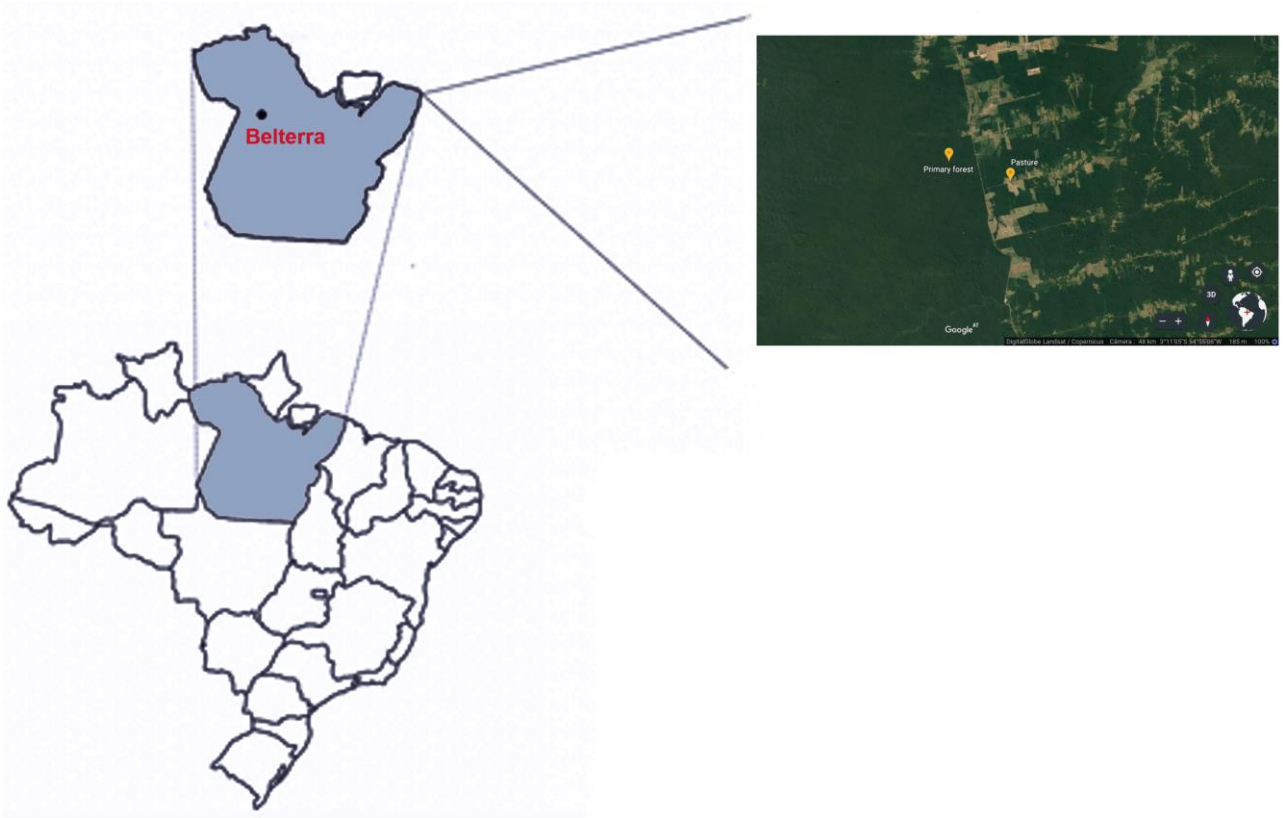
| | | | | | |
|---|---|---|---|---|---|
| | Glycine (glycine hydroxymethyltransferase) | COG0112 | 1 | 1 | 1 |
| | Valine (valine transaminase) | COG3977 | 0 | 0 | 0 |
| | Tyrosine (Aspartate/tyrosine/aromatic aminotransferase) | COG1448 | 0 | 0 | 0 |
| Fermentation | Lactato (Malate/lactate dehydrogenase) | COG0039 | 1 | 1 | 1 |
| De novo synthesis of pyrimidine | Carbamoyl phosphate synthetase (CPS, EC 6.3.5.5) | COG0458 | 0 | 0 | 0 |
| | Aspartate transcarbamoylases (ATC, EC 2.1.3.2) | COG0540 | 0 | 0 | 0 |
| | Dihydroorotase (DHO, EC 3.5.2.3) | COG0418 | 0 | 0 | 0 |
| | Dihydroorotate dehydrogenase (DODH; EC 1.3.99.11) | COG0167 | 1 | 0 | 1 |
| | Uridine 5′-monophosphate synthase (UMPS, EC 2.4.2.10 plus 4.1.1.23) | COG0284 | 1 | 0 | 0 |
| | UMP kinase (UMPK, EC 2.7.4.4) | COG0572 | 0 | 0 | 0 |
| | Nucleoside diphosphate kinase (NDPK, EC 2.7.4.6) | COG0105 | 1 | 1 | 1 |
| | CTP synthetase (CTPS, EC 6.3.4.2) | COG0504 | 1 | 1 | 1 |
| De novo synthesis of purine | Phosphoribosylpyrophosphate synthetase | COG0462 | 0 | 1 | 1 |
| | IMP dehydrogenase | COG0516 | 0 | 0 | 0 |
| | GMP synthetase | COG0518 | 1 | 0 | 1 |
| | sAMP synthetase | COG0104 | 0 | 0 | 0 |
| | sAMP lyase | COG0015 | 0 | 0 | 0 |

Supplementary Table 4. Single-cell genomes described by Choi et al. (2016)

| NCBI ID | Total Contigs > 2,200 bp | Total Assembled Length (bp) | Maximum Contig Length (bp) | Completeness (%) | Contamination (%) |
|---|---|---|---|---|---|
| LSSX00000000 | 200 | 2.503.189 | 85.609 | 47.75 | 0.0 |
| LSSY00000000 | 69 | 709.610 | 29.805 | 0.0 | 0.0 |
| LSSZ00000000 | 126 | 1.547.648 | 52.772 | 4.17 | 0.0 |
| LSTA00000000 | 93 | 1.170.883 | 60.329 | 22.86 | 0.0 |
| LSTB00000000 | 86 | 1.314.140 | 84.789 | 41.18 | 0.0 |
| LSTC00000000 | 162 | 2.314.649 | 140.388 | 52.67 | 1.21 |
| LSTD00000000 | 316 | 4.337.762 | 105.657 | 28.69 | 0.0 |
| LSTE00000000 | 125 | 1.462.098 | 57.674 | 66.95 | 0.0 |
| LSTF00000000 | 123 | 2.058.214 | 85.098 | 56.24 | 0.74 |
| LSTG00000000 | 50 | 281.276 | 17.877 | 9.48 | 0.0 |
| LSTH00000000 | 237 | 2.314.199 | 44.079 | 33.2 | 0.0 |
| LSTI00000000 | 82 | 995.360 | 43.153 | 24.14 | 0.0 |
| LSTJ00000000 | 154 | 1.513.572 | 42.901 | 20.85 | 0.0 |
| LSTK00000000 | 69 | 615.657 | 28.470 | 4.17 | 0.0 |

# Appendix C. Supplementary Material of Chapter 5

## Supplementary Figure




**Supplementary Figure 1. Sampling map from Belterra municipality, in the state of Pará, Brazil.** Soil sampling was carried out in the Tapajós National Forest (3°17'44.4"S, 54°57'46.7"W) a well-preserved primary forest and a cattle pasture (3°18'46.7"S, 54°54'34.8"W) next to the forest.

**Supplementary Table 1. Accumulated emissions of methane $CH_4$ (ng C-$CH_4$ soil $g^{-1}$) in forest and pasture soils from Eastern Amazon under different moisture levels (control, 60, 80, and 100% of field capacity) in a microcosm experiment carried out for 30 days.** Adapted and reprinted with permission from VENTURINI (2019).

| Treatment | Forest | Pasture |
|---|---|---|
| Control | -693,4±415,1***** | 337,3±242,4 |
| 60% | -1042,2±715,6 | -1909,7±1378,8 |
| 80% | -294,4±345,5 | 68,6±374,5 |
| 100% | 287,3±298,1 | 8013,7±1379,3 |

*****Mean±Standard error

# Appendix D. Scientific and Teaching Contributions

## 6.1. Scientific papers and book chapters

Below are the contributions to scientific knowledge as a direct (marked with an asterisk) or indirect consequences of the work developed during this thesis:

1. LEMOS, L. N. et al. Bioinformatics for Microbiome Research: Concepts, Strategies, and Advances. In: PYLRO, V.; ROESCH, L. (Eds.). **The Brazilian Microbiome: Current Status and Perspectives**. Cham: Springer International Publishing, London, 2017. p. 111–123. *

2. LEMOS, L. N. et al. Metagenome sequencing of the microbial community of two Brazilian anthropogenic Amazon dark earth sites, Brazil. **Genomics Data**, Amsterdam, v. 10, p. 167–168, 2016. *

3. NAVARRETE, A. A. et al. Zinc concentration affects the functional groups of microbial communities in sugarcane-cultivated soil. **Agriculture, Ecosystems & Environment**, Amsterdam, v. 236, p. 187–197, 2017.

4. ABDALLA FILHO, A. L. et al. Diets based on plants from Brazilian Caatinga altering ruminal parameters, microbial community and meat fatty acids of Santa Inês lambs. **Small Ruminant Research**, Amsterdam, v. 154, p. 70–77, 2017.

5. ARAUJO, A. S. F. DE et al. Protist species richness and soil microbiome complexity increase towards climax vegetation in the Brazilian Cerrado. **Nature Communications Biology**, London, v. 1, n. 1, p. 1–8, 2018.

6. BRESCIANI, L. et al. Draft Genome Sequence of "Candidatus Spirobacillus cienkowskii," a Pathogen of Freshwater Daphnia Species, Reconstructed from Hemolymph Metagenomic Reads. **Microbiology Resource Announcements**, New York, v. 7, n. 22, p. e01175-18, 2018.

7. LUPATINI, M. et al. Moisture Is More Important than Temperature for Assembly of Both Potentially Active and Whole Prokaryotic Communities in Subtropical Grassland. **Microbial Ecology**, Amsterdam, v. 77, n. 2, p. 460–470, 2019.

8. MIRANDA, A. R. L. et al. Dynamics of archaeal community in soil with application of composted tannery sludge. **Scientific Reports**, London, v. 9, n. 1, p. 1–8, 2019.

9. LEMOS, L. N. et al. Genomic signatures and co-occurrence patterns of the ultra-small Saccharimonadia (CPR/Patescibacteria phylum) suggest a symbiotic lifestyle. **Molecular Ecology**, Amsterdam, 2019.

### 6.2. Ad-hoc journal Reviewer

PeerJ; FEMS Microbiology Ecology; Parasitology; Frontiers in Environmental Science; BMC Microbiology. PLoS One; Scientific Reports; Genomics

### 6.3. Teaching experience

● Invited Teacher (Dec/2017 – Dec/2017) - Bioinformatics course for Next-generation sequencing data in Microbial Ecology - Federal University of Sao Carlos (UFScar/Brazil): Postgraduate Program in Biotechnology.

● Invited Teacher (Aug/2017 – Aug/2017) - Bioinformatics course for Next-generation sequencing data in Microbial Ecology - Federal University of Viçosa (UFV/Brazil): Postgraduate Program in Microbiology.

● Invited Teacher (2015 – 2017) – Bioinformatics, Genomics and microbial ecology – University of Sao Paulo (USP/Brazil): Bachelor in Agronomy and Bachelor in Life Sciences.